electronics

MDPI

Article

Correspondence Learning for Deep Multi-Modal Recognition and Fraud Detection †

Jongchan Park [1], Min-Hyun Kim [2] and Dong-Geol Choi [3,*]

1 Lunit Inc., Seoul 06241, Korea; jcpark@lunit.io
2 Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea; minhyun@kaist.ac.kr
3 Department of Information and Communication Engineering, Hanbat National University, Daejeon 34014, Korea
* Correspondence: dgchoi@hanbat.ac.kr
† This paper is an extended version of our paper published in 2019 International Conference on Electronics, Information, and Communication (ICEIC).

Abstract: Deep learning-based methods have achieved good performance in various recognition benchmarks mostly by utilizing single modalities. As different modalities contain complementary information to each other, multi-modal based methods are proposed to implicitly utilize them. In this paper, we propose a simple technique, called correspondence learning (CL), which explicitly learns the relationship among multiple modalities. The multiple modalities in the data samples are randomly mixed among different samples. If the modalities are from the same sample (not mixed), then they have positive correspondence, and vice versa. CL is an auxiliary task for the model to predict the correspondence among modalities. The model is expected to extract information from each modality to check correspondence and achieve better representations in multi-modal recognition tasks. In this work, we first validate the proposed method in various multi-modal benchmarks including CMU Multimodal Opinion-Level Sentiment Intensity (CMU-MOSI) and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) sentiment analysis datasets. In addition, we propose a fraud detection method using the learned correspondence among modalities. To validate this additional usage, we collect a multi-modal dataset for fraud detection using real-world samples for reverse vending machines.

Keywords: deep learning; pattern recognition; multi-modal learning; classification

Citation: Park, J.; Kim, M.-H.; Choi, D.-G. Correspondence Learning for Deep Multi-Modal Recognition and Fraud Detection. *Electronics* **2021**, *10*, 800. https://doi.org/10.3390/electronics10070800

Academic Editor: Krzysztof Szczypiorski

Received: 23 February 2021
Accepted: 25 March 2021
Published: 28 March 2021

1. Introduction

Advances in deep learning [1,2] have shown state-of-the-art performances in various recognition tasks [3–5]. Thanks to open-sourced deep learning frameworks [6–8], commercial applications [9,10] based on deep learning are made possible.

On the other hand, individual sensors have limited information, and different sensors have complementary information to provide. Therefore, multi-modal systems with multiple sensors have been developed to exploit the complementary information [11–14]. For example, RGB camera sensors provide rich information under sufficient lighting but may fail during night-time. Therefore, thermal imaging sensors and LiDAR sensors can be used for a more robust autonomous driving system [12]. In action recognition tasks [15–17], initial approaches use RGB image sequences and optical flow sequences as model inputs, as RGB images provide contextual information and optical flow images provide motion information. To combine the two modalities, a naive and effective approach, called late-fusion, is to ensemble two separate model outputs. A recent dataset [18] for action recognition has shown that some actions are only recognizable with an audio modality. Apart from naive multi-modal approaches in which individual models are trained for each modality, there are studies using a single model with multi-modal inputs. A pioneering

work [19] proposes a model with multi-modal inputs for multiple tasks, including image and text inputs, classification, detection, translation and captioning tasks. Although the final performances are weaker than the state-of-the-art of each task, it is a proof-of-concept for utilizing multi-modal inputs.

While the aforementioned tasks can achieve high performance with single-modality inputs, sentimental analysis tasks [20,21] require the use of multiple modalities. Sentimental analysis is a task performed to predict the sentiment of a person in a given video clip. Three heterogeneous modalities can be used for this task: RGB frames, an audio sequence, and spoken sentences. The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset [21] provides samples to explicitly show that the modalities are complementary and thus are essential to accurately classify those samples. As a result, recent studies [21–23] propose ways to fuse multi-modal inputs to utilize complementary information in multiple modalities.

Multiple modalities can also be used for fraud detection. Face anti-spoofing is a widely-known task to identify forgery inputs. The face anti-spoofing benchmarks [24,25] use RGB, depth and IR sensors to identify fake inputs for face identifications. A single-modality system can be easily fooled. In a famous incident, called "facegate" [26], a printed face on a paper or a 3D mask was able to fool an RGB-based system. On the other hand, simultaneously fooling multiple modalities is much more difficult, and we propose a simple method to detect fake data using the learned correspondence among multiple modalities.

Inspired by a recent study in self-supervised learning [27], we propose a correspondence learning scheme to exploit the relationship among multiple modalities. The correspondence is defined according to whether the multiple modalities are taken from the same sample—that is, the modalities from the same sample have positive correspondence; when one or more modalities are taken from a different sample, the modalities have negative correspondence. The overall process is illustrated in Figure 1. Each modality has its own feature extractor, and correspondence learning is added as an auxiliary task on top of the original task. After the feature extraction, we synthesize negative correspondence samples by swapping features from other samples and train a sub-network to predict the correspondence of given features. Positive correspondence samples are the un-swapped, original paired features. In this way, the sub-network is trained to predict the correspondence among modalities, and the extracted multi-modal features contain information about the correspondence with each other. We empirically show that correspondence learning can significantly improve the performance of models with multi-modal inputs in sentimental analysis benchmarks. In addition, the learned correspondence among modalities can be used for fraud detection, and we can effectively filter out fake inputs. As camera modality is frequently used as inputs, we focus on preventing look-alike frauds in reverse vending machine cases. Nevertheless, the idea can easily be extended to other situations. In the reverse vending machine dataset, the naive joint learning may not fully utilize multi-modal information, so we additionally use an attention mechanism to keep the performance on par with the conventional approach and simultaneously detect fraud inputs. Please note that this is an extended version of our conference paper [28].

The paper contributes in the following ways:

1. We propose correspondence learning (CL), a novel and simple technique to explicitly learn the relationship among modalities;
2. In sentimental analysis benchmarks, we show that CL significantly improves performances with the simple auxiliary correspondence learning task;
3. In the garbage classification task, we show that single-modality-based models are vulnerable to fraud inputs and unseen class objects (out-of-distribution), and the learned correspondence can be used for fraud detection with high detection rates. We also show that the material classification is possible with non-contact ultrasound sensors.

