Classifying Sidewalk Materials Using Multi-Modal Data

Thesis

Submitted in partial fulfillment of the requirements for the degree

Master of Data Science and Engineering

At

The City College of the City University of New York

By

Jiawei Liu

May 2023

Approved by:

Professor Hao Tang — Professor Zhigang Zhu

Thesis Advisor

Professor Michael Grossberg — Professor Zhigang Zhu

Co-Directors, Data Science and Engineering Program

Classifying Sidewalk Materials Using Multi-Modal Data

Jiawei Liu

Data Science and Engineering Program

Thesis Advisor: Professor Hao Tang, Professor Zhigang Zhu

Abstract

Navigating safely and independently presents considerable challenges for people who are blind or have low vision (BLV), as it requires a comprehensive understanding of their neighborhood environment. Our user study reveals that materials and objects on sidewalks play a crucial role in navigation tasks. Unfortunately, current methods for assessing sidewalk materials are suboptimal, often relying on labor-intensive and expensive manual assessments that fail to capture the full range of sidewalk features critical to individuals with BLV.

In response to this problem, this master's thesis investigates deep learning approaches specifically designed for the classification of multi-modal sidewalk materials. The proposed framework aims to empower individuals with BLV to automatically gather information about sidewalk materials while navigating their surroundings. This innovative solution comprises two primary components. (1) First, the study focuses on designing a lightweight data collection methodology that involves attaching an inertial measurement unit (IMU) and a microphone to the white cane. This sensor design enables the measurement of the haptic and audio feedback, represented by acceleration data and audio data, respectively, as the white cane interacts with the sidewalk surface. The collected acceleration and acoustic signal data effectively capture the unique characteristics of different sidewalk materials. Utilizing this novel data collection method, we have successfully generated a multi-modal sidewalk material (MSM) dataset, encompassing a wide range of sidewalk material categories. (2) the research develops a deep learning-based classifier to identify different sidewalk materials using this multi-modal data. We investigate two model architectures: the ResNet-Encoder model and the Transformer-Encoder model to understand their efficacy in sidewalk material classification. Experimental results indicate that the ResNet-Encoder model provides superior performance, achieving an optimal accuracy of 83% when trained with 4-second-long data clips.

In summary, our research has significant implications for the development of AI-based assistive navigation solutions for individuals with BLV. It contributes to both the methodology for sidewalk material data collection and the algorithm of deep learning for sidewalk

material classification. By employing the proposed multi-modal deep learning approach, BLV people can effortlessly acquire information about sidewalk materials. Furthermore, this data can be utilized to generate an urban accessibility geospatial map, thereby facilitating independent travel for individuals with BLV.

Keywords: Machine Learning, Sensor Design, Multi-modal Data, Audio Data Processing, Assistive Navigation, Blind or Low Vision

Acknowledgements

I would first and foremost like to express my deepest gratitude to my thesis advisors, Professor Hao Tang and Professor Zhigang Zhu. Their expertise, understanding, and guidance have been the compass that steered my academic journey towards this significant milestone. They have provided me with unparalleled support, nurtured my curiosity, and challenged me in ways that have cultivated my growth as a researcher.

I am particularly indebted to Professor Hao Tang, who has been much more than an academic advisor. His unwavering belief in my capabilities, even in times when I doubted myself, has been a source of strength and inspiration. Beyond the confines of academia, Professor Hao Tang has generously imparted invaluable life lessons and wisdom. His mentorship has had a profound impact not only on my academic track but also on my personal development. My sincere thanks also go to my fellow lab members at the City College of Visual Computing Lab (CCVCL) for their feedback, support, and friendships.

In addition, I must extend my appreciation to my family, who have been my backbone throughout this endeavor. Their love, encouragement, and faith in my abilities have been a driving force behind my motivation and determination. They have stood by me unflinchingly, weathering the storms of this arduous process. I am truly blessed to have them in my life.

Lastly, to all those who have directly or indirectly supported me throughout this journey, I extend my heartfelt gratitude. This thesis is not just a testament to my hard work, but also a reflection of the collective efforts of those who have been part of my academic voyage. Thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The World Health Organization (WHO) estimates that there are 285 million people with visual impairment worldwide, among whom 39 million are totally blind [31]. People who are blind or have low vision (BLV) face many challenges in their daily lives, including the difficulty of navigating safely and independently [11]. In order to navigate effectively, individuals with BLV need to acquire as much spatial information as possible from their surroundings, including information about sidewalk materials and defects [21]. Regrettably, most existing advanced applications [2, 5, 7] do not provide sufficient functionality to help BLV people collect landmark information and understand sidewalk conditions. Mobile navigation applications with GPS and mapping services (such as Google Maps and Apple Maps), mainly focus on finding efficient, short navigation routes. However, while this is beneficial for the average user, it is insufficient for BLV people [1, 4]. Their preferences tilt towards paths rich in tactile landmarks and minimal sidewalk defects, prioritizing safety and reliability over shorter distances.

1.2 Problem Statement

As for the BLV individuals, they often rely on white canes to scan their surroundings, receiving auditory and haptic feedback that not only enhances their spatial awareness but also assists in self-localization [21]. For example, they can follow the tactile shoreline in their travel by identifying surface material changes, such as grass edges or raised curbs. Many street intersections are equipped with tactile pavements of varying materials and patterns, designed to aid BLV people in identifying important locations like street crossings, bus stops, and the direction of streets. These surface materials serve as invaluable landmarks and their inclusion in accessible maps is crucial. These maps prove to be incredibly useful to BLV people, facilitating real-time navigation and also serving as tools for orientation and mobility training. To better understand the challenges the BLV people are facing in their life, we also conducted an informal user study, which reveals that materials and objects on sidewalks play a crucial role in navigation tasks.

Although New York City has comprehensive sidewalk and intersection designs, but current regulations do not provide sufficient data resources for sidewalk materials [8]. This oversight leaves a huge gap in the creation of fully accessible maps that are essential to BLV people. Given the laborious and time-consuming nature of large-scale data collection, the urgency for a more efficient sidewalk material data collection solution is becoming increasingly apparent.

1.3 Overview of Proposed Solution

To address this problem, we proposed an Artificial Intelligence (AI)-based data collection framework that enables BLV people to independently survey sidewalk material information during their travels. This innovative solution comprises two primary contributions.

The first contribution is the design of a lightweight data collection methodology dedicated to the acquisition of non-visual information of sidewalk materials. While deep learning methods have shown impressive performance in numerous image recognition applications [18, 36], acquiring image data is a challenge for individuals with BLV, especially in cluttered sidewalk environments. This highlights the need for alternative non-visual information sources. Notably, BLV people rely heavily on non-visual sensory feedback from the white canes to discriminate various landmarks, including surface materials, suggesting that non-visual information holds key landmark characteristics. To capitalize on this, we equipped the white cane with an inertial measurement unit (IMU) and microphone that captures haptic and auditory feedback in the form of acceleration and audio data as the cane interacts with the sidewalk surface. In addition, we have guided a group of volunteer students to collect a large amount of sidewalk material data by applying this innovative data collection method, leading to the generation of a high-quality, multi-modal sidewalk material (MSM) dataset.

The second contribution focuses on the algorithm development of a deep learning-based classifier, which is able to identify different sidewalk materials using multi-modal data. We investigate two main model architectures, the ResNet-encoder model and the Transformer-encoder model, to understand their efficacy in sidewalk material classification. The choice of these models is guided by their specific computational strengths. The ResNet-Encoder model uses its convolutional layers to learn the regional information of the input data

[32]. This ability to extract, process, and classify local patterns from the data makes it a suitable candidate for our solutions, where understanding the specific characteristics of the sidewalk materials is crucial [10, 22, 27, 28, 42, 43]. On the other hand, the transformer-encoder model emphasizes the learning of temporal information and is able to identify the relationships between the various temporal observations from our input data [41]. This ability to discern temporal relationships has the potential to capture the complex nuances of various material sounds over time, which is a key aspect in understanding different sidewalk materials [12, 15, 20, 26, 48].

Through a comprehensive study of these models, we aim to investigate their effectiveness in sidewalk material classification, offering valuable insights into the potential of deep learning techniques in enhancing the robustness of our AI-enabled data collection framework.

1.4 Outline of the Thesis

The remainder of this thesis is structured as follows: Chapter 2 reviews related work in accessibility data collection, material recognition, and audio classification, providing crucial context for our study. Chapter 3 details our innovative data collection approach, introducing the lightweight collection equipment, assistive mobile application, and the multi-modal sidewalk material dataset. Chapter 4 describes the sidewalk material classification process, introduces the training dataset, and describes the technical aspects of our two models, the ResNet-Encoder Model and the Transformer-Encoder Model. We conduct a series of experiments to investigate their efficacy and conclude with a comparative analysis. In Chapter 5, we present our conclusions and future work, starting with a comprehensive summary of our findings followed by a reflection on unexpected results.

We discuss the limitations identified and suggest potential improvements and present a detailed plan for evaluating the performance of our model on unlabeled continuous data. The chapter concludes by exploring potential applications of our findings to real-world scenarios.

Chapter 2

Related Work

2.1 Crowdsourcing and Accessibility Data Collection

Local and state governments often need to collect data on street accessibility [6]. Thanks to the ubiquitous Internet and mobile technologies, data can be collected more efficiently and economically using crowdsourcing methods [29]. A wide range of studies has highlighted the efficacy of crowdsourcing methods to collect a large amount of data on street-level images, where they focus on urban road construction and beautification [39, 47]. Beyond the field of urban construction improvement, there are several studies that focus on data collection on sidewalk accessibility. Several papers have used crowdsourcing and Google Street View (GSV) to allow people to remotely identify bus stops and curb ramps [17, 37]. One notable study designed a web-based crowdsourcing platform that collaboratively leverages GSV and image detection models to collect storefront accessibility data [25]. Moreover, a methodology similar to our approach was introduced in another study, where a tri-axial accelerometer was fitted under a wheelchair seat. This setup was utilized to infer sidewalk accessibility features such as slope and curb presence from

the behavior of the wheelchair [44]. Analogously, in our project, we mounted a wearable sensor and mini microphones on a white cane and collected sidewalk material data while the BLV people walked with the white cane.

