SMDAF: A Scalable Sidewalk Material Data Acquisition Framework with Bidirectional Cross-Modal Knowledge Distillation

Jiawei Liu[1], Wayne Lam[3], Zhigang Zhu[1,3], Hao Tang[1,2]
[1] The Graduate Center - CUNY, [2] Borough of Manhattan Community College - CUNY
[3] The City College of New York - CUNY

Abstract

Ensuring safe and independent navigation poses considerable difficulties for individuals who are blind or have low vision (BLV), as it requires detailed knowledge of their immediate environment. Our research highlights the critical need for accessible data on sidewalk materials and objects, which is currently lacking in existing map services. To bridge this gap, we present the Sidewalk Material Data Acquisition Framework (SMDAF), designed for large-scale data collection. This framework includes (1) a lightweight data collection system embedded in a white cane, which captures audio data through the interaction of the cane tip with the sidewalk surface, and a mobile app that facilitates data storage and management, resulting in a novel multimodal dataset comprising both image and audio data; and (2) a unique Cross-Modal Knowledge Distillation (CMKD) technique for an enhanced audio material classifier. Our CMKD approach employs an image-based model as the teacher to improve the audio model, incorporating an Enhanced Bidirectional learning method with an intuitive filtering technique: Bidirectional Correct Sample Filtering (BCSF). BCSF filters correct samples to prevent the distillation of incorrect knowledge, addressing the issue of inaccurate cross-modal learning. This novel approach has resulted in a 1.84% improvement in Macro Accuracy, achieving an overall accuracy of 87.62%, surpassing all state-of-the-art KD and CMKD methods. This study underscores the efficacy of SMDAF and provides a practical CMKD technique for future cross-modal learning tasks. Code and dataset are available (https://github.com/FgSurewin/SMDAF-CMKD).

1. Introduction

Navigating urban environments presents significant challenges for blind and low vision (BLV) individuals. As of the most recent data, approximately 1 million adults in the US are classified as blind, with over 6 million Americans experiencing some form of vision loss [11].

The task of navigating city streets requires BLV individuals to continuously query their immediate environment for spatial information [18]. Unlike sighted individuals who rely on visual landmarks, BLV individuals depend heavily on tactile and auditory cues to orient themselves and avoid obstacles [18]. Previous research [5, 12, 40] mostly focuses on 2D and/or 3D estimation of objects such as curbs and ramps. Understanding of sidewalk materials, which serve as critical tactile landmarks that help guide navigation, is rarely studied. Variations in surface textures, such as the difference among various sidewalk materials (e.g., concrete, brick, etc.) and sidewalk landmarks (e.g., manhole cover, and tactile pavement, etc.), can signal important changes in the walking environment and aiding in orientation and movement. Hence these variations should be considered important urban accessibility data.

Despite the importance of detailed environmental information for BLV individuals, current map applications such as Google Maps [2] and Apple Maps [1] fall short in providing the necessary data to support safe and independent navigation. These applications are primarily designed to offer the shortest or fastest routes, often neglecting the accessibility and safety considerations essential for BLV users. Although some studies [5, 40] have proposed tools to collect information about sidewalks using deep learning, their focus is on urban planning rather than addressing the specific needs of BLV individuals. Government agencies also manage information about public facilities, including records of tactile pavement [9]. However, this data collection is limited to certain types of infrastructure and often suffers from significant delays in updates. Consequently, the existing information is neither comprehensive nor timely enough to be effectively utilized for BLV navigation.

This gap in data collection raises a critical question: *how can we **effectively** collect **large-scale** sidewalk material data?* Addressing this issue helps BLV individuals navigate the urban environment safely and independently. To bridge this gap, we present the Sidewalk Material Data Acquisition Framework (SMDAF), a novel solution for large-scale

street-level sidewalk material data collection (Figure 1).

Large-scale Data Collection. To ensure a large-scale data collection approach, we develop a lightweight crowdsourcing data collection system integrated into a white cane, an essential tool for BLV individuals. This system captures audio information when the cane tip interacts with various sidewalk materials. This allows BLV individuals to seamlessly collect data as they travel, making the process unobtrusive and natural. Our approach leverages the travel patterns of BLV individuals who frequently visit commercial districts and other populated areas. This targeted data collection ensures that the information gathered is relevant to in-demand navigation scenarios for BLV individuals.

To enhance preliminary data collection, an iOS mobile app is developed to record real-time video and GPS information, which is used to train the material classifier and generate a city-scale accessibility map layer (Figure 1, Map in Part A). Collecting visual data requires mounting or hand-holding video equipment, which poses an increased risk of theft for BLV individuals [6]. In this project, image data is used only during the initial model training phase (Figure 1, Part B), not during the inference phase. Hence, BLV individuals are not required to capture visual data while traveling, making the data collection process feasible and scalable to many BLV individuals. In this study, an initial dataset is collected and used to train our audio material classifier to identify and label sidewalk material data. Additionally, the design of our framework naturally supports "*training in the loop*", akin to [21]. As future work, when more and more data is collected by BLV individuals during their travels and automatically labeled using the proposed audio material classifier, the classifier can be further enhanced through iterative training with new data. Such an iterative approach enables scalable and sustainable crowdsourced data collection.

Effective Data Collection. To ensure the effectiveness of our data collection, we develop an audio material classifier that effectively distinguishes between different materials with high precision.