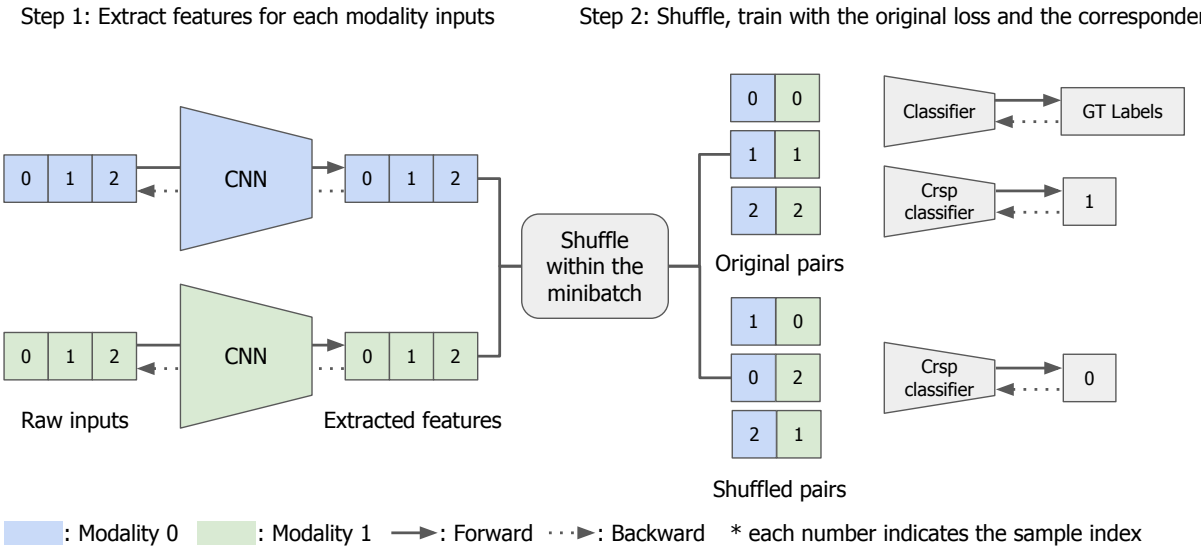

Figure 1. Overall architecture of the proposed correspondence learning approach. In the figure, two different modalities are shown in two colors (blue/green), and three samples are given for training. The overall architecture consists of three parts: a feature extractor, classifier and correspondence classifier. The training process consists of two parts: (1) each modality input in each sample is given to the modality-specific feature extractor (written as CNN), (2) multi-modal features are concatenated and trained with the ground-truth labels and are used as positive correspondence samples (**top**), while shuffled samples are used as the negative correspondence pairs (**bottom**). Note that there are two losses used: the cross-entropy loss for classification, and the binary cross-entropy loss for correspondence learning.

2. Related Works

After the success of deep learning in single modality tasks [1,2] with a large scale dataset [3], a number of datasets have been proposed with multiple modalities in action recognition [15,16,18], sentiment analysis [20,29] and face anti-spoofing [24,25]. In action recognition datasets, there are three different modalities to be used: visual (RGB sequence), motion (optical flow) and audio. Until recently, only visual and motion modalities have been used, and a common approach is to train one model for each modality and ensemble via late-fusion [30]. The two modalities are crucial yet complementary to each other: the visual modality contains the contextual information, and the the optical flow modality contains the motion information. A recent study [31] shows that the complementary information can be partially distilled from motion to visual modality, and the single visual modality can achieve comparable results to the two-stream approaches. While modalities may share information on the target task, there are fundamental differences in different modalities: RGB sensors cannot defend against 3D masks in face identification [25] and some actions can only be recognized with audio modality [18] (e.g., snapping), while some sentiments can only be expressed via tone(audio), words(language) or facial expressions(visual) [29].

A widely-used approach for multi-modal recognition is late-fusion [30], in which one model is trained for each modality and the predictions from multiple modalities are combined. While aligning visual and motion modalities is straightforward [32], as they are similar in terms of spatial characteristics, aligning heterogeneous modalities, such as visual, word and audio, requires sophisticated techniques [21–23]. The proposed correspondence learning can be used as an auxiliary task along with any of the fusion methods. It can be regarded as semi-supervised learning with self-supervision from cross-modal correspondence. In the experiments, we show that the auxiliary task of CL can improve the performance.

Methods to fool deep learning models, or adversarial attacks [33–35], have been actively developed to identify the vulnerability of the deep learning models and make them robust. Adversarial samples are easy to synthesize with minimal noises that are not visible

to human eyes [34]. Surprisingly, the adversarial samples can be extended to the physical world, and adversarial patches can fool the recognition models [35]. A real world case [26], in which a face recognition system was fooled by pictures, indicated the vulnerability of single modality inputs. To mitigate this issue, several benchmarks have been proposed to detect fraud inputs [24,25]. Previous approaches on fraud detection (i.e., face anti-spoofing) are only applicable with spatially-aligned image modalities. In contrast, our proposed method is simple and widely applicable without any constraints on the inputs.

Several previous works have tried to exploit cross-modal correspondence [27,36]. SoundNet [36] uses a teacher–student framework to distill the discriminative information from a visual model into an audio model and achieves a new state-of-the-art in audio classification benchmarks. Relja Arandjelović et al. [27] use the correspondence between audio and visual modalities to train a cross-modal retrieval system, where images can be retrieved with audio inputs, and vice versa. The proposed method in this manuscript also exploits cross-modal correspondence, but it differs from all previous works in several aspects. First, the previous works focus on using the learned correspondence for a specific task. For example, Relja Arandjelović et al. [27] use the learned correspondence for sound-image retrieval and sound localization in the given video frames. On the other hand, we propose correspondence learning as an auxiliary feature for another task and aim to improve the performance of the original task. In the fraud detection task, we can keep the original task of garbage classification and add an important feature of fraud detection at the same time with only a marginal overhead. Furthermore, in contrast to SoundNet [36], which only distills the rich information from a visual modality to sound modality, the proposed method jointly learns the correspondence among multiple modalities and automatically learns information from other modalities.

We propose correspondence learning (CL), which is an auxiliary task of classifying whether modalities are coming from the same sample or not. CL is inspired by representation learning [27] using visual and audio modalities. To show the efficacy of CL for multi-modal recognition tasks, we utilize CL in two multi-modal sentiment analysis benchmarks [20,29] with several state-of-the-art baselines. In addition, the learned correspondence among modalities can detect any inconsistency among modalities, which can be an indication of fraud inputs. Thus, we propose a method and dataset to detect fraud samples without any fraud samples in the training set.