2.2 Material Recognition

Most existing research on material recognition relies heavily on visual cues, with some progress in deep learning approaches. One notable study [40] achieved significant results by focusing on three key elements: material image datasets, contextual influences, and unique descriptors of material appearance. In addition, numerous studies have explored the utility of light field (LF) images for material identification [23]. LF images provide richer light information, thus enriching the scope of vision-based measurement applications, including material identification. An alternative view of material recognition has been proposed through the use of haptic acceleration signals. In combination with surface images, a fully convolutional network has been deployed for joint surface material recognition [46]. In contrast, our project mainly utilizes non-visual data, specifically acceleration and audio data. Our deep learning classifier aims to use both forms of data to discriminate sidewalk materials, thus providing a new perspective in the field of material recognition.

2.3 Audio Classification

The popularity of deep learning has increased dramatically in recent years, emerging as a reliable approach for a wide range of machine learning tasks, including audio classification. Various audio classification tasks are addressed by deep learning algorithms,

such as speech recognition, music classification, and environmental sound classification [13, 16]. A prevalent trend in this domain involves the preprocessing of raw audio data to convert it into spectrograms, including Mel-Spectrogram and Mel-frequency cepstral coefficients (MFCC). These characteristic representations then serve as inputs to intricate network models for training. Several studies have affirmed the effectiveness of Convolutional Neural Network (CNN) based models when applied to spectrograms [28, 42]. Remarkably, most state-of-the-art results have been achieved through transfer learning, employing pre-trained CNN models like ResNet50 [22]. Interestingly, one notable study indicated that CNNs pre-trained with regular images, such as ImageNet, remain proficient at extracting critical features from audio spectrograms [33]. On the contrary, a part of the research still advocates a traditional approach, using recurrent neural network (RNN) based models to explore continuous audio information [12, 15, 20, 26, 48]. Recognizing the power and popularity of transformer structures, a recent study introduced an audio spectrogram transformer model that achieved state-of-the-art results in different classification tasks [15]. Following these advances, our project explores two main model architectures: the ResNet-Encoder model and the Transformer-Encoder model. We aim to gauge their effectiveness in the specific task of sidewalk material classification.

Chapter 3

Data Collection

3.1 Introduction of Data Collection Methodology

One of the main contributions of this project is to design a data collection methodology for BLV people to collect the data on sidewalk materials by themselves. People with BLV typically navigate various sidewalk surfaces with their white canes, which interact with a range of distinct sidewalk materials, generating distinct haptic and acoustic feedback in the process. Each type of feedback is uniquely representative of the sidewalk material, playing a crucial role in providing BLV individuals with essential information about their surroundings.

The translation of sidewalk material characteristics into distinguishable haptic and auditory feedback forms the basis of our data collection methodology. These two feedbacks are subjective, relying on individual BLV people's experience. For example, when the cane strikes the sidewalk surface, it generates a sensory response in the form of haptic and acoustic feedback. These responses convey crucial information about the sidewalk material type, texture, and condition. Consequently, our data collection methodology

is intentionally designed to capture these two determinative feedbacks. Specifically, our data collection methodology addresses two key challenges:

1. **Lightweight Data Collection:** We needed to ensure that the equipment used to collect haptic and acoustic feedback did not burden the BLV individual. Thus, a primary concern was the design and selection of lightweight collection equipment that did not interfere with the cane's use or compromise the user's comfort.

2. **Simultaneous Multi-modal Data Collection:** The integrity of our data hinges on the precise and simultaneous collection of multi-modal data, capturing both haptic and acoustic feedback. Hence, a significant aspect of our methodology focuses on ensuring the accuracy and completeness of the data collected, while still remaining user-friendly and unobtrusive to the BLV people.

3.1.1 Light-Weight Collection Equipment

As previously stated, our data collection methodology is fundamentally designed to acquire haptic and acoustic feedback. It is crucial to understand the inherent significance of these two forms of feedback and the methodology we have implemented for their precise measurement.

Haptic feedback is a sensory reflection of sidewalk material characteristics. As BLV people swing their canes over various sidewalk surfaces, the interactions with different materials generate varied degrees of resistance. These changes in resistance are discerned by the user as distinct tactile sensations, representing the haptic feedback. In this project, we have devised a system to convert these sensory experiences into quantifiable data. To do so, we employ an Inertial Measurement Unit (IMU) on the white cane to record the acceleration of the white cane in a three-dimensional (3D) space. The recorded

acceleration data serves as a proxy for the resistance felt by the user and, thus, for the haptic feedback. Streamlining the data acquisition process is crucial for the practical implementation of our methodology. Therefore, to ensure the stability of the IMU and the quality of the recorded data, we have set a configuration of 400 Hz for the IMU. This configuration effectively balances the need for high-quality data and the practical constraints of data acquisition, paving the way for the effective collection of acceleration data for our study.

Acoustic feedback is another crucial piece of information we want to collect in our data collection methodology. When the white cane interacts with different sidewalk mounted a microphone on the white cane, specifically positioned near the cane tip to maximize the clarity of recorded sounds while minimizing ambient noise.

FIGURE 3.1: The white cane is equipped with a microphone and IMU sensor (the white device is the IMU sensor and the black unit is a mini microphone attached to a blue sponge)

With this configuration (Fig 3.1), we can efficiently collect acoustic and haptic feedback, represented by audio data and acceleration data, respectively. This dual mode of data

collection allows for the creation of a rich, high-quality, multi-modal dataset. Moreover, the lightweight nature of both the wearable IMU sensor and the mini microphone ensures mobility and comfort for the user, allowing BLV individuals to integrate this solution seamlessly into their daily routines.

3.1.2 Accessible Mobile Application

To facilitate the simultaneous collection of multi-modal data, audio data, and acceleration data, we designed a dedicated iOS mobile application called "SidewalkVacuum". This application constitutes a key element of our data collection methodology, providing an efficient and user-friendly platform for BLV people to engage in self-directed data collection.

A key feature of the "SidewalkVacuum" application is the incorporation of a third-party API from the IMU platform. This integration permits the application to interface directly with the IMU sensor, effectively managing its operations. As such, the application can remotely control the initiation, pausing, and termination of the sensor's data acquisition process, ensuring an effortless interaction between the user and the sensor. Building on this seamless integration, we devised an innovative algorithm to simultaneously capture video and acceleration data, which is useful to our multi-modal data collection framework. The video data includes audio signals from the sidewalk environment, serving as auditory feedback, while the acceleration data acquired from the IMU sensor act as a proxy for the haptic feedback as experienced by BLV people.

Notably, we've also embedded annotation functionality into our application, empowering sighted data collectors to label the data in real time during the collection process. This

feature enables us to collect an extensive and high-quality multi-modal dataset of sidewalk materials, which is a vital training resource for our deep-learning classifier. We will discuss this dataset in more detail in the subsequent section. However, it's important to note that for BLV people, the annotation function is disabled, meaning all data collected by them remains unlabeled and requires a deep learning classifier for interpretation. Through meticulous system design and integration, we've succeeded in interconnecting all facets of our application, thereby ensuring a seamless and efficient user experience in multi-modal data collection.

After successfully acquiring the data, our application will also store the geographic coordinates of the current location. Later it will package the location information, video, and acceleration data in preparation for transferring them to our secure cloud-based database. To accommodate user preferences and data limitations, we provide an option for users to postpone the data upload process until a WIFI connection is available, thus saving the use of cellular data. To further improve transfer speed and data security, we developed a dedicated algorithm to split long data records into manageable 1-minute segments. This segmentation not only speeds up the upload process but also adds an additional layer of data integrity, minimizing the risk of data loss during transmission.

Once in the cloud, these data will be saved under different users' folders. This cloud-based architecture ensures convenient access to the dataset for subsequent analysis and model training, facilitating scalability and efficiency. It also guarantees data safety and allows for the rapid accumulation of information over time.

3.2 Multi-modal Sidewalk Material Dataset

While we have established a comprehensive and robust data collection framework, it constitutes only half of our overarching objective. The second and equally critical part involves the development of a deep-learning classifier. This AI-driven classifier was programmed to automatically identify unlabeled data collected by the community of BLV people. Therefore, it becomes critical to produce a large-scale, high-quality multi-modal sidewalk material (MSM) dataset. This dataset is indispensable because it provides foundational training material for refining the accuracy and reliability of our deep learning classifier.

To assemble the MSM dataset, a large-scale data collection effort was initiated involving 27 student volunteers. These students were tasked with collecting extensive sidewalk material data in three different boroughs: Brooklyn, Manhattan, and Queens. To ensure accuracy and relevance, we classified the collected sidewalk material data strictly according to the criteria specified in the New York City Street Design Manual [8] and the Guidebook for Accessible Sidewalk and Street Intersection Information [3].

After several months of intensive data collection, we successfully released the MSM dataset with a total duration of 7 hours. This dataset represents a diverse and comprehensive encapsulation of 11 primary categories and 2 secondary categories of sidewalk materials. The primary categories span a variety of common sidewalk materials including concrete, asphalt, dirt, grass, metal, manhole, granite, tactile pavement, brick, subway grate, and cellar door. To further enrich our dataset, we also introduced two secondary categories: concrete mixed with stone and concrete tactile pavement. The relationship between these categories is illustrated in the following figure (Fig 3.2).

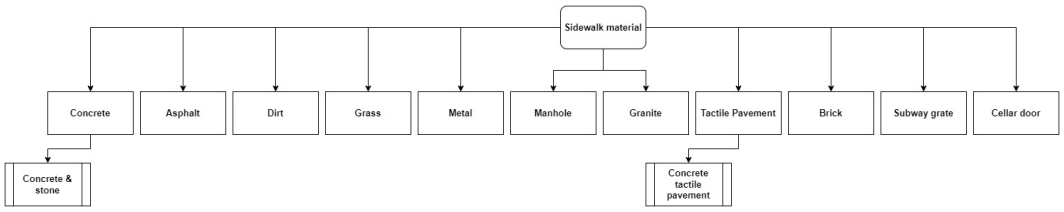

FIGURE 3.2: Tree structure of sidewalk material categories

Considering the uncommon nature of some sidewalk materials, it is reasonable to assume that the distribution of the data we collected is unbalanced. The figure below (Fig 3.3) demonstrates the distribution of all sidewalk material types, where the y-axis represents the duration in minutes. From the figure, we can see that the main sidewalk material types such as concrete, subway grate, etc. are still relatively easy to collect. Conversely, certain subcategories, such as concrete tactile pavement, presented a more substantial collection challenge due to their rarity.