Previous research has shown the feasibility of using Convolutional Neural Network (CNN)-based and Transformer-based models, pre-trained on ImageNet, to extract audio features [14, 19, 29, 37]. To enhance the performance of the proposed audio material classifier, we apply Cross-Modal Knowledge Distillation (CMKD) [15]. This technique allows the transfer of visual information to the audio model, leveraging the superior performance of image-based models. Although the modality gap [24, 38, 44] is a known challenge, our novel approach mitigates this issue by fine-tuning both the teacher (image model) and the student (audio model) simultaneously. We introduce an Enhanced Bidirectional CMKD approach, akin to deep mutual learning (DML) [47].

A recent study [24] reveals knowledge misalignment and introduces an On-the-Fly Selection Distillation (OFSD) strategy to address it by filtering misaligned samples using the Kendall Rank Correlation Coefficient (KRC). To further tackle this issue, we propose a filtering method called Bidirectional Correct Sample Filtering (BCSF). This technique filters correct samples from the teacher when the student learns and vice versa, preventing the distillation of incorrect knowledge. BCSF has proven to be a simple yet effective method, resulting in a 1.84% improvement in macro accuracy and achieving an overall accuracy of 87.62%, surpassing all state-of-the-art KD and CMKD methods. Through these components, our SMDAF provides an effective solution for large-scale street-level data collection on sidewalk materials, ultimately enhancing the safety and independence of BLV individuals. Our main contributions can be summarized as follows:

- **A Novel Scalable Sidewalk Material Data Acquisition Framework (SMDAF).** We introduce a lightweight data collection system to gather audio data on various sidewalk materials, along with a mobile app that records GPS and visual data. Utilizing SMDAF, we generate and publish an initial multimodal sidewalk material dataset composed of both image and audio data for future study by other researchers.

- **Advanced Audio Material Classification with a novel Enhanced Bidirectional CMKD approach.** We develop a specialized audio material classifier for identifying sidewalk materials, leveraging CMKD training techniques with Bidirectional Correct Sample Filtering (BCSF). Our Enhanced Bidirectional CMKD method utilizing the BCSF technique addresses the main challenges in the orignal CMKD, namely the modality gap and knowledge misalignment. This study serves as a practical guideline for general CMKD tasks.

2. Related Work

2.1. Crowdsourcing Urban Data Collection

Local and state governments often collect data on street accessibility [4]. With ubiquitous Internet and mobile technologies, data can be collected more efficiently and economically using crowdsourcing methods [30]. Several studies use crowdsourcing and Google Street View (GSV) to allow people to identify bus stops, curb ramps and storefront accessibility data remotely [16, 28, 32]. Another novel setup mounts a tri-axial accelerometer under a wheelchair seat [39] to infer sidewalk accessibility features such as slope and curb presence from the behavior of the wheelchair. Analogously, we mount a microphone onto a white cane

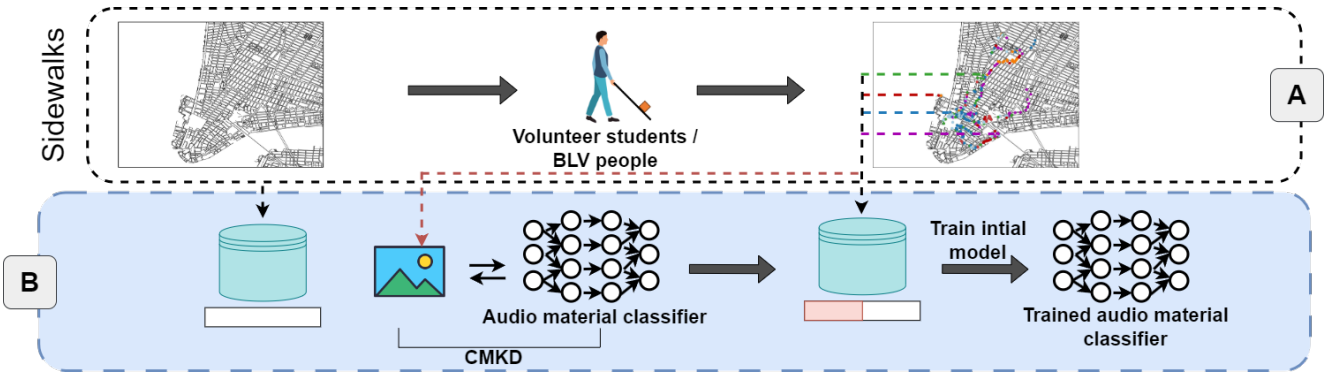

Figure 1. Overview of the proposed SMDAF. A) The diagram illustrates the initial data collection phase, where both sighted volunteers and BLV individuals gather data across diverse environments. B) This comprehensive dataset serves as the foundation for training the audio material classifier, which is enhanced by utilizing CMKD.

and collect sidewalk material data while the BLV people walk with a white cane.

Beyond data collection, the analysis of sidewalk material is crucial. Prior studies in the environmental research community [13, 42] show the physical properties of sidewalks—such as material type—are crucial for accessibility. Certain materials, such as uneven surfaces or soft materials like rubber tiles, directly impact mobility for individuals with disabilities. We build on the findings from domain experts by analyzing sidewalk landmarks, such as tactile pavements, which are particularly useful for BLV people in navigation.

2.2. Material Recognition and Audio Classification

Material recognition in the past usually relies on visual cues from image data [7, 35]. One study [36] achieves results by focusing on material image datasets, contextual influences, and unique descriptors of material appearance. In another study, sidewalk materials are classified utilizing street-level images from GSV [21]. In contrast, the proposed framework utilizes audio data for material recognition as there have been studies illustrating the efficacy of using object-to-surface acoustic signals for event recognition tasks [25].

In terms of audio classification in general, two primary architectures are commonly used:

CNN Architectures. Convolutional neural network (CNN) architectures are successful in audio classification tasks when classifying spectrograms derived from transformed raw audio data [29]. In [19], the study leverages transfer learning by employing a pre-trained Resnet50 model for audio classification.