3. Method: Correspondence Learning

In this section, we explain the motivation, the concept and the detailed implementation of the proposed correspondence learning approach.

3.1. Motivation and Initial Observations

Multiple modalities contain rich information that partially overlap and are partially complementary to each other. For example, in the garbage classification introduced in the following sections, the high-level class information can be contained in the RGB image of the object and also in the non-contact ultrasound signals, but in two different aspects; that is, the RGB images contain appearance information of the class, and the ultrasound signals contain material information. The two aspects are complementary to each other, so a multi-modal recognition model is expected to fully exploit the complementary information across all modalities and utilize the dynamic relationship among modalities.

For the proposed garbage classification task, we expect the multi-modal model to exploit the relationship among modalities: if the appearance is an aluminum can, then the ultrasound signal should contain a metallic characteristic. There should be shared knowledge distilled among different modalities, and this can be partially trained by learning the correspondence of the modalities; that is, learning whether the modalities are from the same sample or not. In addition, this correspondence idea can be simply extended to fraud detection tasks. If there are negative relationships, then we expect the model to have lower confidence regarding the prediction. However, due to a common phenomenon of

deep neural networks—i.e., over-confidence [37]—the model outputs confident predictions regardless of the inter-modality relationship. In our initial experiments for the garbage classification task, we observe that a confidence thresholding cannot achieve satisfactory results in fraud detection.

To this end, we propose an auxiliary correspondence learning task for multi-modal recognition tasks. We show that the auxiliary CL can improve multi-modal recognition tasks and can be effectively used for fraud detection.

3.2. Multi-Modal Correspondence Learning

In this section, we explain how correspondence learning is implemented. As an example, we assume there are two modalities, illustrated in Figure 2. Note that correspondence learning is a binary classification task to predict whether the multiple modality features are extracted from the same data. The features from the same sample are positive pairs for correspondence learning; the features from different samples are negative pairs. Note that the positive pairs are already given in the data, and the negative samples are generated by simple shuffling; no extra data are required for the proposed correspondence learning. Multi-modal models usually have modality-specific feature extractors and merge the features or predictions afterwards. When the target task includes fraud detection, we can treat the unmatched pairs as an extra class, and the implementation is straightforward; if it is used as an auxiliary task, we can simply add a branch to the network and train the branch to learn the binary classification of positive and negative correspondences.

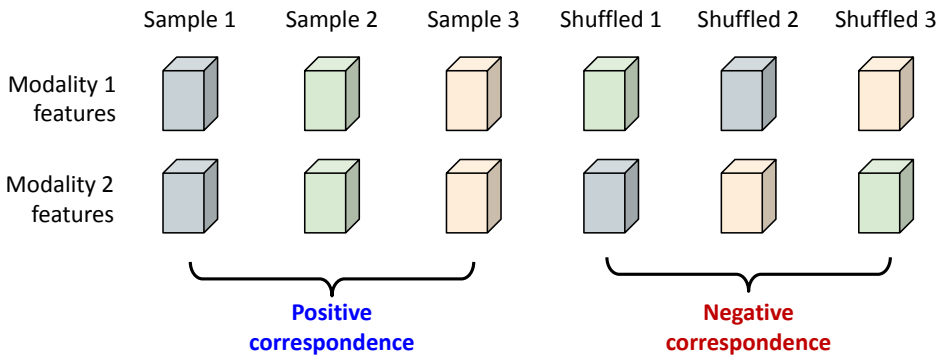

Figure 2. An illustration of the sample generation process for correspondence learning. Two modalities are shown, and the feature pair from the same sample (the leftmost three columns) is regarded as positive correspondence; after shuffling, the feature pair for which features are not from the same sample is regarded as negative correspondence. Note that all feature pairs are used for the auxiliary correspondence learning, and only the leftmost three feature pairs are used to train the original task.

Specifically, in this work, when we choose to add an auxiliary task of CL, we add an auxiliary model after using modality-specific feature extractors. The feature pairs may be sampled from the same object or synthetically paired by shuffling. The model design is very simple: concatenate all modality features and apply a multi-layered perceptron. The predicted correspondence score is normalized with the sigmoid function. The hidden layer sizes are identical to the input feature size for design simplicity.

$$p(y_{ij}) = Classifier_{CL}(concat(f_1[i], f_2[j])) \tag{1}$$

$$L_{CL} = -\frac{1}{N}\sum_{i,j} BCE(p(y_{ij}),\ y_{ij}), \quad \text{where } y_{ij} = \begin{cases} 1 & \text{if } i == j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

As shown in Equation (1), i and j are random indices in the minibatch, and the concatenation is done in the channel dimension. Equation (2) shows how we calculate

the losses for the correspondence learning. i, j are the indices of the samples for different modalities. The correspondence loss is a simple binary cross entropy, where the label indicates whether the modality features are sampled from the same data point or not.

The correspondence loss L_{CL}, is added to the original task's loss as below:

$$L_{total} = L_{task} + \alpha L_{CL} \tag{3}$$

L_{task} is the original loss for each task. α is a hyper-parameter used to tune the magnitude of the correspondence loss. We have tuned α by a simple grid search from 0.1 to 1.0 for each task. As a simple yet comprehensive explanation, Algorithm 1 shows the pseudo code of the method explained above. Note that the original task part (upper part of the code) can be adapted to any kind of task, and the method can be easily extended to more than two modalities. Throughout this paper, we assume that there are two modalities for simplicity.

Algorithm 1 Pseudo code for the proposed method.