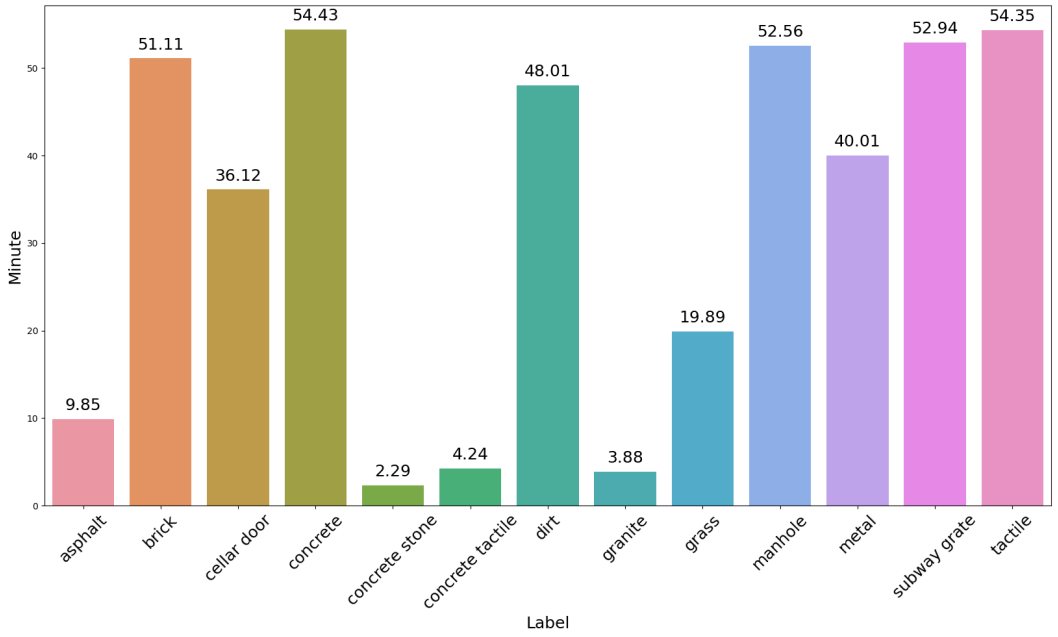

FIGURE 3.3: Distribution of multi-modal sidewalk material dataset

In summary, despite the inherent challenges associated with certain material properties, our robust MSM dataset provides an inclusive representation of typical and atypical sidewalk materials, thereby creating a valuable resource for further research and applications in this field.

Chapter 4

Sidewalk Material Classification

4.1 Dataset

4.1.1 Overview of Training Dataset

Training a deep learning model necessarily requires a large amount of data. In this study, we strategically used the top nine categories with the largest amount of data in the MSM dataset as our training dataset. These categories (Fig 4.1) include a wide variety of sidewalk materials: concrete, tactile pavement, subway grate, manholes, bricks, dirt, metal, cellar doors, and grass.

FIGURE 4.1: Images of nine classes

Figure 4.2 shows the data distribution for these categories. The figure highlights data-rich categories such as concrete, tactile pavement, subway fencing, manholes, and bricks that collectively represent the prevalent types of sidewalk materials. Each of these categories encompasses data with a duration of around 50 minutes, creating a robust foundation for our model training. In contrast, specific categories, such as dirt, metal, cellar door, and grass have less data, with grass having the least amount of data at 19.89 minutes. This discrepancy is largely due to the challenges inherent in collecting high-quality data for these specific sidewalk material types. Materials like metal, cellar doors, and especially grass often prove to be more difficult to collect data on than concrete or brick, which is more commonly encountered.

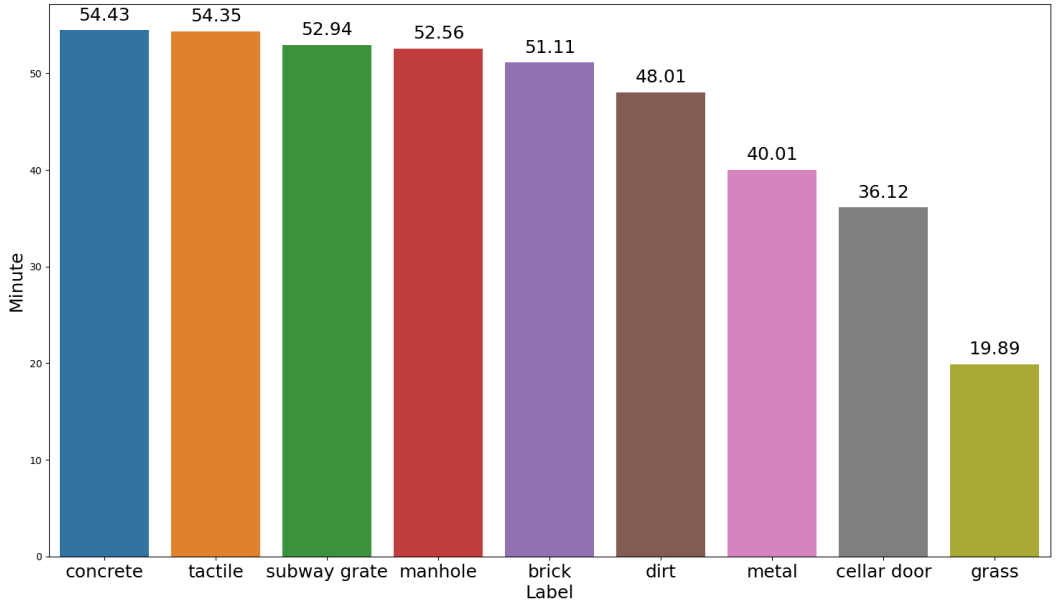

FIGURE 4.2: The distribution of the training dataset

It is worth noting that although this training dataset is not perfectly balanced, it does provide a great deal of diversity for our deep learning model. The nine categories do not fully cover all possible types of street material, but the available data still provide a solid foundation for the training process.

4.1.2 Data Preprocessing

Data preprocessing is a crucial step in any machine learning project for transforming raw data into a form that machine learning models can learn effectively. In this project, the goal of data preprocessing is twofold: to slice our multi-modal data into manageable, trainable pieces, and to convert these pieces into a format suitable for deep learning classifiers to learn. Figure 4.3 provides a schematic diagram of the data preprocessing pipeline used in this study. The pipeline consists of three main components: data preparation (Fig 4.3, Part I), data slicing (Fig 4.3, Part II), and data transformation (Fig 4.3, Part III). These components are introduced in detail in the following subsections.

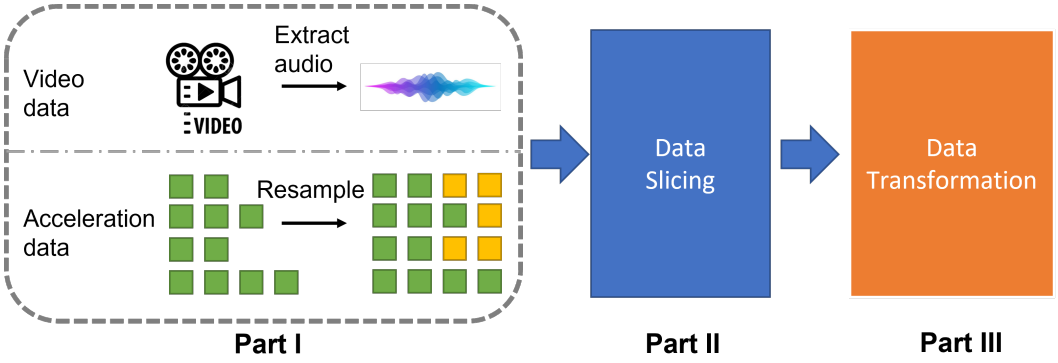

FIGURE 4.3: Schematic diagram of data preprocessing pipeline

4.1.2.1 Data Preparation

The first step in our data preprocessing pipeline is data preparation (Fig 4.3, Part I), including the extraction of audio data from the video data and the resampling of acceleration data to a uniform frequency of 400 Hz. Although our IMU was configured to collect data at 400 Hz, it still occasionally records at a slightly lower frequency. To maintain uniformity, we resample the data using the Sinc interpolation method [45].

The Sinc function is a well-acknowledged technique in signal processing, mainly used for its accuracy in reproducing the original signal's frequencies and its robustness to missing

or irregularly-spaced samples. The Sinc function is given by:

$$\text{sinc}(x) = \begin{cases} 1, & \text{if } x = 0 \\ \frac{\sin(\pi x)}{\pi x}, & \text{otherwise} \end{cases} \tag{4.1}$$

Applying Sinc interpolation allowed us to resample the acceleration data to 400 Hz without causing significant distortion or information loss, ensuring consistent and uniform data for model training.

4.1.2.2 Data Slicing

The next step is data slicing (Fig 4.3, Part II), which is the process where we decompose the original multi-modal data into manageable segments that are suitable for training our deep learning classifier. The sliding window technique operates by first establishing a window of a fixed length that moves across the data sequence with a determined step size. Each shift of this window generates a new data segment, enabling the extraction of localized features from the time series data. Notably, the choice of window length and step size directly affects the amount of data samples. The shorter the window length and step length, the larger the number of data samples. Figure 4.4 demonstrates the process of data slicing.

This method is particularly useful for dealing with sequential data such as audio data, where temporal dependencies exist. By maintaining a fixed window length, the sliding window technique ensures that each sliced data segment encompasses a complete cycle of the sequential pattern present in the dataset. This encapsulation is critical for preserving the inherent temporal correlations in the data and allows the model to learn from the sequential characteristics of the audio and acceleration signal.