Transformer Architectures. Audio Spectrogram Transformer (AST) and other transformer based architectures have recently achieved state-of-the-art results in audio classification tasks [8, 14]. Like the CNN model architectures, transformer architectures evaluate audio data transformed into spectrograms. These architectures are based on the Data Efficient Image Transformer (DeiT) [34] architecture, which itself is based on the Vision Transformer (ViT) [10] architecture, while allowing variable size spectrograms to be evaluated [14] as opposed to a standard DeiT model which evaluates images of a fixed size [34].

2.3. Knowledge Distillation

Knowledge distillation (KD) techniques transfer knowledge from a pre-trained teacher model to a student model by reducing the difference between the logits of the teacher model and student model [20], the extracted features of the models [45], or the models' output relations of data samples [31]. This distillation is enforced by applying an additional loss, $\mathcal{L}_{dist}$, to the total loss function, $\mathcal{L}_{total}$.

Unimodal Knowledge Distillation. In traditional KD, the teacher and student models are trained on the same modality of data. This methodology is often employed so that a large cumbersome teacher can transfer knowledge to a smaller more efficient student [20].

Cross-Modal Knowledge Distillation. Another form of KD utilizes a teacher model pre-trained on one modality in order to teach a student model in another modality [15, 44]. While it is often utilized in situations where a higher accuracy teacher model teaches a lower accuracy student model [15], it is also sometimes favorable to train the models bidirectionally [24, 26, 47].

3. Sidewalk Material Data Acquisition

3.1. Design of the SMDAF

The Sidewalk Material Data Acquisition Framework (SMDAF) is designed to facilitate the collection and continuous improvement of labeling sidewalk material data, which can be utilized in developing future navigation applications for BLV individuals. The framework operates in two key phases (Figure 1):

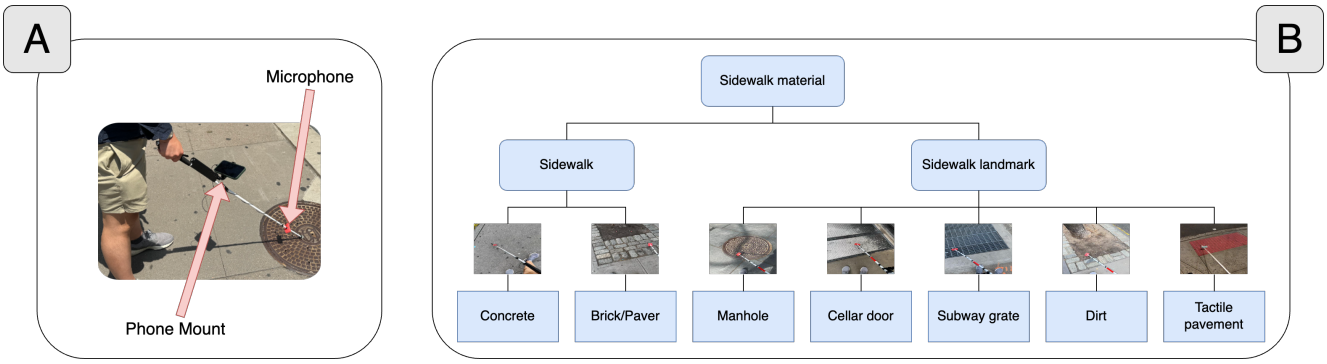

Figure 2. **A)** The modified white cane for multimodal sidewalk material initial data collection. **B)** The categories of the data inventory. The categories are divided into 2 *sidewalk* categories and 5 *sidewalk landmark* categories.

- **Initial Data Acquisition.** The initial dataset is collected by both sighted volunteers and BLV individuals. To support this process, a mobile application is developed to facilitate multimodal data collection (further details in Section 3.3).

- **Data Inventory Establishment and Audio Material Classifier Training.** The collected data is first categorized into *sidewalk* and *sidewalk landmarks*, establishing a robust sidewalk material data inventory (further details in Section 3.4). A third-party annotation application is utilized to label the data manually. Following the establishment of the inventory, an initial dataset is generated and used by an audio material classifier to train using the CMKD technique. The classifier enables the automated labeling of new sidewalk material data as it is collected in the future.

3.2. Data Acquisition Equipment

The data acquisition system is designed to collect acoustic feedback and video data when the cane interacts with various sidewalk surfaces using a modified white cane (Figure 2 A). While various canes are used by BLV individuals, we select a cane obtained from the National Federation of the Blind with a metal tip.

As a white cane interacts with different materials, distinct acoustic signals are produced by the cane tip. To capture those differences, a wired microphone is positioned near the cane tip and clipped to a foam ring to maximize the clarity of recorded sounds while minimizing ambient noise and cane vibration noise.

Additionally, the white cane is equipped with a mount to securely hold an iPhone, enabling the simultaneous recording of video data. This video data, captured using an iOS mobile app, complements the acoustic data and is particularly valuable during the initial training phase of the classifier model. The combined use of audio and visual data enhances the robustness of the dataset, facilitating more accurate and comprehensive sidewalk material classification.

3.3. Mobile App for Data Acquisition

To facilitate the collection of data, an iOS mobile app is developed. The application is designed to capture video and audio data simultaneously, leveraging the microphone attached to the cane for auditory feedback. Additionally, the geo-location of the sidewalk material data is recorded by leveraging the built-in GPS function on smartphones.