Require: Multi-modal inputs x_1, x_2 with minibatch size n, target labels t

The original task for multi-modal classification
$f_1 = F_1(x_1)$
$f_2 = F_2(x_2)$ ▷ Extract feature for modality 1 and 2
$p_{task} = Classifier(concat(f_1, f_2))$
$L_{task} = CE(p_{task},\ t)$ ▷ Cross-entropy loss for the original task

Proposed correspondence learning
$ind = [0, 1, 2, \ldots, n-1]$
$ind_1 = concat(ind,\ shuffle(ind))$
$ind_2 = concat(ind,\ shuffle(ind))$ ▷ Half in the original order, and half shuffled
$p_{CL} = Classifier_{CL}(concat(f_1[ind_1],\ f_2[ind_1]))$
$L_{CL} = BCE(p_{CL},\ ind_1 == ind_2)$ ▷ Learn if the features come from the same sample

$L_{all} = L_{task} + \alpha L_{CL}$

4. Garbage Classification Task for Fraud Detection

In this work, we propose a multi-modal garbage classification task to evaluate the robustness of multi-modal recognition systems against fraud inputs. Three different modalities are recorded for each sample: an RGB image, a non-contact ultrasound signal and the weight. The dataset consists of three garbage types (can, PET and glass bottle) and fraud examples. The task is to classify a given sample among three garbage types and reject any fraud samples. The fraud samples are defined as visually similar samples (VS) or non-target samples. VS samples are intentionally crafted to confuse the recognition system, and non-target samples are any objects that are not included in the three classes.

A reverse vending machine (RVM) collects empty, recyclable containers from users and gives out rewards. There are several products in operation, such as TOMRA [38], RVM Systems [39] and Superbin [40]. Photos of commercial RVM systems are shown in Figure 3. Previous systems often used UPC or bar code scanners to specifically identify the incoming containers. However, such systems require a huge and up-to-date database of containers and cannot handle deformed (crumpled) containers for which UPC or bar codes are not identifiable. To handle such problems, we have built a simple vision-based system with deep convolutional neural networks for garbage classification; it has shown over 99% accuracy for classification. Previously built systems used image inputs only and were vulnerable to fraud inputs such as look-alike samples.

Since an automated RVM gives back immediate rewards, it is crucial to not give a false positive classification; that is, to identify a non-target object as one of the target class. The system must reject any non-target inputs and ask the users to input target class objects. If the system accepts non-target objects, this vulnerability may be abused by malicious

users and can lead to huge losses for the company. This is a fundamental threat to the RVM business model.

(a) (b)

Figure 3. Commercial reverse vending machines (RVM). (**a**) The RVM from TOMRA and Superbin. (**b**) View from inside the RVM. Images are taken from TOMRA and Superbin's official websites (https://www.tomra.com, accessed on 23 February 2021, http://superbin.co.kr, accessed on 23 February 2021).

4.1. Hardware Settings

In this section, we introduce the data acquisition system and the types of databases for our experiments.

The hardware setup for data acquisition is shown in Figure 4a,b. We used a single pair of transmitter/receiver ultrasonic sensors (HG-M40TN2/HG-M40RN2, Hagisonic), a USB webcam sensor and a 5 kg load cell sensor. We used a controller (compactRIO-9036, National Instruments) to trigger and receive raw signals of the ultrasonic and load cell sensors. We triggered the ultrasonic sensor transmitter every 200 ms and recorded the raw input in the receiver at 1 mega samples per second. We recorded the load cell signal simultaneously. We acquired the image data with the USB webcam. Everything was controlled by the laptop computer. The controller and the USB webcam were connected to the laptop.

4.2. Dataset Composition

4.2.1. Raw Material Samples

To build the databases for our multi-modal classification task, we acquired sensor inputs from various objects using ultrasonic, camera and load cell sensors. There were two types of databases: the raw material database in which the target objects had the same shapes and different material types, as shown in Figure 4c,d; and the real object database, in which the target objects were real world objects including our target class objects (can, PET bottles, glass bottles), fraud inputs and non-target objects, as shown in Figure 5. The raw material types were stainless steel, aluminum, poly-carbonate and polyvinyl chloride. To learn material features that were robust to sizes and shapes, we made the objects for the raw material database in various shapes and sizes. We used three shapes: flat, cuboid and cylinder. Flat shapes had widths of 80 mm, 100 mm, 120 mm and 140 mm, heights of 100 mm, 200 mm and 300 mm, and 3T of thickness. Cuboids had square bases with 50 mm, 75 mm and 100 mm sides and 100 mm, 200 mm and 300 mm heights. Cylinders had circular bases with 50 mm, 75 mm and 100 mm diameters and 100 mm, 200 mm and 300 mm heights.

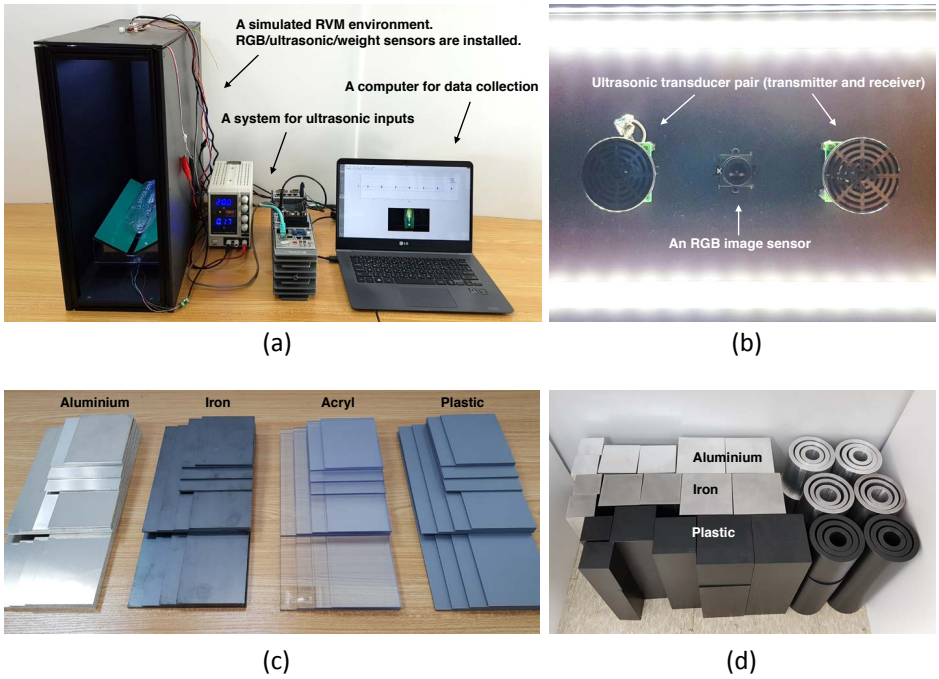

Figure 4. The hardware setup for data acquisition and raw materials used for the material database. (**a**) Overview of the setup. From left: the box for object placement, the power supply, the controller and the laptop. Ultrasonic, camera (RGB) and load cell sensors are attached to the box. (**b**) Inner-upper side of the box, where LED bars, ultrasonic transmitter/receiver and the camera sensor were attached. (**c**) Flat shapes and (**d**) cuboid and cylinder shapes.