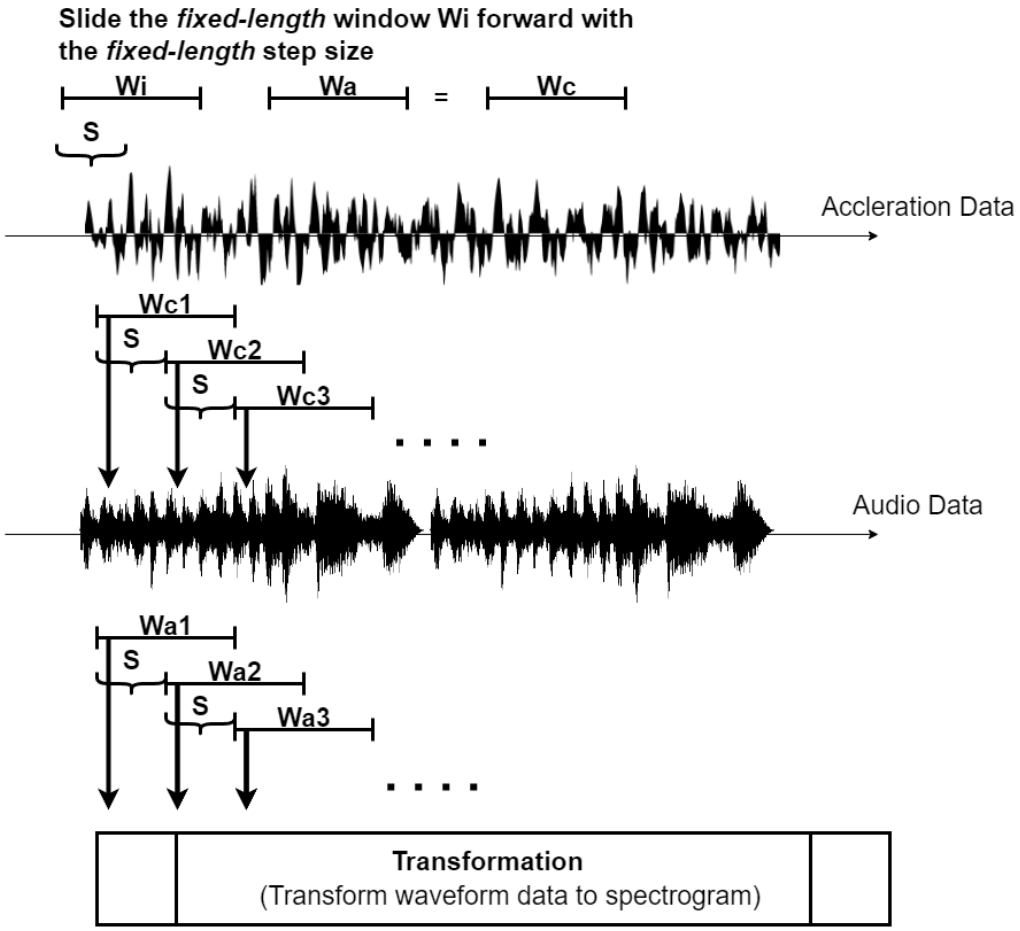

FIGURE 4.4: Schematic diagram of data slicing

4.1.2.3 Data Transformation

The final step is data transformation (Fig 4.3, Part III) where we performed a Mel-spectrogram transformation on the segmented audio and acceleration data. This transformation maps raw data onto a two-dimensional grid, with the horizontal axis representing time and the vertical axis denoting frequency. The Mel-spectrogram efficiently captures both the spectral and temporal properties of the raw audio and acceleration signals, which are crucial for our sidewalk material classification task. There are three important steps for Mel-spectrogram transformation.

First, we apply a Short-Time Fourier Transform (STFT) [34] to the windowed signal which is audio data and acceleration data in our case. STFT is denoted as below, where

$W[t]$ is the window function:

$$STFT(x(t)) = X(m, \omega) = \sum_{n=0}^{N-1} x[n] \cdot W[t] \tag{4.2}$$

Once we obtain the spectrogram by applying STFT, we then map the result onto the Mel scale, which approximates the human ear's response more closely than the linearly-spaced frequency bands. The formula to convert the frequency f to Mel scale m is given by:

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \tag{4.3}$$

Next, we apply Mel filter banks [30] to the spectrogram. These filter banks are triangular filters that are overlapped such that each filter's peak corresponds to the center frequency of the previous filter. The Mel-filter bank function is given by:

$$H_m(k) = \begin{cases} 0, & \text{if } k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)}, & \text{if } f(m-1) \leq k \leq f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)}, & \text{if } f(m) \leq k \leq f(m+1) \\ 0, & \text{if } k > f(m+1) \end{cases} \tag{4.4}$$

Finally, we take the logarithm of all the Mel-filter banks to obtain the final Mel-spectrogram. Figure 4.5 showcases the Mel-spectrograms of audio data clips for nine classes.

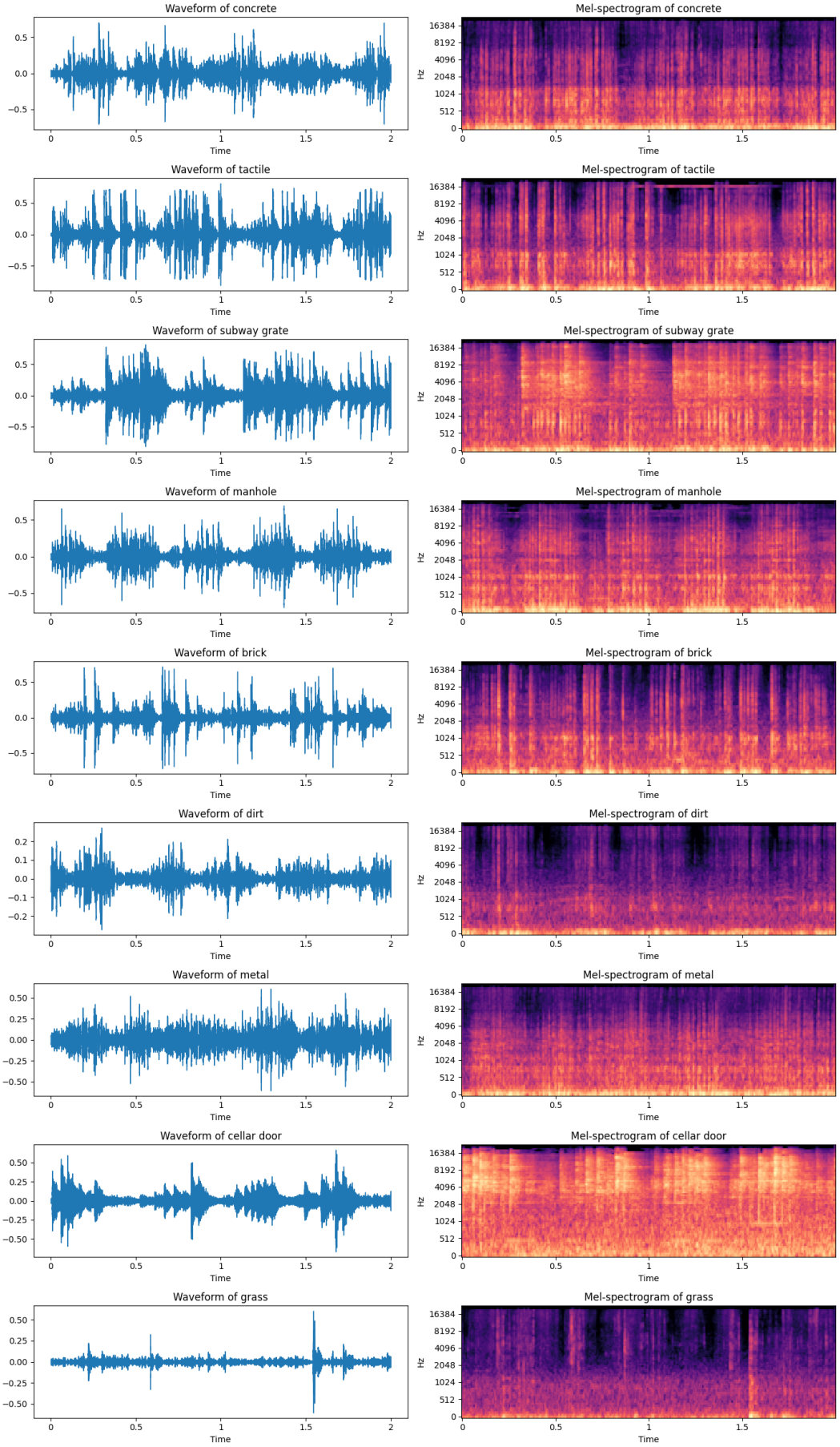

FIGURE 4.5: Mel-spectrograms of audio data clips for nine classes (the clip length is 2-second long)

4.1.3 K-Fold Cross-Validation

To achieve a rigorous and accurate assessment of model performance, we used a K-fold cross-validation approach [9]. In the scope of this study, our dataset was divided into eight folds, each of which was systematically stratified to include all eight categories. For instance, when we slice the audio and acceleration data with a step size of 0.5 seconds and a window length of 2 seconds, we are able to get the dataset shown in the following table (Tab 4.1). Notably, this dataset will be used as a training dataset in the following experiments.

K-fold	**#Samples**	**#Classes**	**#Collectors**
0	6402	9	19
1	5715	9	19
2	5964	9	18
3	5700	9	20
4	5959	9	21
5	5915	9	18
6	5166	9	20
7	5675	9	18

TABLE 4.1: Overview of data distribution among k-folds with respect to sample count, class diversity, and collector numbers.

As shown, each folder contains approximately 5,000 to 6,000 data samples distributed in equal proportions across nine different categories. In addition, the data in each folder was compiled by a different collector, thus ensuring the generalizability and breadth of the data. Importantly, our implementation of the K-fold cross-validation technique is performed on the original, full-length video recordings prior to any sliced data. This ensures that sliced data segments are isolated in their respective folds and do not bleed into other folds, thus maintaining data integrity and preventing any potential cross-fold contamination.