The video data recorded by the iPhone mounted on the cane provides complementary visual information for the corresponding sidewalk material audio data. This multimodal data is crucial for the initial training of the classifier model. In addition, the visual information serves as a reference for annotations, aiding in the continuous and accurate labeling of sidewalk materials. By integrating visual, audio, and location data, the mobile app ensures a comprehensive and robust dataset, essential for developing reliable navigation aids for BLV individuals.

3.4. Data Inventory and Management

To assemble the sidewalk material data inventory, a large-scale data collection effort is initiated involving student volunteers. The data is gathered from diverse environmental conditions in New York City, from commercial districts to residential neighborhoods and during medium traffic business hours to high traffic rush hours. The data collection is from two different modes: *static* and *continuous*.

- **Static data collection.** 27 sighted volunteers collect a substantial amount of sidewalk material data by swiping a cane on the sidewalk surface while standing. Each data record contains over 30 seconds of a single sidewalk material. The mobile app allows users to annotate the sidewalk material data in real-time.

- **Continuous data collection.** 6 sighted volunteers and 1 BLV individual collect sidewalk material data while walking along longer sections of sidewalk to better emulate the real-world conditions experienced by BLV

users. In this mode, each data record includes multiple categories of sidewalk materials. The video data collected in this manner is manually annotated.

To ensure accuracy and relevance, the collected multimodal sidewalk material data is divided into 7 categories modified from the NYC Street Design Manual [3] and through common sidewalk material encounters during data collection. The *sidewalk* and *sidewalk landmark* categories are illustrated in Figure 2 Part B, where 50 to 60 minutes of data is collected for each category.

- **Sidewalk.** The sidewalk materials we categorize as *sidewalk* are directly derived from the NYC Street Design Manual [3] and categorized as concrete or brick.
- **Sidewalk Landmark.** The sidewalk materials that are categorized as *sidewalk landmark* are commonly encountered materials during data acquisition that are distinct from the *sidewalk*. The categories include manhole covers, cellar doors, subway grates, dirt patches, and tactile pavement.

4. Audio Material Classification with CMKD

4.1. Baseline Audio Material Classification Models

The identification of potential teacher and student architectures is the initial task for training an audio material classifier. We explore both CNN-based and Transformer-based models, focusing specifically on DeiT models [34], which have shown strong performance in audio classification tasks [8, 14, 19, 29, 37]. The goal is a student model that is computationally efficient and suitable for future deployment on mobile devices. This consideration is to ensure that the model can be used in real-world scenarios by BLV individuals without excessive resource demands.

We evaluate several popular lightweight models to serve as the student model for audio material classification. Table 1 presents the results of our evaluation, including the number of parameters (Params), floating-point operations per second (FLOPS), Macro Accuracy (Macro Acc.), Macro F1-Score (Macro F1.) and Micro Accuracy (Micro Acc.).

Model	Macro Acc.	Macro F1.	Micro Acc.	Params (M)	FLOPS (M)
AST-Tiny224 [14]	85.78%	85.46%	84.95%	5.79	881.2
AST-Base224 [14]	86.17%	85.78%	85.31%	86.86	13,971.7
ResNet18 [17]	84.75%	84.39%	84%	11.18	833.5
MobileNetV3 [22]	80.93%	80.90%	80.38%	1.53	29.3
EfficientNetV2 [33]	85.62%	85.22%	85.54%	23.52	1,871.9

Table 1. Performance and computational efficiency comparison of baseline audio material classification models.

As shown in Table 1, the models vary significantly in terms of parameters and computational requirements. DeiT-based models, such as AST (Audio Spectrogram Transformer) [14], demonstrate higher performance with a Macro Accuracy of 86.17% and a Macro F1-Score of 85.78%, but at the cost of increased computational complexity. Among the CNN-based models, ResNet18 and MobileNetV3 are considered due to their relatively low computational demands. ResNet18 provides a good balance between performance and efficiency, with a Macro Accuracy of 84.75%, a Macro F1-Score of 84.39% and relatively low FLOPS(833.5M), making it a strong candidate for the student model. MobileNetV3, while the most lightweight, shows a lower Macro Accuracy of 80.93%.

Considering the trade-offs between computational efficiency and accuracy, we select ResNet18 and AST-Tiny224 as our baseline models for audio material classification. These baseline models provide a solid benchmark for the subsequent experiments to assess the performance gains achieved by the proposed Enhanced Bidirectional CMKD.

Student (Audio)	Teacher (Image)	Macro Acc.	Macro F1.	Micro Acc.	Change
Resnet18 [17]	-	84.75%	84.39%	84%	-
AST-Tiny224 [14]	-	85.78%	85.46%	84.95%	-
Resnet18	Resnet50	83.72%	83.46%	82.65%	-1.03%
AST-Tiny224	Resnet50	84.68%	84.42%	83.82%	-1.1%
AST-Tiny224	DeiT-Tiny224	85.44%	85.07%	84.64%	**-0.34%**
AST-Tiny224	DeiT-Base224	85.49%	85.21%	84.68%	**-0.29%**

Table 2. Performance comparison of various teacher-student architectures under knowledge distillation (KD) using KL-Divergence loss. The "Change" column indicates the percentage difference in Macro Accuracy, which applies across all following tables.

4.2. Preliminary Analysis of CMKD

Traditional knowledge distillation [20], Figure 3.4(A), is a technique where a larger, pre-trained model (teacher) is used to improve the performance of a smaller model (student). The goal is to transfer the knowledge learned by the teacher to the student model, which is typically more lightweight and suitable for deployment in resource-constrained environments. The distillation process involves two types of losses: hard loss and soft loss. The hard loss is the standard classification loss calculated between the student's predictions and the true labels. The soft loss is calculated between the teacher's and student's output probabilities (logits) using Kullback-Leibler (KL) divergence [20]. The overall loss function is a weighted sum of these two losses.