Figure 5. Captured images in the dataset. (**a**) is a target class sample of a can. (**b**) is a visually similar fraud sample of printed can. (**c**) is a non-target sample of glove.

4.2.2. Real World Targets and Fraud Samples

In order to evaluate the robustness of any multi-modal approach against fraud inputs, we collected as many real world samples as possible to ensure the diversity of the target class objects. We acquired real-world garbage samples from a local recycling facility. In total, 167 cans, 141 PET bottles and 228 glass bottles were collected with a multi-modal system. In addition, we collected fraud inputs to validate the robustness of a multi-modal recognition system, including visually similar (VS) samples and non-target samples. Note that the fraud inputs were only included in the validation set.

VS samples can be viciously manipulated to confuse the recognition system. We assumed that the visual information is easy to confuse and collected 60 visually-similar samples by printing out the target class objects. As shown in Figure 5, the printed objects were realistic enough to "fool" a deep neural network system. Our initial experiment showed that a model using image inputs was able to reject 8.3% of the VS samples, leaving the rest of the VS samples to be mis-classified.

Although not viciously manipulated, non-target samples can also be potentially misclassified with high confidence. Therefore, we collected 29 non-target samples, with the objects used in our daily life, such as paper cups, gloves, plastic bags, human arms or clothes.

The classification was very trivial, and the final accuracy was 98.0% using only the image modality. At the same time, the rejection rates of VS and non-target samples were 8.3% and 6.9%, respectively. The limitation of single-modality recognition is clear and is discussed in the experiment section, along with the efficacy of the proposed correspondence learning.

5. Experiment Setups

5.1. Datasets

The proposed method was evaluated with state-of-the-art approaches in multi-modal recognition tasks. First, we used CMU Multimodal Opinion-Level Sentiment Intensity (CMU-MOSI) [20] and CMU-MOSEI [29] sentiment analysis datasets with Tensor Fusion Networks (TFN) [22] and Multilogue-Net [23]. Second, the correspondence learning technique was applied to fraud detection in garbage classification, while no fraud samples were available during training.

5.2. CMU-MOSI and CMU-MOSEI Dataset

The CMU Multimodal Opinion-Level Sentiment Intensity (CMU-MOSI) dataset is a collection of YouTube videos in which people express their opinions on various subjects. In total, 93 videos were collected, and the spoken words were transcribed and aligned with the video with multiple verification stages. In total, 2199 sentiment segments were extracted, and each segment was annotated with one of seven sentiments from highly negative to highly positive. The intensity of each sentiment was also annotated. The dataset was split into 52 training videos with 1151 segments, 10 videos with 296 segments and 31 videos with 752 segments. The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset [29] consists of 3228 monologue videos from YouTube. In total, 1000 unique identities are included in the dataset. A total of 22,676 sentences were extracted from the videos. The annotation method for sentiments was identical to CMU-MOSI dataset, but has five classes. In addition, the CMU-MOSEI dataset includes emotion annotations of happiness, sadness, anger, fear, disgust and surprise, along with likely annotations. The training, validation and test sets comprised 16,216, 1835, and 4625 sentences, respectively.

In our following experiments, we first reproduced the baseline performances using official implementations of the state-of-the-art methods on CMU-MOSI and CMU-MOSEI and showed the efficacy of CL. We focused on the sentiment analysis performance, and the binary performance indicated the classification between positive and negative sentiments. Except for the mean absolute error (MAE) metric, higher scores for all metrics indicate better results.

5.3. Baseline for the Garbage Classification Task

In this section, we briefly explain the baseline settings for the garbage classification task for fraud detection. The network design comprised three parts: feature extractors, attention layers and the classifiers. In addition, if not otherwise specified, the networks were trained with the Adam optimizer, with a learning rate of 1×10^{-4}.

5.3.1. Feature Extractors

For the image modality, we used ResNet-18 up to stage 4 as the feature extractor, followed by a global average pooling layer. A linear layer was added at the end and output a 1D feature vector of size 512. For the ultrasound modality, we used time–frequency data in the range (30 kHz, 50 kHz) as input. The feature extractor consisted of four 1D convolutions with a kernel size and stride of [(201, 5), (51,1), (51,1), (51,1)] with ReLU. Similar to the image feature extractor, a linear layer was added at the end and output a 1D

vector of size 512. For the weight modality, we used one-hot encoding where the bin size was 3 g per bin and the maximum weight was 600 g. The feature extractor consisted of three linear layers, where the hidden sizes were [512, 512, 512] with batch normalization and ReLU.

5.3.2. Attention Module

Features from different modalities were fed into the attention layers for feature refinement. The attention layer generated gates for each modality feature. The attention module consisted of three linear layers with output sizes [1536, 1536, 1536, 1536]. The last output was normalized with a sigmoid layer. The normalized output was divided into three vectors, and the three vectors were regarded as attention vectors for three modalities. Each modality feature vector was multiplied with the attention vector for feature refinement. There are two types of attention: cosine similarity and joint attention. When generating cosine attention for one modality feature, the other two modalities were used as inputs. Cosine similarity attention was generated by the element-wise multiplication of the two modality features. The joint attention was generated by MLPs where all features were used as the input. This simultaneously generated an attention map for all modalities. A sigmoid layer was used to normalize the attention gates' range to [0, 1]. Each modality feature was multiplied with the generated attention map for feature refinement.

5.3.3. Classifiers

The refined modality features were concatenated and fed into the joint classifier. The joint classifier was a four-layer multilayer perceptron (MLP) with batch normalization and ReLUs, with hidden sizes of [768,768,768]—the last output was the number of target classes. When we used the synthesized negative correspondence class, one extra class was added. In order to train each modality feature well, we also assigned separate classifiers for different modalities. In this way, even when a negative correspondence class instance was fed into the network, we could train separate branches with the real labels of each type of modality data. Note that for the fraud detection task, we used an extra class instead of an auxiliary task or model. This simplified the final prediction, as we only needed to choose the argmax of the final four class predictions; if we chose an auxiliary model, then we may have needed to choose a good threshold for the binary classification task. Each modality-specific classifier contained three linear layers with hidden sizes of [256,256]; the last output was the number of target classes. For modality-specific classifiers, we were not able to use the extra negative correspondence class.