Considering computational constraints and time efficiency, we did not perform exhaustive cross-validation for all eight folds. Instead, each model was cross-validated using three

unique folds chosen at random. In the design of this project, seven of the eight folds have been allocated for model training, with one fold reserved for testing purposes. The training data is derived from 80% of the seven designated training folds, while the remaining 20% is earmarked for validation. This approach ensures a reasonable estimate of model performance, maintaining an essential equilibrium between computational demands and the practical constraints of the study.

4.2 ResNet-Encoder Model

4.2.1 The Overview of the ResNet-Encoder Model

An overview of the ResNet-Encoder model is depicted in Figure 4.6. Within the ResNet-Encoder model, we employ the transfer learning technique to expedite the learning process, effectively leveraging pre-trained models to extract generic featuresr. Considering the remarkable generality and power of the ResNet model [19], we chose ResNet-50 as the encoder to extract features from the input data.

Our study involves multi-modal data consisting of two input sources: audio and acceleration data. Each of these inputs was transformed into a Mel-spectrogram, a format that facilitates our classification task. In the designed model, each ResNet-50 encoder is assigned an input data type that extracts the underlying representation from the corresponding Mel-spectrogram (Fig 4.6, Part I). The features extracted by each encoder are combined into a composite vector, which is then imported into a fully connected layer, often known as the fusion layer (Fig 4.6, Part II). With respect to the structure of the model, we determined that the output representation of each ResNet encoder is (512*1), thus defining the size of the fusion layer as (1024*1).

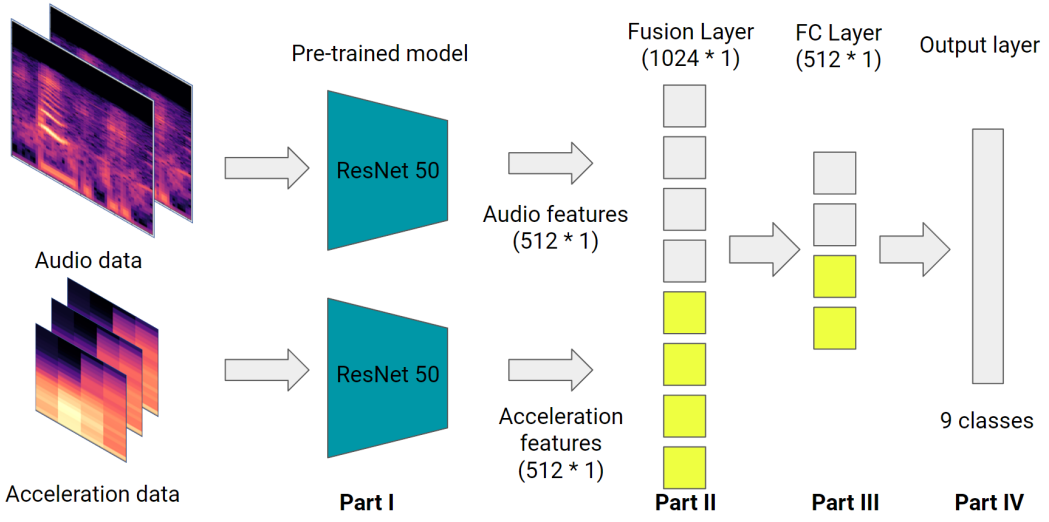

FIGURE 4.6: RetNet-Encoder model architecture

A critical element of our model structure is the inclusion of dropout and batch normalization techniques after each fully connected layer. This technique serves as a protection against overfitting. Specifically, we used a dropout rate of 0.2, a strategy that involves randomly deactivating 20% of the neurons during the training process. Simultaneously, batch normalization was implemented to stabilize the learning process and aid in preventing overfitting. Batch normalization standardizes the inputs to each layer of the model, reducing internal covariate shifts. As for the activation function, the Rectified Linear Unit (ReLU) was selected for its simplicity and effectiveness in deep neural networks. ReLU is computationally efficient and helps in mitigating the vanishing gradient problem, facilitating a more effective learning process within our model's architecture.

After passing through the fusion layer, the representation is passed to another fully connected layer (Fig 4.6, Part III), configured in the shape of (512*1). Subsequently, the output is directed to the output layer (Fig 4.6, Part IV) of the model. For the loss function, we use a cross-entropy loss function, a choice based on its ability to handle multi-class classification tasks.

4.2.2 Experiment 1: Assessing the Impact of Individual and Combined Inputs on Model Performance

The first experiment we conducted aims to understand the relative contribution and combined potential of acceleration and audio data in identifying sidewalk materials when fed into our ResNet-Encoder model. Since our data collection method involves the capture of these two types of data, it is crucial to investigate how each, and their combination, affects the performance of the model. This investigation can help us understand how the model learns and interprets this interconnected data.

The following table (Tab 4.2) presents the accuracy and F1-scores for the ResNet-Encoder model when trained with different input types:

Input	**Accuracy**	**F1-Score**
Acceleration data	46.83%	46.59%
Audio data	76.24%	76.55%
Acceleration and audio data	77.64%	77.68%

TABLE 4.2: Model performance of ResNet-Encoder model with varying input data

The findings from this experiment provide several interesting insights. When the model was trained with acceleration data alone, it achieved an accuracy of 46.83% and an F1-score of 46.59%, demonstrating a modest performance. However, when trained solely with audio data, the model showed a significant improvement in both accuracy (76.24%) and F1-score (76.55%). This suggests that the audio data may contain more distinctive features or patterns that aid in classifying sidewalk materials than the acceleration data. Perhaps the most important finding from this experiment was that using a combination of acceleration and audio data further improved the model's performance, with the accuracy and F1-score rising to 77.64% and 77.68%, respectively. This indicates that while the acceleration and audio data individually provide certain insights for the classification

task, they also complement each other when used together, capturing a more comprehensive representation of sidewalk materials. Therefore, this investigation substantiates the effectiveness of a multi-modal data approach for the sidewalk material classification task.

4.2.3 Experiment 2: Evaluating the Effect of Freezing Pre-trained Layers on Model Performance

The motivation behind this experiment was to examine how freezing different numbers of pre-trained layers of the ResNet-Encoder model impacts the classification performance. Pre-trained models carry learned features from large datasets that might not align perfectly with the features of our specific task. Consequently, tweaking the extent to which we allow our model to adjust these pre-existing features during training, by freezing a certain number of layers, may lead to a better match with our task-specific feature characteristics and thereby enhance model performance [38].

The table (Tab 4.3) below summarizes the results of this experiment, detailing the accuracy and F1-score achieved by the model with varying numbers of frozen layers and for different input types:

Input	**#Freezed Layers**	**Accuracy**	**F1-Score**
Acceleration data	0	48.71%	48.39%
Acceleration data	1	48.84%	48.39%
Acceleration data	2	46.83%	46.59%
Audio data	0	78.02%	78.56%
Audio data	1	78.24%	78.93%
Audio data	2	76.24%	76.55%
Acceleration and audio data	2	77.64%	77.68%
Acceleration and audio data	1	79.24%	79.68%

TABLE 4.3: Model performance of ResNet-Encoder model with varying numbers of frozen layers and input data

The results highlight some intriguing trends. When training on acceleration data alone, freezing one layer resulted in a slight improvement in accuracy (48.84%), but the F1-score remained consistent (48.39%). However, freezing two layers led to a decrease in both accuracy and F1-score (46.83% and 46.59%, respectively). This suggests that freezing too many layers can prevent the model from adequately learning from acceleration data. A similar pattern is observed with audio data. Freezing one layer led to a marginal improvement in both accuracy and F1-score (78.24% and 78.93%, respectively). However, when two layers were frozen, the model's performance dropped slightly.

Most notably, when combining acceleration and audio data, freezing one layer of the model resulted in the best performance (accuracy of 79.24% and F1-score of 79.68%), surpassing the performance obtained with two frozen layers. We will use this configuration as the default setting in the next experiment.

4.2.4 Experiment 3: Investigating the Impact of Data Clip Length on Model Performance

The objective of this experiment was to examine the impact of data clip length on the performance of our ResNet-Encoder model. This analysis is essential as it helps in understanding how the model's ability to identify sidewalk materials changes with variations in the amount of temporal information contained within each data clip. It is hypothesized that data clips of longer duration could potentially provide the model with more context and detailed information from audio and acceleration data, thereby influencing its ability to classify sidewalk materials accurately. In this experiment, we will use the multi-modal ResNet-Encoder model, which means that the input data includes both audio and acceleration data.

The table (Tab. 4.4) below presents the experimental results, summarizing the accuracy and F1-score achieved by the ResNet-Encoder model when trained with 2-second and 4-second data clips:

Clip Length	**#Samples**	**Accuracy**	**F1-score**
2-second	46496	79.24%	79.68%
4-second	43924	83.25%	83.43%

TABLE 4.4: Model performance of ResNet-Encoder model with varying clip length

From the results, it is apparent that the data clip length plays a substantial role in the performance of the ResNet-Encoder model. Despite having a larger number of samples, the 2-second data clips yielded a lower accuracy and F1-score (79.24% and 79.68%, respectively) compared to the 4-second data clips (83.25% accuracy and 83.43% F1-score). This finding suggests that the longer clips provided a richer, more detailed representation of the sidewalk materials, allowing the model to capture subtler patterns and distinctions that were possibly missed in the shorter clips. This experiment, therefore, illustrates the significance of data clip length as a parameter to consider during data preprocessing. Future work might explore an even wider range of clip lengths to identify the optimal temporal scope for classifying sidewalk materials using audio and acceleration data.

4.3 Transformer-Encoder Model

4.3.1 The Overview of the Transformer-Encoder Model

An overview of the Transformer-Encoder model is depicted in Figure 4.7. In the Transformer-Encoder model, we use the Transformer architecture as an encoder to extract features from the input data. With the attention mechanism, Transformer-Encoder is able to identify the relationships between the various temporal information. This ability to discern temporal relationships has the potential to capture the complex nuances of various

material sounds over time, which is essential to our study. In terms of the architecture, this Transformer model is designed with the same architecture as the traditional, scalable Natural Language Processing (NLP) transformer architecture [41]. In this study, the Transformer-Encoder is configured with two layers for efficiency in computation. Moreover, the self-attention mechanism within the encoder uses eight attention heads. Likewise, the type of input data considered in this model is multimodal, including audio data and acceleration data, both of which are uniquely processed.