In the context of CMKD, the main difference is that the teacher and student models operate on data from different modalities. This introduces additional challenges due to the inherent differences between modalities - **modality gap**. Directly transferring knowledge across modalities can be ineffective due to this gap. Previous studies have introduced the Modality Focusing Hypothesis (MFH), which posits that Modality-General Decisive Features (MGDF) are crucial for successful knowledge transfer during distillation [44]. However, directly identifying MGDF is impractical and alternative approaches to maximize the transfer of

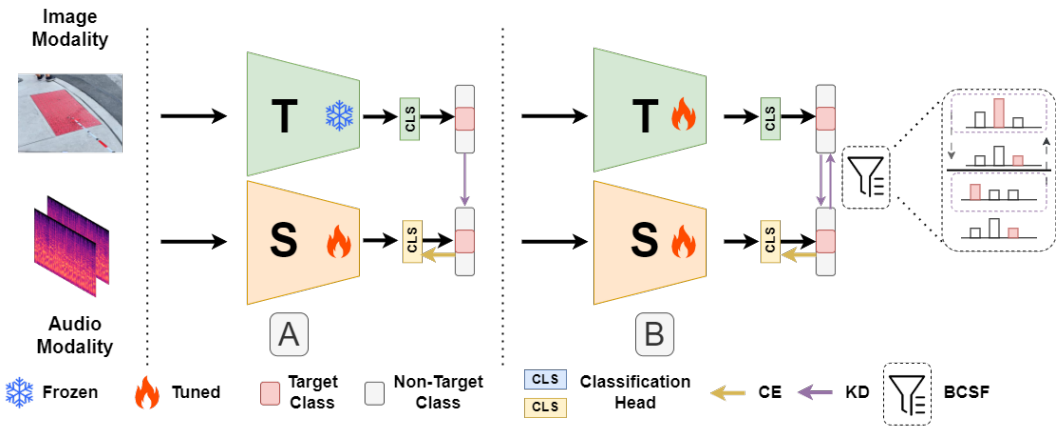

Figure 3. of CMKD. (A) **CMKD with Traditional KD [20].** Traditional KD is applied where the pre-trained teacher model is frozen. (B) **Proposed Enhanced Bidirectional CMKD.** Both pre-trained teacher and untrained student are tunable, leading to mutual learning from each other through Bidirectional KD and BCSF.

useful knowledge across different modalities are necessary. One strategy is to carefully select teacher-student architectures that can bridge the modality gap effectively.

Teacher Model (Image)	Macro Acc.	Macro F1.	Micro Acc.
Resnet50 [17]	95.28%	95.36%	95.35%
DeiT-Tiny224 [34]	95.86%	95.96%	95.89%
DeiT-Base224 [34]	95.96%	96.04%	96.02%

Table 3. Performance comparison of CNN-based and DeiT-based teacher models for image modality.

To investigate the impact of different teacher-student architectures on CMKD, we experiment with various combinations of models using KL-divergence for unidirectional KD loss. Table 2 shows the performance of different architectures and the results indicate that all the student-teacher combinations did not improve upon the baseline models, highlighting the modality gap in CMKD, although the teacher model (Table 3) is stronger than the student model. Specifically, AST-Tiny224 paired with ResNet50 shows a larger decrease of 1.1% with respect to Macro Accuracy. However, the combination of AST-Tiny224 with DeiT-Base224 shows the smallest decrease in Macro Accuracy at 0.29%, indicating a relatively better knowledge transfer. This modest decrease suggests that DeiT-based architectures may be more effective for CMKD with our data, due to their ability to capture more generalizable features across modalities. Despite the modality gap, these preliminary results guide our decision to adopt AST-to-DeiT structures for further experiments.

4.3. Enhanced Bidirectional CMKD

As the preliminary analysis shown above, traditional KD methods face significant challenges in the context of cross-modal knowledge transfer due to the disparity in the data modalities utilized by the teacher and student models. To address these challenges and enhance the performance of CMKD, we propose an Enhanced Bidirectional CMKD approach. This approach consists of two main modules: Bidirectional KD and Bidirectional Correct Sample Filtering (BCSF), Figure 3.4 Part B.

4.3.1 Bidirectional KD

Cross-modal KD presents challenges because the teacher and student models learn from different modalities, leading them to rely on distinct feature representations for classification. In order to alleviate this issue, bidirectional KD is employed, allowing both models to adjust to each other, enhancing mutual learning [24,26,47]. The concept of bidirectional KD stems from deep mutual learning (DML) [47], where both the teacher and student models learn collaboratively through a composite distillation loss $\mathcal{L}_{dist}$ consisting of two soft losses:

$$\mathcal{L}_{dist} = \tau^2 \mathcal{D}(f_s, f_t) + \tau^2 \mathcal{D}(f_t, f_s) \quad (1)$$

where f_s and f_t are student logits and teacher logits respectively, $\mathcal{D}$ is selected distillation technique (e.g. KL-divergence), and τ is the settable *temperature* hyperparameter and squared as suggested by [20]. Pioneering work like SHAKE demonstrates the feasibility of this approach by using a proxy teacher to learn from the student [26]. The advantage of the proxy teacher structure is lightweight and less computationally expensive compared to the original teacher, but in our case, it doesn't quite reflect the same performance as the original teacher (Table 5). In this study, we employ bidirectional KD between a pre-trained teacher and an untrained student model. During training, the student is initialized from scratch and learns from the pre-trained teacher, while the teacher model is simultaneously fine-tuned to adapt to the student's modality (Figure 3.4, Part B).