Since neural networks are not usually designed for fraud detection, we used a heuristic method of fraud detection. Neural networks are usually trained with known classes and are not aware of unseen class instances. In classification networks, the output is softmax normalized and the answer is the maximum-likelihood output. In order to detect fraud inputs, we used a heuristic threshold for the likelihood: when the maximum-likelihood output was below the threshold, we regarded the input as a fraud input. In addition, in cases in which a negative correspondence class was used for training, the test inputs classified as the negative correspondence class were also regarded as fraud inputs.

6. Experiment Results

6.1. CMU-MOSI and CMU-MOSEI

In this section, we summarize the experimental results in the two sentiment analysis tasks. Two baselines were used for the CMU-MOSI dataset: Tensor Fusion Network (TFN) [22] and Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling (HPFN) [41]. We used the official implementations (https://github.com/Justin1904/TensorFusionNetworks, accessed on 23 February 2021), (https://github.com/jiajiatang0000/HPFN, accessed on 23 February 2021) to reproduce the baseline performances and added correspondence learning. Similarly, we reported CMU-MOSEI results using a strong baseline Multilogue-Net [23] and its official implementation (https://github.com/amanshenoy/

multilogue-net, accessed on 23 February 2021) for reproduction. The hyper parameter α in Equation (3) was searched from 0.1 to 1.0 with 0.1 intervals. Since we used the official implementation provided by the authors [22,41], the single best performance was reported after the grid search on α. We report the binary classification and regression results.

As shown in Tables 1 and 2 in both datasets and all baselines, correspondence learning resulted in significant performance improvements in the classification tasks. CL helped the model to learn representations that were more helpful for discriminatory tasks more effectively. Note that there are overheads in correspondence learning during training, as there are some auxiliary layers and an additional loss to be used. However, the auxiliary parts can be removed during evaluation, so there is no overhead during testing.

Table 1. CMU Multimodal Opinion-Level Sentiment Intensity (CMU-MOSI) experiment results. All the results are reproduced, except for TFN we are using the performance from the original paper. MAE: Mean Absolute Error.

Method	Binary	Regression	
	Acc (%)	MAE	r
Random	50.1	1.86	0.057
TFN [22]	77.1	0.87	0.70
TFN [22] + CL	78.6	0.79	0.70
HPFN [41]	77.16	0.984	0.66
HPFN [41] + CL	77.97	0.995	0.63
Human	87.5	0.71	0.82

Table 2. CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) experiment results. All the results are reproduced. CL: correspondence learning. MAE: Mean Absolute Error.

Method	Binary	Regression	
	Acc (%)	MAE	r
Multilogue-Net [23]	78.06	0.609	0.46
Multilogue-Net [23]+CL	80.19	0.605	0.48

6.2. Raw Materials with Ultrasound

Then, a single pair of non-contact ultrasonic sensors and a 1D CNN were used to show that raw material classification is viable, especially when the objects are in various shapes and poses. In our target task of reverse vending machines, the object material is important. It has been shown that material classification is viable with ultrasonic signals [42,43], so we decided to use ultrasonic sensors as a new modality.

However, the experiments conducted in [42,43] were highly controlled in that the target objects were flat board shapes with the same pose and distance from the sensors. In real-world cases, the target objects would be in various shapes, sizes and poses. Therefore, we needed to show that ultrasonic signals still contained enough information in such challenging cases.

In the experiment, we used the raw material dataset acquired in Section 4.1 with various shapes, sizes and poses. The feature extractor and the classifier were the same as those specified in Section 5.3. According to the result in Table 3, we empirically verified that material classification is possible with a single pair of non-contact ultrasonic sensors and a 1D CNN.

Table 3. Raw material classification with ultrasound.

Material Type	**Accuracy (%)**	**Material Type**	**Accuracy** (%)
Acryl	100.0	Aluminum	100.0
Aluminum	100.0	Plastic	91.6
Iron	100.0	Iron	91.8
Plastic	96.0		
Avg acc	99.0	Avg acc	94.4
2D shapes		3D shapes	

6.3. Fraud Detection Using Real-World Data

In this section, we show that our proposed model was able to learn to classify target inputs and detect fraud inputs. First, we show the effects and the limitations of the naive use of multi-modal inputs. Next, we show that the two proposed techniques achieved a high fraud detection rate while maintaining high accuracy for target class objects.

6.3.1. Multi-Modal Inputs

When multiple modality inputs are used together, we expect a better performance of DNNs in general. As shown in Table 4, the joint use of multiple modalities was able to achieve a higher fraud detection rate for both visually similar inputs and non-target inputs. The change in target class object accuracy was negligible. In terms of the fraud detection rate (visually similar inputs and non-target inputs), the efficacy of multiple modalities can be observed. Fraud inputs are all unseen classes for DNNs, and the features are different from those of target class objects. We conjecture that multi-modal inputs will show more differences in features compared to single modal cases. Therefore, multi-modal inputs achieve a higher fraud detection rate.

Table 4. Classification results using multi-modal inputs in a real-world database. W denotes the weight modality, target denotes the accuracy in target class objects in Figure 5, VS denotes the fraud detection rate for visually similar fraud inputs, and non-target denotes the fraud detection rate for non-target inputs.

Modality	**Target (%)**	**VS (%)**	**Non-Target (%)**
Image (IMG)	98.0	8.3	6.9
Ultrasound (US)	82.3	15.0	6.9
IMG + US	96.5	15.0	6.9
IMG + US + W	97.5	18.3	13.7

6.3.2. Multi-Modal Attention

The purpose of multi-modal attention is to refine the concatenated multi-modal features by self-attention. As all modality feature vectors are used to generate attention masks for each other, we expect the network to learn better representations. As shown in Table 5, multi-modal attention achieved a higher target class accuracy and a higher fraud detection rate for both visually similar fraud inputs and non-target inputs. Generally, improved performance indicates better representations. For fraud inputs, the inter-modal relationships were different from those of target class objects. As for attention jointly using multi-modal features, we suspect that the network detects fraud inputs using the changes in the inter-modal relationship.

Table 5. Classification results using correspondence learning and multi-modal attention in a real-world database. CL denotes correspondence learning and Att denotes multi-modal attention.