With respect to audio data, it is transformed into a Mel-spectrogram, which is a two-dimensional representation of the signal. This representation conveys the power distribution of the signal over time, across different frequencies. Each Mel-spectrogram is split into multiple frames, with each frame corresponding to a specific temporal window. The frames contain the number of Mel-frequency bands that capture the frequency components of the signal within that specific window. In our case, each frame has 64 Mel-frequency bands. In addition, it is important to add a positional encoding (Fig 4.7, Part I) to each frame before it passes through the Transformer Encoder (Fig 4.7, Part II). Positional encoding helps maintain this sequence information, assigning a unique indicator to each position in the input sequence, thereby preserving the sequential context. The processed information from the Transformer-Encoder, which incorporates both feature and positional information, is then passed to an aggregation layer (Fig 4.7, Part III). This layer condenses the high-dimensional output into a one-dimensional vector. This vector serves as a significant representation of the audio data.

For the acceleration data, we process it using standard statistical methodologies. Notably, we apply the feature extraction methodology on each axis (x-axis, y-axis, and z-axis) from the data clip. The methods are listed below.

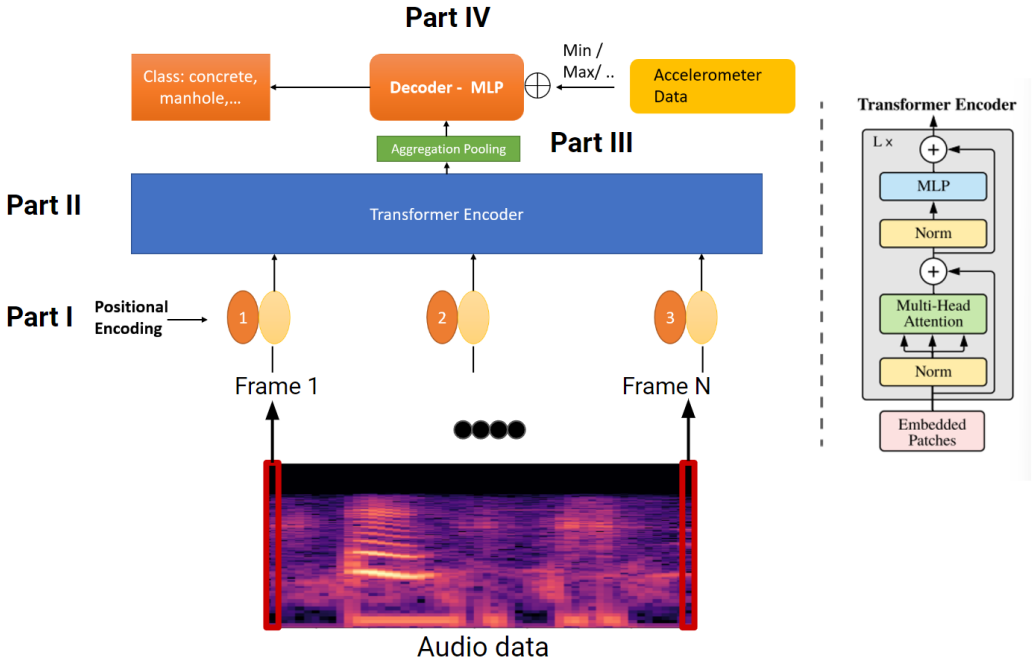

FIGURE 4.7: Transformer-Encoder model architecture

- **Mean:** The mean of each axis of the acceleration data is computed as the average of all data points.
- **Standard Deviation:** The standard deviation measures the dispersion of the data points for each axis.
- **Median Absolute Deviation:** This measure provides an understanding of the statistical dispersion of each acceleration axis.
- **Value-Mean Difference:** This feature provides a measure of the deviation of each data point from the mean.
- **Min, Max:** These values represent the minimum and maximum data points for each axis.
- **Max-Min Difference:** This measures the range of the data for each axis.

- **Interquartile Range:** This is a measure of statistical dispersion, computed as the difference between the 75th percentile (Q3) and 25th percentile (Q1) of the data.
- **Negative Count, and Positive Count:** These are the total counts of negative and positive data points, respectively.
- **Values Above Mean:** This is the count of data points exceeding the mean for each axis.
- **Number of Peaks:** This represents the total count of peak values in the acceleration data.
- **Skewness:** This is a measure of the asymmetry of the data distribution for each acceleration axis.
- **Energy:** This quantifies the total energy of the acceleration signal.
- **Kurtosis:** This measures the "tailedness" of the distribution.
- **Signal Magnitude Area (SMA):** This represents the integral of the absolute value of the acceleration data over an interval.

These features will be combined to form a comprehensive representation of acceleration data, which is then processed further for classification tasks.

Once we have these two important representations for both audio and acceleration data, we then merge them and input them to the decoder layer (Fig 4.7, Part IV) for classification purposes. Likewise, this model utilizes the cross-entropy loss function for loss computation.

4.3.2 Experiment: Comparative Analysis of Different Aggregation Pooling Methods on Model Performance

The purpose of this experiment is to investigate the effect of different aggregation pooling methods on the performance of the transformer-encoder model. Aggregation pooling is an important component for processing the feature sequences obtained from the Transformer encoder output data. Specifically, we will use aggregation methods to extract a one-dimensional representation vector from the Transformer encoder output data that contains the most important features of the input data. The method of aggregation affects the model's ability to capture relevant temporal patterns in the sequence, affecting its overall performance. The purpose of this experiment is to compare three common aggregation pooling methods - maximum, average, and self-attention pooling [24] - and to understand how they affect the classification of sidewalk material.

The following table (Tab 4.5) summarizes the results of this experiment, showing the accuracy and F1-scores of different aggregation pooling methods and clip lengths:

Aggregation pooling methods	**Clip Length**	**Accuracy**	**F1-score**
Max	4-second	70.93%	71.02%
Mean	4-second	71.25%	71.34%
Self-Attention Pooling	2-second	73.02%	73.45%
Self-Attention Pooling	4-second	73.31%	74.47%

TABLE 4.5: Model performance of Transformer-Encoder model with varying clip length and aggregation pooling method

The results of this experiment reveal distinct patterns concerning the model's performance across different aggregation pooling methods. The Max and Mean pooling methods achieved similar accuracy and F1-score, with Mean pooling marginally outperforming Max pooling. However, the Self-Attention Pooling method demonstrated superior performance, achieving the highest accuracy and F1-score for both 2-second and 4-second data clips.

Interestingly, the clip length did not have a significant impact on the model's performance with Self-Attention Pooling, with both 2-second and 4-second data clips yielding similar accuracy and F1-score. This suggests that the self-attention pooling method is efficient in capturing and summarizing the relevant temporal patterns within the data clips, regardless of their length.

4.4 Comparative Analysis of ResNet-Encoder Model and Transformer-Encoder Model

The purpose of this section is to provide an in-depth comparative analysis of the two complex deep learning architectures used in our study, the ResNet-Encoder and Transformer-Encoder models. Both models were put through their paces in multi-modal sidewalk material classification and their performance was evaluated based on two key metrics - accuracy and F1-score.

Our comparison is primarily based on the accuracy and F1-score of the models, which are commonly used performance metrics in classification tasks. Accuracy offers a measure of the overall correctness of the models' predictions, while the F1-score delivers a balance between the precision and recall of the model. This provides a more nuanced picture of the model's ability to correctly identify each class while minimizing false positives and negatives. The following table (Tab 4.6) illustrates the best results for both models:

Model	**Clip Length**	**Accuracy**	**F1-Score**
Transformer-Encoder Model	4-second	73.31%	74.47%
ResNet-Encoder Model	4-second	83.25%	83.43%

TABLE 4.6: The best model performance from Transformer-Encoder model and ResNet-Encoder model

4.4.1 Loss Comparison

An additional avenue for comparing the ResNet-Encoder and Transformer-Encoder models is to consider their training and validation losses over epochs. This comparison provides vital insights into the models' learning behaviors and capacity to generalize to unseen data. A model is considered to be overfitting if it shows low training loss but a comparatively high validation loss - indicating that it performs well on the training data, but less so on new, unseen data. Conversely, underfitting occurs when the model cannot achieve a low loss on the training set, indicating that it has not learned the data's underlying patterns effectively.