4.3.2 Bidirectional Correct Sample Filtering

Bidirectional KD alone is insufficient to address the challenges in CMKD. One significant problem is **knowledge misalignment**, where the knowledge learned by the teacher does not align well with the student's learning needs. Hence, a practical approach for filtering non-distillable samples is demanded. In this context, several forms of filtering in determining whether distillation soft loss should be applied to an individual data sample [24, 43]. The SOTA CMKD method, C2KD [24], addresses this issue by introducing On-the-Fly Selection Distillation (OFSD) utilizing the Kendall Rank Correlation (KRC) [24] for filtering.

Formulation of Bidirectional Correct Sample Filtering (BCSF). While OFSD and similar methods focus on filtering based on logit alignment, we propose filtering out samples based on the condition of correctness. During BCSF, distillation loss from teacher-to-student is only applied when the teacher is correct in classification and distillation loss from student-to-teacher is only applied when the student is correct in classification:

$$\eta_m = \begin{cases} 1 & \text{if } \hat{y}_m = y \\ 0 & \text{otherwise} \end{cases} \quad \text{for } m \in \{t, s\} \tag{2}$$

In Equation 2, the filter η_t is applied to teacher-to-student soft loss and the filter η_s is applied to student-to-teacher soft loss; $\hat{y}_t$ being the teacher prediction, $\hat{y}_s$ being the student prediction, and y being the ground-truth label. Derived from Equation 1, the resultant distillation loss is:

$$\mathcal{L}_{dist} = \eta_t \tau^2 \mathcal{D}(f_s, f_t) + \eta_s \tau^2 \mathcal{D}(f_t, f_s) \tag{3}$$

The total loss, $\mathcal{L}_{total}$, is the sum of the cross-entropy hard loss, $\mathcal{L}_{CE}$, and distillation loss $\mathcal{L}_{dist}$.

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{CE} + \beta \mathcal{L}_{dist} \tag{4}$$

where, α and β are settable hyperparamters. We set α, β as 1 and temperature, τ, as 2 following C2KD [24].

Enhanced Bidirectional CMKD is more efficient than the SOTA C2KD method due to the use of BCSF as a filtering method, which operates in constant time by filtering only correct samples. This makes the method computationally more efficient compared to KRC-based filtering in C2KD.

5. Experiments

5.1. Dataset

The dataset we select contains 7 of the most collected classes from the data inventory: *concrete*, *tactile pavement*, *subway grate*, *manholes*, *bricks*, *dirt*, and *cellar doors*. Each class contains nearly 60 minutes of data, ensuring a robust foundation to train on. Additionally, the dataset is split such that entire recorded data segments are placed in either the training set (81%) or the validation/test set (19%), preventing data pollution.

5.2. Implementation Details

Data Pre-Processing. Akin to [27], both modalities of data are sliced into 1-second duration clips utilizing a sliding window technique advancing the start of each 1-second long clip 0.5 seconds ahead of the previous slice. The audio clips are then transformed into Mel-Spectrograms as suggested by [29] while an image is extracted from the center frame of the 1-second video clip.

Models Implementation. We select AST-Tiny224 for the audio student model and DeiT-Base224 for the image teacher model, considering the highest macro accuracy among the models tested. For the AST model, we follow the implementation details provided in the original paper [14], while the DeiT-based models are implemented using the *timm* library [41].

Training Configurations. All models are trained on Nvidia 4090 GTX. We apply the same training strategy utilizing the Adam optimizer with a learning rate of 0.0001, a learning scheduler of *ReduceLROnPlateau* with patience set at 5 epochs, and early stop with patience set at 20 epochs.

5.3. Result

Investigate Performance of Advanced KD Methods. To further investigate the performance of advanced KD techniques in a cross-modal setting, we evaluate several KD loss functions. The results, shown in Table 4, indicate that not all advanced KD methods are effective in cross-modal situations. Many of these methods are originally designed for same-modality scenarios, where the teacher and student models perceive similar types of data. Even in cases involving multimodal data, such as depth and RGB images, the modalities are still relatively similar. Surprisingly, Relational Knowledge Distillation (RKD) [31] demonstrated impressive results, outperforming other advanced KD methods. However, our proposed Enhanced Bidirectional CMKD method strongly outperforms these advanced KD methods.

KD Loss	Macro Acc.	Macro F1.	Micro Acc.	Change
w/o	85.78%	85.46%	84.95%	-
KL	85.49%	85.21%	84.68%	-0.29%
NKD [46]	85.43%	85.15%	84.68%	-0.35%
RKD [31]	86.27%	86.07%	85.54%	0.49%
DIST [23]	85.75%	85.33%	84.95%	-0.03%
Ours	**87.62%**	87.47%	87.08%	**1.84%**

Table 4. Performance of various advanced KD-Loss with the architecture of DeiT-Base224-to-AST-Tiny224. The proposed Enhanced Bidirectional CMKD (Ours) demonstrates the highest improvement.

Comparison with Other Bidirectional KD Methods. We compare our Enhanced Bidirectional CMKD method

with other SOTA bidirectional KD methods, including SHAKE and C2KD, as shown in Table 5. C2KD primarily focuses on CNN-based models and describes the use of proxy teacher and student structures only within CNN architectures. Therefore, we compare the change rate to evaluate the effectiveness of our approach. We apply our BCSF filtering method to C2KD (denoted as C2KD*) and observe better performance, demonstrating the robustness of BCSF.