Modality	CL	Att	Target (%)	VS (%)	Non-Target (%)
IMG + US + W	-	-	97.5	18.3	13.7
IMG + US + W	-	✓	99.5	21.7	20.7
IMG + US + W	✓	-	81.8	86.7	93.1
IMG + US + W	✓	✓	94.0	91.7	93.1

6.3.3. Correspondence Learning

Correspondence learning explicitly trains the network to learn the correlation among modalities, and a much higher fraud detection rate was achieved, as shown in Table 5. However, there was a decrease in target class accuracy. We argue that the fraud detection rate improved because the classifier learned the correlation between modalities through correspondence learning. As fraud detection is crucial to the RVM business model, this large improvement of the fraud detection rate is remarkable.

The result agrees with the results from [44], as the network learns to accept modality-matched inputs and reject negative correspondence inputs. Fraud inputs have negative correspondence modalities since visually similar inputs have the visual appearance of various classes but do not have matched ultrasonic or weight inputs.

6.3.4. Final Model

Finally, we combined all the techniques. The last row of Table 5 is the final model we propose, in which all the proposed techniques are used. It achieved high accuracy with a high fraud detection rate. When correspondence learning was used, the fraud detection rate became very high, while the target class accuracy was the most compromised value. The attention mechanism improved the fraud detection rate while maintaining the target class accuracy. We argue that this was due to better feature learning resulting from the multi-modal attention mechanism. When the two techniques were combined, the final model preserved high accuracy while detecting most of the fraud inputs. This is a remarkable improvement since fraud examples are hard to distinguish using only visual modality only, as shown in Figure 5 and Table 4.

Lower target class accuracy may be a concern, but a slight compromise is not a problem in the reverse vending machine task. Most mis-classifications are classified as negative correspondence, and users will be asked to try again. As the single trial accuracy is 94%, the success rate of multiple trials is high.

7. Conclusions

In this work, we propose correspondence learning (CL) for multi-modal object recognition tasks. The block diagram in Figure 6 shows a general architecture of a multi-modal recognition system with correspondence learning. In such systems, there are two benefits of using CL: first, it can efficiently improve the recognition performance by learning the cross-model relationship throughout the correspondence; second, the learned correspondence can be used to effectively filter out fraud inputs. When improving the overall performance, CL can be treated as an auxiliary task during training and can be removed during inference, so there will be no extra cost for inference. When fraud inputs should be detected, a minimal branch will be added at the very end of the network, so the inference overhead is only a small multilayer perceptron.

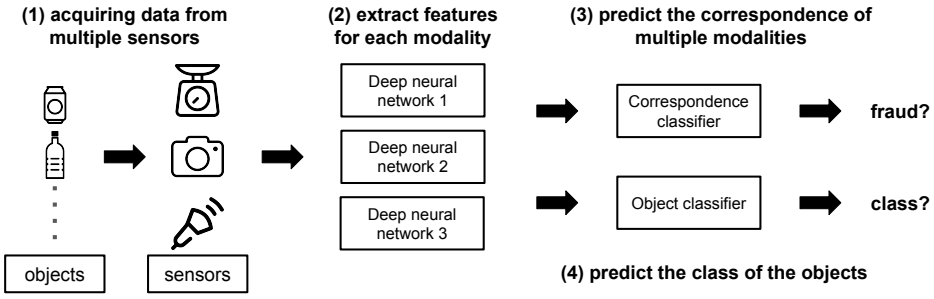

Figure 6. Block diagram of the multi-modal system with correspondence learning.

The efficacy of CL in the two use-cases is empirically validated. First, we add CL to state-of-the-art methods in sentimental analysis, where multiple heterogeneous modalities are used. In the CMU-MOSI [20] and CMU-MOSEI [29] datasets, there are consistent performance improvements across multiple baselines [22,23,41] and datasets. Second, we collect a dataset for garbage classification and show the learned correspondence can effectively filter out real-world fraud inputs. Since no previous works clearly show that non-contact ultrasonic inputs can be used for material classification, we collected raw materials and validated that the non-contact ultrasonic inputs contain sufficient information for garbage classification. Next, we collected real-world samples comprising three target classes (can/PET/glass) and two types of fraud inputs (visually similar and out-of-distribution samples). In this dataset, the fraud detection rate of the baseline (without CL) was very low (20.7% for out-of-distribution (OOD) and 21.7% for VS); with CL, we were able to effectively identify both types of fraud inputs (93.1% and 91.7%, respectively).

Other than the improved performance, the advantages of CL are two-fold: first, the proposed CL is lightweight and simple—CL can be easily integrated into any DNN-based multi-modal systems and can be jointly trained in an end-to-end manner; second, the high-performance fraud detection feature can be trained without any extra data collection for fraud samples.

There are several limitations of the proposed CL which lead to future research directions. First, CL only exploits the mutual information contained among the input modalities. While the proposed CL encourages the feature extractors to learn the mutual information among modalities, encouraging non-mutual information can further improve the recognition performance by fully exploiting the each modality information. Second, in this work, we only used real-world inputs for the fraud detection purpose. Recent adversarial attack methods [33–35] have identified the vulnerability of deep neural networks, and methods such as virtual adversarial training [45] have shown that adversarial inputs can be used to improve the performance. As a future direction, correspondence learning can be extended in combination with adversarial inputs to improve the robustness of the whole system.

8. Relevance to Electronics Journal

This manuscript is submitted for the special issue of "Deep Learning Based Object Detection II". As stated in the instruction, "This Special Issue will cover the most recent technical advances in all deep learning-based object recognition aspects", and the topics include "Sensor fusion for object detection using deep learning", and "Semi-supervised learning for object detection". Our proposed method is a generally applicable technique for all deep learning tasks with multiple sensor inputs, and it can be seen as a semi-supervised learning technique in which the learning signal comes from the cross-modal correspondence. In conclusion, we consider that this manuscript fits the purpose of the special issue.

Author Contributions: Conceptualization, J.P.; methodology, J.P.; software, J.P.; validation, J.P.; formal analysis, J.P.; investigation, J.P.; resources, J.P., M.-H.K. and D.-G.C.; data curation, M.-H.K.; writing—original draft preparation, J.P.; writing—review and editing, M.-H.K. and D.-G.C.;

visualization, J.P.; supervision, D.-G.C.; project administration, D.-G.C.; funding acquisition, D.-G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant from Defense Acquisition Program Administration and Agency for Defense Development (UD180045RD) and Hanbat National University (No.202003290001).