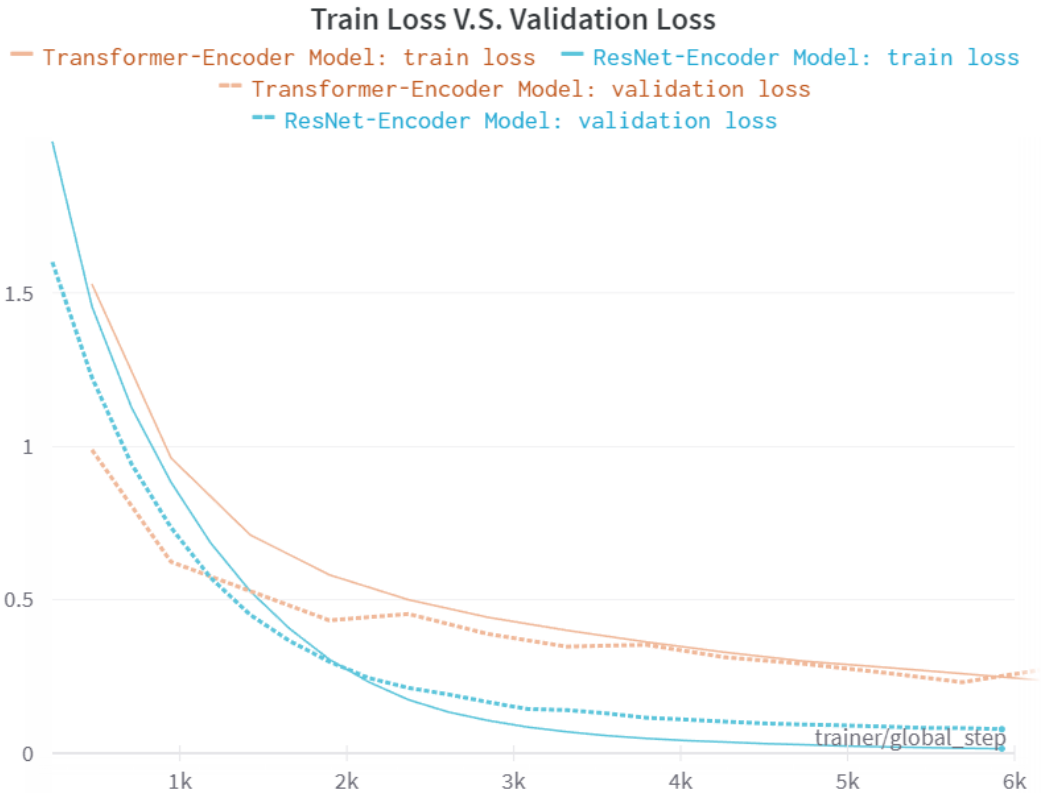

FIGURE 4.8: Comparison of training and validation loss between Transformer-Encoder model and ResNet-Encoder model

From the figure (Fig 4.8) above, we know that both the ResNet-Encoder and Transformer-Encoder models have successfully converged. At the point of convergence, the ResNet-Encoder model exhibits a training loss of 0.01476 and a validation loss of 0.07476, indicating an impressive performance with a modest gap between training and validation losses. This moderate difference suggests a good balance between bias and variance, a prerequisite for effective generalization on unseen data. In contrast, the Transformer-Encoder model shows higher loss values, which implies poorer learning compared to the ResNet-Encoder model.

These observations align with the performance metrics we derived earlier, where the ResNet-Encoder model outperformed the Transformer-Encoder model on both accuracy and F1-score. The lower loss values of the ResNet-Encoder model suggest that it has more effectively captured the underlying patterns in the training data and that this learning translates better to new data, as indicated by its superior validation loss and classification metrics.

4.4.2 Detailed Comparison

The confusion matrices (Fig 4.9 and Fig 4.10) and the table (Tab 4.7) of classification reports further expound the performance of the models across the nine classes: concrete, tactile, subway grate, manhole, brick, dirt, metal, cellar door, and grass.

For the Transformer-Encoder model, it is evident from the confusion matrix that it had some difficulties distinguishing between certain classes. For instance, the model often misclassified the manhole class as grass, and the metal class as manhole. The table of classification report reinforces this, revealing that the model's precision and recall are particularly low for the manhole and metal classes.

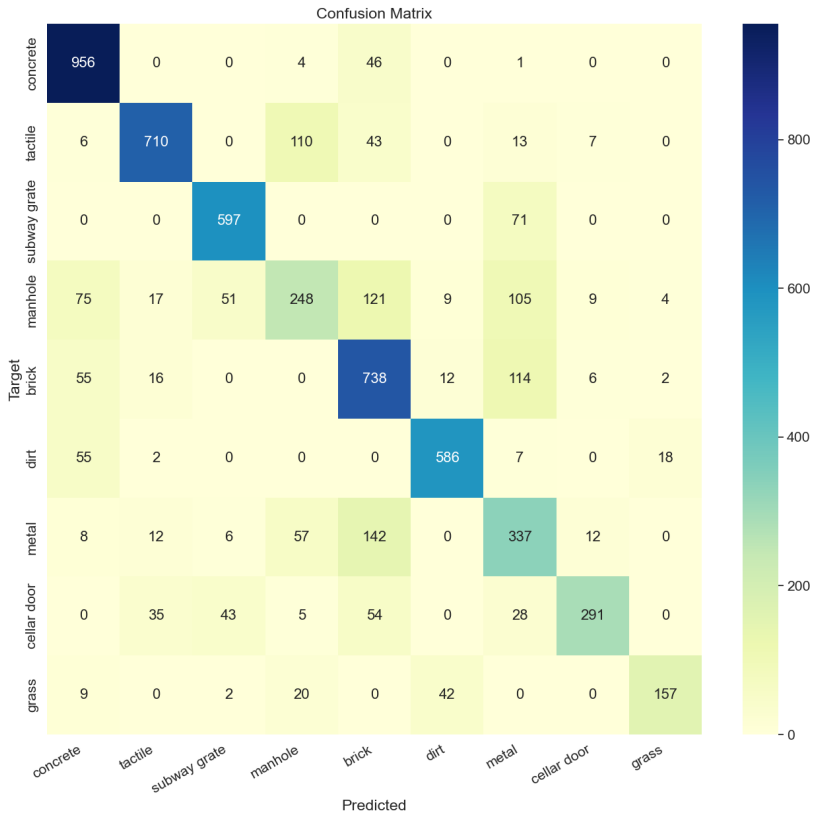

FIGURE 4.9: Confusion matrix chart of Transformer-Encoder model

Label	**ResNet-Encoder Model**		**Transformer-Encoder Model**	
	Precision	**Recall**	**Precision**	**Recall**
Concrete	91.86%	88.48%	82.13%	94.94%
Tactile	91.73%	87.29%	89.65%	79.87%
Subway Grate	85.40%	98.05%	85.41%	89.37%
Manhole	69.57%	83.72%	55.86%	38.81%
Brick	83.97%	91.62%	64.51%	78.26%
Dirt	92.50%	88.62%	90.29%	87.72%
Metal	72.90%	52.96%	49.85%	58.71%
Cellar Door	83.22%	77.19%	89.54%	63.82%
Grass	87.79%	81.30%	86.74%	68.26%

TABLE 4.7: Classification report for Transformer-Encoder model and ResNet-Encoder model

Contrastingly, the ResNet-Encoder model displayed a more balanced performance across all classes. Its confusion matrix shows a high degree of correct classifications along the diagonal, with fewer misclassifications. While some confusion occurred between the metal and manhole classes, it was noticeably less than that seen with the Transformer-Encoder

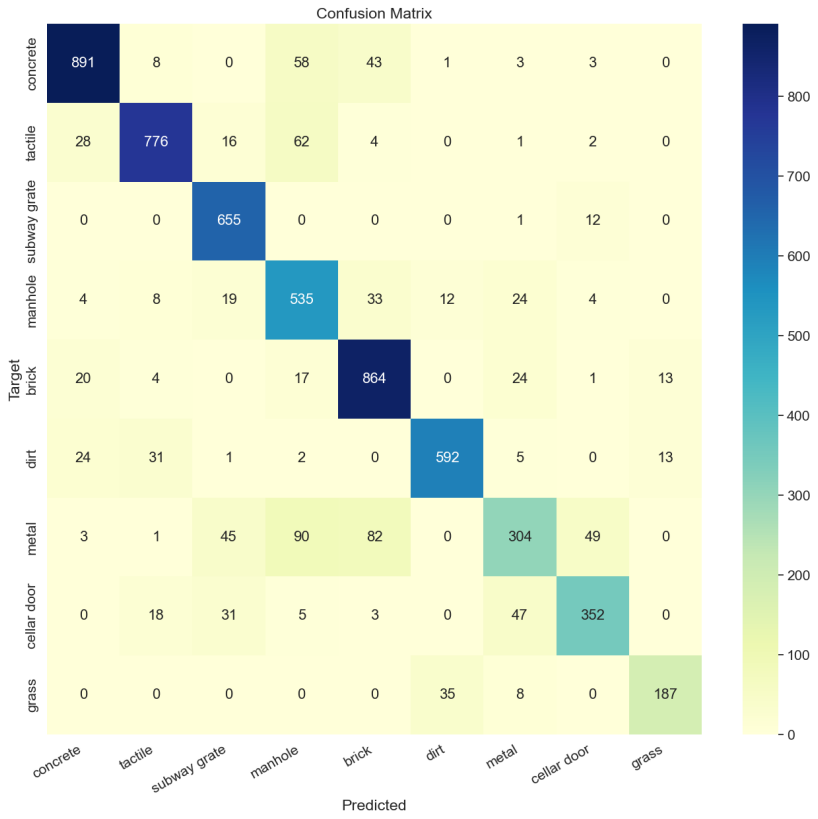

FIGURE 4.10: Confusion matrix chart of ResNet-Encoder model

model. The classification report table corroborates these findings, with relatively high precision and recall across all classes.

4.4.3 Discussion

From the results, the ResNet-Encoder model excels over the Transformer-Encoder model in terms of both accuracy and F1-score. One plausible reason could be that the ResNet-Encoder, with its pre-training phase, is better equipped to discern effective feature representations, thereby enhancing its classification accuracy. Nevertheless, the Transformer-Encoder model still demonstrates potential, despite overall lower performance. It presents a unique strength in handling shorter data clips efficiently, maintaining almost similar performance for both 2-second and 4-second clips.

Chapter 5

Conclusion and Future Work

5.1 Summary

Our innovative data collection method, which incorporated an inertial measurement unit (IMU) and microphones on the white canes, successfully led to the creation of a comprehensive, multi-modal sidewalk material (MSM) dataset. The success of this approach highlights the untapped potential of non-visual data to help individuals with blind or have low vision (BLV) navigate.

For the sidewalk material classification, we applied two distinct deep learning models: the ResNet-Encoder model and the Transformer-Encoder model. Our experiments indicated a superior performance of the ResNet-Encoder model, achieving an accuracy of 83% when trained with a 4-second-long data clip, outperforming the Transformer-Encoder model, which yielded an accuracy of 73.31%. The difference in performance can be attributed to the lack of a pre-trained model for the Transformer encoder.