Our proposed Enhanced Bidirectional CMKD method significantly outperforms these approaches. With the AST-Tiny224 student and DeiT-Base224 teacher, our method achieves a Macro Accuracy of 87.62%, which is an improvement of 1.84% over the AST-Tiny224 baseline. Even with the smaller DeiT-Tiny224 teacher, our method (denoted as Ours†) achieves a Macro Accuracy of 86.74%, representing a 0.96% improvement. These results clearly demonstrate that our Enhanced Bidirectional CMKD method effectively addresses the challenges of cross-modal knowledge distillation and significantly improves performance in comparison to SOTA methods.

KD Loss	Baseline	Macro Acc.	Macro F1.	Micro Acc.	Change
w/o	△	84.75%	84.39%	84%	-
w/o	□	85.78%	85.46%	84.95%	-
SHAKE [26]	△	81.55%	81.26%	80.61%	-3.20%
C2KD [24]	△	85.54%	85.17%	84.73%	0.79%
C2KD*	△	85.58%	85.18%	84.73%	0.83%
Ours	□	**87.62%**	87.47%	87.08%	**1.84%**
Ours†	□	**86.74%**	86.43%	85.95%	**0.96%**

Table 5. Performance comparison of SOTA bidirectional KD methods. "Baseline" indicates which baseline to compare against, where △ represents ResNet18 and □ represents AST-Tiny224. "C2KD*" denotes C2KD with BCSF filtering. "Ours" refers to AST-Tiny224 student and DeiT-Base224 teacher, and "Ours†" refers to AST-Tiny224 student and DeiT-Tiny224 teacher, both trained using the proposed Enhanced Bidirectional CMKD.

5.4. Ablation Study

To further understand the contributions of each component in our Enhanced Bidirectional CMKD method, we conduct an ablation study (Table 6). We analyze the impact of Correct Sample Filtering (CSF) with unidirectional KD (teacher to student), Bidirectional KD alone and proposed Enhanced Bidirectional CMKD method utilizing Bidirectional KD and BCSF on the student-teacher architecture as AST-Tiny224-to-DeiT-Tiny224.

CSF + KD	Bidirectional KD	BCSF	Macro Acc.	Macro F1.	Micro Acc.	Change
			85.78%	85.46%	84.95%	-
✓			85.76%	85.45%	84.92%	-0.02%
	✓		85.99%	85.68%	85.18%	0.21%
	✓	✓	86.74%	86.43%	85.95%	**0.96%**

Table 6. Ablation study on each module with the architecture of DeiT-Tiny224 teacher and AST-Tiny224 student.

As we can see from the table, applying bidirectional KD increases the Macro Accuracy to 85.99%, which is a 0.21% improvement. This indicates that allowing the teacher to learn from the student in a bidirectional manner contributes positively to the model's performance. However, incorporating CSF alone with unidirectional KD, resulted in a slight decrease. These results underscore the importance of using both bidirectional KD and BCSF, which increases the Macro Accuracy to 86.74% with 0.96% improvement.

5.5. Intuitive Justification

To intuitively explain why unidirectional KD with CSF fails to improve performance, consider the analogy of a tennis coach (image modality) teaching a soccer player (audio modality). While the coach filters out irrelevant techniques using CSF, the remaining guidance is still tailored to tennis, making it difficult for the soccer player to apply effectively. Without feedback from the soccer player, the coaching doesn't fully translate to soccer skills. However, with Bidirectional KD, the soccer player (student) can inform the coach (teacher), allowing the coach to adjust their instructions to better fit the context of soccer. In our model, Bidirectional KD enables this mutual refinement: the teacher model adapts its guidance based on the student's feedback, and vice versa. The inclusion of BCSF ensures both models learn the correct knowledge from each other, addressing the knowledge misalignment problem and ultimately boosting performance.

6. Conclusion

In this paper, we present the Sidewalk Material Data Acquisition Framework (SMDAF) to collect large-scale sidewalk material data, which is an important landmark for BLV individuals in urban navigation. We propose an Enhanced Bidirectional Cross-Modal Knowledge Distillation (CMKD) method, incorporating Bidirectional Correct Sample Filtering (BCSF) to improve the performance of audio material classifier in SMDAF. Our extensive experiments demonstrate that our method significantly outperforms SOTA, achieving higher accuracy and robustness in cross-modal knowledge transfer.

Future work will focus on refining the data acquisition process and improving model performance through advanced CMKD techniques using *training in the loop*. Additionally, a BLV user study will be conducted to determine an acceptable accuracy for BLV individuals with respect to sidewalk landmark identification.

7. Acknowledgments

The work is supported by the National Science Foundation through awards CNS-2131186, CNS-1827505, and the US Air Force Office of Scientific Research (AFOSR) through award FA9550-21-1-0082.

References

[1] Apple maps. https://www.apple.com/maps/. Accessed: 2024-07. 1

[2] Google map. https://www.google.com/maps. Accessed: 2024-07. 1

[3] Nyc dot - street design manual. https://www.nycstreetdesign.info/material/sidewalks. Accessed: 2024-07. 5

[4] Nyc mayor's office for people with disabilities (mopd). https://www1.nyc.gov/site/mopd/index.page. Accessed: 2024-07. 2

[5] Alex; Murray Dennis; Abbott, Andrew; Deshowitz and Eric C. Larson. Walknet: A deep learning approach to improving sidewalk quality and accessibility. *SMU Data Science Review*, 2018. 1

[6] Tousif Ahmed, Patrick Shaffer, Kay Connelly, David Crandall, and Apu Kapadia. Addressing physical safety, security, and privacy for people with visual impairments. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 341–354, Denver, CO, June 2016. USENIX Association. 2

[7] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015. 3

[8] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5178–5193. PMLR, 23–29 Jul 2023. 3, 5