Data Availability Statement: The sentimental analysis datasets (CMU-MOSI, CMU-MOSEI) used in the manuscript are publicly available at https://github.com/A2Zadeh/CMU-MultimodalSDK, accessed on 23 February 2021. The garbage classification dataset is not publicly available.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
2. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
3. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
4. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 740–755.
5. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; Parikh, D. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
6. Chen, T.; Li, M.; Li, Y.; Lin, M.; Wang, N.; Wang, M.; Xiao, T.; Xu, B.; Zhang, C.; Zhang, Z. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. *arXiv* **2015**, arXiv:1512.01274.
7. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2019; pp. 8024–8035.
8. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: tensorflow.org (accessed on 23 February 2021).
9. Google Cloud Vision. Available online: https://cloud.google.com/vision/ (accessed on 21 December 2017).
10. Papago. Available online: https://papago.naver.com/ (accessed on 21 December 2017).
11. John, V.; Mita, S. Deep Feature-Level Sensor Fusion Using Skip Connections for Real-Time Object Detection in Autonomous Driving. *Electronics* **2021**, *10*, 424. [CrossRef]
12. Choi, Y.; Kim, N.; Hwang, S.; Park, K.; Yoon, J.S.; An, K.; Kweon, I.S. KAIST Multi-spectral Day/Night Dataset for Autonomous and Assisted Driving. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 934–948. [CrossRef]
13. Bednarek, M.; Kicki, P.; Walas, K. On Robustness of Multi-Modal Fusion—Robotics Perspective. *Electronics* **2020**, *9*, 1152. [CrossRef]
14. Bodapati, J.D.; Naralasetti, V.; Shareef, S.N.; Hakak, S.; Bilal, M.; Maddikunta, P.K.R.; Jo, O. Blended Multi-Modal Deep ConvNet Features for Diabetic Retinopathy Severity Prediction. *Electronics* **2020**, *9*, 914. [CrossRef]
15. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
16. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.
17. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
18. Monfort, M.; Andonian, A.; Zhou, B.; Ramakrishnan, K.; Bargal, S.A.; Yan, T.; Brown, L.; Fan, Q.; Gutfruend, D.; Vondrick, C.; et al. Moments in Time Dataset: One million videos for event understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 502–508. [CrossRef] [PubMed]
19. Kaiser, L.; Gomez, A.N.; Shazeer, N.; Vaswani, A.; Parmar, N.; Jones, L.; Uszkoreit, J. One model to learn them all. *arXiv* **2017**, arXiv:1706.05137.
20. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv* **2016**, arXiv:1606.06259.
21. Zadeh, A.; Liang, P.P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.

22. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 1103–1114. [CrossRef]
23. Shenoy, A.; Sardana, A. Multilogue-Net: A Context Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation. *arXiv* **2020**, arXiv:2002.08267.
24. Li, A.; Tan, Z.; Li, X.; Wan, J.; Escalera, S.; Guo, G.; Li, S.Z. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. *arXiv* **2020**, arXiv:2003.05136.
25. Zhang, S.; Wang, X.; Liu, A.; Zhao, C.; Wan, J.; Escalera, S.; Shi, H.; Wang, Z.; Li, S.Z. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 919–928.
26. Video Shows Galaxy S8 Facial Recognition Tricked By A Photo. Available online: http://www.gizmodo.co.uk/2017/03/video-shows-galaxy-s8-facial-recognition-tricked-by-a-photo/ (accessed on 4 February 2018).
27. Arandjelović, R.; Zisserman, A. Objects that Sound. *arXiv* **2017**, arXiv:1712.06651.
28. Park, J.; Kim, M.H.; Choi, S.; Kweon, I.S.; Choi, D.G. Fraud detection with multi-modal attention and correspondence learning. In Proceedings of the 2019 International Conference on Electronics, Information, and Communication (ICEIC), Auckland, New Zealand, 22–25 January 2019; pp. 1–7.
29. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2236–2246. Volume 1: Long Papers.
30. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *arXiv* **2014**, arXiv:1406.2199.
31. Crasto, N.; Weinzaepfel, P.; Alahari, K.; Schmid, C. Mars: Motion-augmented rgb stream for action recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7882–7891.
32. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
33. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial Examples in the Physical World. In Proceedings of the ICLR Workshop, Toulon, France, 24–26 April 2017.
34. Athalye, A.; Engstrom, L.; Ilyas, A.; Kwok, K. Synthesizing Robust Adversarial Examples. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
35. Brown, T.; Mane, D.; Roy, A.; Abadi, M.; Gilmer, J. Adversarial Patch. *arXiv* **2017**, arXiv:1712.09665.
36. Aytar, Y.; Vondrick, C.; Torralba, A. Soundnet: Learning sound representations from unlabeled video. *arXiv* **2016**, arXiv:1610.09001.
37. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 1321–1330.
38. TOMRA. Available online: https://www.tomra.com/en/ (accessed on 20 December 2017).
39. RVM Systems. Available online: http://www.reversevending.co.uk/ (accessed on 4 February 2018)..
40. Superbin. Available online: http://www.superbin.co.kr/new/index.php (accessed on 20 December 2017).
41. Hou, M.; Tang, J.; Zhang, J.; Kong, W.; Zhao, Q. Deep multimodal multilinear fusion with high-order polynomial pooling. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 12136–12145
42. Ohtani, K.; Baba, M. A Simple Identification Method for Object Shapes and Materials Using an Ultrasonic Sensor Array. In Proceedings of the 2006 IEEE Instrumentation and Measurement Technology Conference Proceedings, Sorrento, Italy, 24–27 April 2006; pp. 2138–2143. [CrossRef]
43. Moritake, Y.; Hikawa, H. Category recognition system using two ultrasonic sensors and combinational logic circuit. *Electron. Commun. Jpn. Part III Fundam. Electron. Sci.* **2005**, *88*, 33–42. [CrossRef]
44. Arandjelović, R.; Zisserman, A. Look, Listen and Learn. In Proceedings of the IEEE International Conference on Computer Vision,Venice, Italy, 22–29 October 2017.
45. Miyato, T.; Maeda, S.i.; Koyama, M.; Ishii, S. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1979–1993. [CrossRef] [PubMed]