5.2 Limitations and Improvements

While our findings exhibit potential, the research conducted possesses certain limitations that warrant discussion. Our ResNet-Encoder Model solely focuses on utilizing ResNet-50 as the encoder, hence, restricting our scope of understanding the performance of other state-of-the-art models, such as CLIP (Contrastive Language–Image Pre-training) [35] and ViT (Vision Transformer) [14]. This limited focus presents an opportunity for potential enhancement of the model's robustness and efficiency. Furthermore, our Transformer-Encoder Model encounters the issue where there is no pre-trained model, which may lead to sub-optimal performance. A reasonable solution to address this limitation is to implement a large public audio dataset for training, which will help to develop an optimized pre-trained transformer model. Therefore, we expect that this could improve the performance and functionality of Transformer-Encoder model.

However, these limitations should not be viewed as setbacks, but rather as opportunities to guide future research. By exploring and implementing advanced machine learning models, we may discover the unexplored potential to improve the performance of the system.

5.3 Future Work

5.3.1 Performance Evaluation on Unlabeled Continuous Data

To further refine and evaluate our classification model, we propose a plan for future work that includes the application of unlabeled continuous data for real-world testing. This step represents an important advancement in our research, providing a practical,

user-oriented approach that allows us to evaluate the performance of the model under real-world, non-laboratory conditions.

Our proposed approach is to create a testbed, which comprises an area of 4X4 streets. We will first collect sidewalk material information from these 4X4 streets, and manually label them. In subsequence, we will invite and guide BLV people to collect continuous and unlabeled sidewalk material data using the proposed data collection framework. With this evaluation, we can realistically test the validity of our models and the robustness of our data collection framework.

We believe this future research direction represents a significant step towards a more comprehensive evaluation of our model. By integrating real-world conditions and user feedback into our assessment process, we anticipate gaining deeper insights into the model's performance, its potential areas of improvement, and its overall value to BLV people. Ultimately, we hope that the insights gleaned from this future work will aid us in refining our classification model and data collection framework, propelling our research closer to a practical solution for BLV individuals navigating urban environments.

5.3.2 Potential Applications

Within the scope of our project, we introduced an AI-based data collection framework designed to enable BLV people to autonomously collect information about sidewalk materials during their journeys. The main purpose of this tool, combined with the active participation of BLV individuals, is to generate a comprehensive, large-scale accessible map layer. The layer aims to encompass all necessary landmarks, such as different sidewalk materials, as well as other important environmental data.

One of the potential applications of this accessible map layer is the development of a novel assistive navigation application. This application could revolutionize how BLV people navigate and interact with their surroundings. Users could strategically plan their travel paths based on the availability and extent of accessible infrastructure, such as tactile pavements. Accordingly, the application would suggest the safest and most convenient routes for users, enhancing their mobility and independence. Additionally, the application could possess the capability to calibrate direction and orientation, leveraging the precise landmarks provided by our accessible map layer. This would allow for an increase in navigational accuracy and dependability, further fostering user trust and usability.

From an extended perspective, our framework's potential applications could also cross the bounds of accessibility and integrate with city planning and maintenance. For instance, city administrators could utilize the detailed data provided by our system to make informed decisions about where to implement accessible infrastructure or maintain existing sidewalk materials. This could not only enhance the city's overall accessibility but also promote a more inclusive living environment.

Bibliography

[1] Ariadne gps. https://www.ariadnegps.eu. Accessed: 2023-03-30.

[2] Blindsquare. http://www.blindsquare.com/. Accessed: 2023-03-30.

[3] Federal highway administration u.s. department of transportation accessible sidewalks and street crossings. https://highways.dot.gov/. Accessed: 2023-03-30.

[4] Google map. https://www.google.com/maps. Accessed: 2023-03-30.

[5] Nearby. https://www.aph.org/nearby-explorer/. Accessed: 2023-03-30.

[6] Nyc mayor's office for people with disabilities (mopd). https://www1.nyc.gov/site/mopd/index.page. Accessed: 2023-03-30.

[7] Seeing. http://www.senderogroup.com/products/seeingeyegps/index.html. Accessed: 2023-03-30.

[8] Transit street design guide. https://nacto.org/publication/transit-streetdesign-guide/transit-lanes-transitways/lane-elements/pavement-materials. Accessed: 2023-03-30.

[9] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, and S. Ridella. The'k'in k-fold cross validation. In *ESANN*, pages 441–446, 2012.

[10] Y. Chen, Y. Wang, Z. Dong, J. Su, Z. Han, D. Zhou, Y. Zhao, and Y. Bao. 2-d regional short-term wind speed forecast based on cnn-lstm deep learning model. *Energy Conversion and Management*, 244:114451, 2021.

[11] N. Donaldson. *Visually impaired individuals' perspectives on obtaining and maintaining employment*. PhD thesis, Walden University, 2017.

[12] L. Dong, S. Xu, and B. Xu. Speech-transformer: a no-recurrence sequence-to-sequence

model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5884–5888. IEEE, 2018.

[13] M. Dong. Convolutional neural network achieves human-level accuracy in music genre classification. *arXiv preprint arXiv:1802.09697*, 2018.

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

[15] Y. Gong, Y.-A. Chung, and J. Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[16] A. Guzhov, F. Raue, J. Hees, and A. Dengel. Esresnet: Environmental sound classification based on visual domain models. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4933–4940. IEEE, 2021.

[17] K. Hara, S. Azenkot, M. Campbell, C. L. Bennett, V. Le, S. Pannella, R. Moore, K. Minckler, R. H. Ng, and J. E. Froehlich. Improving public transit accessibility for blind riders by crowdsourcing bus stop landmark locations with google street view: An extended analysis. *ACM Transactions on Accessible Computing (TACCESS)*, 6(2):1–23, 2015.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

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

[20] X. He, Y. Chen, and Z. Lin. Spatial-spectral transformer for hyperspectral image classification. *Remote Sensing*, 13(3):498, 2021.

[21] J. F. Herman, S. P. Chatman, and S. F. Roth. Cognitive mapping in blind people: Acquisition of spatial relationships in a large-scale environment. *Journal of Visual Impairment & Blindness*, 77(4):161–166, 1983.

[22] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal,

D. Platt, R. A. Saurous, B. Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[23] Z. Hu, X. Chen, H. W. F. Yeung, Y. Y. Chung, and Z. Chen. Texture-enhanced light field super-resolution with spatio-angular decomposition kernels. *IEEE Transactions on Instrumentation and Measurement*, 71:1–16, 2022.

[24] J. Lee, I. Lee, and J. Kang. Self-attention graph pooling. In *International conference on machine learning*, pages 3734–3743. PMLR, 2019.

[25] J. Liu, H. Tang, W. Seiple, and Z. Zhu. Annotating storefront accessibility data using crowdsourcing. 2022.

[26] X. Liu, H. Lu, J. Yuan, and X. Li. Cat: Causal audio transformer for audio classification. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[27] Y. Liu, J. Cai, and G. Tan. Multi-level circulation pattern classification based on the transfer learning cnn network. *Atmosphere*, 13(11):1861, 2022.

[28] A. Maccagno, A. Mastropietro, U. Mazziotta, M. Scarpiniti, Y.-C. Lee, and A. Uncini. A cnn approach for audio classification in construction sites. *Progresses in Artificial Intelligence and Neural Systems*, pages 371–381, 2021.

[29] G. Marzano, J. Lizut, and L. O. Siguencia. Crowdsourcing solutions for supporting urban mobility. *Procedia Computer Science*, 149:542–547, 2019.

[30] A. Meghanani, C. Anoop, and A. Ramakrishnan. An exploration of log-mel spectrogram and mfcc features for alzheimer's dementia recognition from spontaneous speech. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 670–677. IEEE, 2021.

[31] W. H. Organization. Global data on visual impairment. https://www.who.int/blindness/publications/globaldata/en. Accessed: 2023-03-30.

[32] K. O'Shea and R. Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[33] K. Palanisamy, D. Singhania, and A. Yao. Rethinking cnn models for audio classification. *arXiv preprint arXiv:2007.11154*, 2020.

[34] M. Portnoff. Time-frequency representation of digital signals and systems based on short-time fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(1):55–69, 1980.

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[36] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV*, pages 525–542. Springer, 2016.

[37] M. Saha, M. Saugstad, H. T. Maddali, A. Zeng, R. Holland, S. Bower, A. Dash, S. Chen, A. Li, K. Hara, et al. Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[38] E. Shi, Y. Wang, H. Zhang, L. Du, S. Han, D. Zhang, and H. Sun. Towards efficient fine-tuning of pre-trained code models: An experimental study and beyond. *arXiv preprint arXiv:2304.05216*, 2023.

[39] A. Torralba, B. C. Russell, and J. Yuen. Labelme: Online image annotation and applications. *Proceedings of the IEEE*, 98(8):1467–1484, 2010.

[40] A. Trémeau, S. Xu, and D. Muselet. Deep learning for material recognition: most recent advances and open challenges. *arXiv preprint arXiv:2012.07495*, 2020.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[42] L. Vrysis, I. Thoidis, C. Dimoulas, and G. Papanikolaou. Experimenting with 1d cnn architectures for generic audio classification. In *Audio Engineering Society Convention 148*. Audio Engineering Society, 2020.

[43] J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang. Dimensional sentiment analysis using a

regional cnn-lstm model. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 225–230, 2016.

[44] T. Watanabe, H. Takahashi, G. Sato, Y. Iwasawa, Y. Matsuo, and I. E. Yairi. Wheelchair behavior recognition for visualizing sidewalk accessibility by deep neural networks. In *Deep Learning for Human Activity Recognition: Second International Workshop, DL-HAR 2020, Held in Conjunction with IJCAI-PRICAI 2020, Kyoto, Japan, January 8, 2021, Proceedings 2*, pages 16–29. Springer, 2021.

[45] p. j. wolfe and j. howarth. nonuniform sampling theory in audio signal processing. *journal of the audio engineering society*, may 2004.

[46] H. Zheng, L. Fang, M. Ji, M. Strese, Y. Özer, and E. Steinbach. Deep learning for surface material classification using haptic and visual information. *IEEE Transactions on Multimedia*, 18(12):2407–2416, 2016.

[47] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[48] Y. Zhuang, Y. Chen, and J. Zheng. Music genre classification with transformer classifier. In *Proceedings of the 2020 4th International Conference on Digital Signal Processing*, pages 155–159, 2020.