[9] NYC Open Data. Pedestrian ramp locations. https://data.cityofnewyork.us/Transportation/Pedestrian-Ramp-Locations/ufzp-rrqu/about_data. Accessed: 2024-07. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[11] The Centers for Disease Control and Prevention. Prevalence estimates for vision loss and blindness. https://www.cdc.gov/vision-health-data/prevalence-estimates/vision-loss-prevalence.html. Accessed: 2024-07. 1

[12] Jon E Froehlich, Mikey Saugstad, Edgar Martínez, and Rebeca de Buen Kalman. Sidewalk accessibility in the us and mexico: Policies, tools, and a preliminary case study. In *CSCW2020 Workshop on Civic Technologies: Research, Practice, and Open Challenges*, 2020. 1

[13] Alexandra-Ioana Georgescu, Hoda Allahbakhshi, and Robert Weibel. The impact of microscale street elements on active transport of mobility-restricted individuals: A systematic review. *Journal of Transport & Health*, 38:101842, 2024. 3

[14] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 3, 5, 7

[15] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2827–2836, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society. 2, 3

[16] Kotaro Hara, Shiri Azenkot, Megan Campbell, Cynthia L Bennett, Vicki Le, Sean Pannella, Robert Moore, Kelly Minckler, Rochelle H Ng, and Jon 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. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 6

[18] James F Herman, Steven P Chatman, and Steven 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. 1

[19] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan 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. 2, 3, 5

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 3, 5, 6

[21] Maryam Hosseini, Fabio Miranda, Jianzhe Lin, and Claudio T. Silva. Citysurfaces: City-scale semantic segmentation of sidewalk materials. *Sustainable Cities and Society*, 79:103630, 2022. 2, 3

[22] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 5

[23] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *arXiv preprint arXiv:2205.10536*, 2022. 7

[24] Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C2kd: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16006–16015, June 2024. 2, 3, 6, 7, 8

[25] Yasha Iravantchi, Yi Zhao, Kenrick Kin, and Alanson P. Sample. Sawsense: Using surface acoustic waves for surface-bound event recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. 3

[26] Lujun Li and Jin Zhe. Shadow knowledge distillation: Bridging offline and online knowledge transfer. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3, 6, 8

[27] J Liu, WP Lam, Z Zhu, and H Tang. Surveying sidewalk materials for and by individuals who are blind or have low vision: Audio data collection and classification. International Conference on SMART MULTIMEDIA, 2024. 7

[28] Jiawei Liu, Hao Tang, William Seiple, and Zhigang Zhu. Annotating storefront accessibility data using crowdsourcing. *Journal on Technology and Persons with Disabilities*, 10:154–170, 2022. 2

[29] Alessandro Maccagno, Andrea Mastropietro, Umberto Mazziotta, Michele Scarpiniti, Yong-Cheol Lee, and Aurelio Uncini. A cnn approach for audio classification in construction sites. *Progresses in Artificial Intelligence and Neural Systems*, pages 371–381, 2021. 2, 3, 5, 7

[30] Gilberto Marzano, Joanna Lizut, and Luis Ocha Siguencia. Crowdsourcing solutions for supporting urban mobility. *Procedia Computer Science*, 149:542–547, 2019. 2

[31] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 3, 7

[32] Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro 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. 2

[33] Mingxing Tan and QV Le. Efficientnetv2: Smaller models and faster training. arxiv 2021. *arXiv preprint arXiv:2104.00298*. 5

[34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers amp; distillation through attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021. 3, 5, 6

[35] Alain Trémeau, Sixiang Xu, and Damien Muselet. Deep learning for material recognition: most recent advances and open challenges. *arXiv preprint arXiv:2012.07495*, 2020. 3

[36] Alain Trémeau, Sixiang Xu, and Damien Muselet. Deep learning for material recognition: most recent advances and open challenges. *arXiv preprint arXiv:2012.07495*, 2020. 3

[37] Lazaros Vrysis, Iordanis Thoidis, Charalampos Dimoulas, and George Papanikolaou. Experimenting with 1d cnn architectures for generic audio classification. In *Audio Engineering Society Convention 148*. Audio Engineering Society, 2020. 2, 5

[38] Guangzhi Wang. Dynamic knowledge distillation with cross-modality knowledge transfer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2974–2978, 2021. 2

[39] Takumi Watanabe, Hiroki Takahashi, Goh Sato, Yusuke Iwasawa, Yutaka Matsuo, and Ikuko Eguchi 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. 2

[40] Galen Weld, Esther Jang, Anthony Li, Aileen Zeng, Kurtis Heimerl, and Jon E. Froehlich. Deep learning for automatically detecting sidewalk accessibility problems using streetscape imagery. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 196–209, New York, NY, USA, 2019. Association for Computing Machinery. 1

[41] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 7

[42] Magdalena Wojnowska-Heciak, Jakub Heciak, and Adam Kłak. Concrete paving slabs for comfort of movement of mobility-impaired pedestrians—a survey. *International journal of environmental research and public health*, 19(6):3183, 2022. 3

[43] Ruobing Xie, Shaoliang Zhang, Rui Wang, Feng Xia, and Leyu Lin. Explore, filter and distill: Distilled reinforcement learning in recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4243–4252, New York, NY, USA, 2021. Association for Computing Machinery. 7

[44] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding cross-modal knowledge distillation. In *ICLR*, 2023. 2, 3, 5

[45] Jing Yang, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. Knowledge distillation via softmax regression representation learning. In *International Conference on Learning Representations*, 2021. 3

[46] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17185–17194, 2023. 7

[47] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 6