MMTM: Multimodal Transfer Module for CNN Fusion

Hamid Reza Vaezi Joze*
Microsoft
hava@microsoft.com

Amirreza Shaban*†
Georgia Tech
amirreza@gatech.edu

Michael L. Iuzzolino†
CU Boulder
michael.iuzzolino@colorado.edu

Kazuhito Koishida
Microsoft
kazukoi@microsoft.com

Abstract

In late fusion, each modality is processed in a separate unimodal Convolutional Neural Network (CNN) stream and the scores of each modality are fused at the end. Due to its simplicity, late fusion is still the predominant approach in many state-of-the-art multimodal applications. In this paper, we present a simple neural network module for leveraging the knowledge from multiple modalities in convolutional neural networks. The proposed unit, named Multimodal Transfer Module (MMTM), can be added at different levels of the feature hierarchy, enabling slow modality fusion. Using squeeze and excitation operations, MMTM utilizes the knowledge of multiple modalities to recalibrate the channel-wise features in each CNN stream. Unlike other intermediate fusion methods, the proposed module could be used for feature modality fusion in convolution layers with different spatial dimensions. Another advantage of the proposed method is that it could be added among unimodal branches with minimum changes in the their network architectures, allowing each branch to be initialized with existing pretrained weights. Experimental results show that our framework improves the recognition accuracy of well-known multimodal networks. We demonstrate state-of-the-art or competitive performance on four datasets that span the task domains of dynamic hand gesture recognition, speech enhancement, and action recognition with RGB and body joints.

1. Introduction

Different sensors can provide complementary information about the same context. Multimodal fusion is the act of extracting and combining relevant information from the different modalities that leads to improved performance over using only one modality. This technique is widely used in various machine learning tasks, such as video classification [1, 2], action recognition [3], emotion recognition [4, 5], and audio visual speech enhancement [6, 7].

*Equal contribution.
†Work done during an internship at Microsoft.

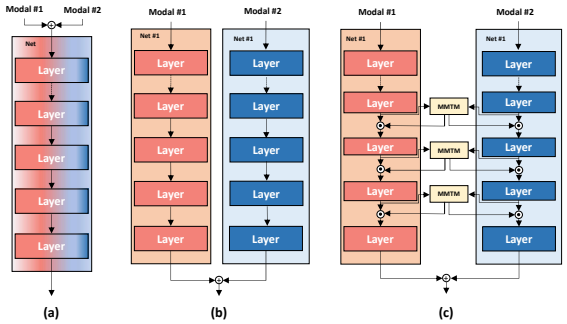

Figure 1. (a) early fusion (b) late fusion (c) intermediate fusion with Multimodal Transfer Module (MMTM). MMTM operates between CNN streams and uses information from different modalities to recalibrate channel-wise features in each modality.

In general, fusion can be achieved at the input level (i.e. early fusion), decision level (i.e. late fusion), or intermediately [8]. Although studies in neuroscience [9, 10] and machine learning [1, 3] suggest that mid-level feature fusion could benefit learning, late fusion is still the predominant method utilized for mulitmodal learning [11–13]. This is mostly due to practical reasons. For example, a simple pooling operator [14, 15] or an attention mechanism [16] can be used to fuse 1-dimensional prediction scores of each stream. However, intermediate level features of different modalities have different or unaligned spatial dimensions making the intermediate fusing more challenging. Another reason for the popularity of late fusion is that the architecture of each unimodal stream is carefully designed over years to achieve state-of-the-art performance for each modality. This also enables the CNN streams of a multimodal framework to be initialized by weights that have been pretrained with a large number of unimodal training samples. However, intermediate fusion requires major changes in the base network architecture, which complicates the use of pretrained weights in most cases and requires the network to be retrained from randomly initialized states [17, 18]. Figure 1 illustrates three common multimodal fusion techniques.

The goal of the proposed method is to overcome the aforementioned problems of intermediate fusion. Inspired by the squeeze and excitation (SE) module [19] for unimodal convolutional neural networks, we propose a multimodal transfer module to recalibrate the channel-wise fea-

tures of different CNN streams. MMTMs can be inserted into intermediate levels of any late fusion backbone architecture. Each MMTM has two units: a) a multimodal squeeze unit that receives the features from all modalities at a given level of representation across the branches, generating a global joint representation of these features, and b) an excitation unit that uses this joint representation to adaptively emphasize on more important features and suppress less important ones in all modalities. The squeeze unit aggregates spatial dimensions, allowing information with global receptive fields from all modalities to be used in the global representation. It also enables learning a joint representation from modalities with different spatial dimensions.

Although the module design is generic and could potentially be added at any level in the network hierarchy, the optimal locations and number of modules are different for each application. We design application specific networks for gesture recognition, audio-visual speech enhancement, and action recognition tasks and study the benefit of adding MMTM in their architectures. We make the following empirical observations from these applications. Firstly, adding MMTM to intermediate and high-level features is beneficial, whereas the same is not true about low-level features. We believe that is because intera-modality correlation in low-level features is lower compared to intermediate and high-level features. This is also highlighted in previous research [20]. Secondly, even in gesture recognition where RGB and depth modalities are spatially aligned and fusion can be done without the squeeze operation, squeezing considerably improves the performance by providing information with a global receptive field. Lastly, excitation by gating operation outperforms the sum operation that is usually used in residual learning, highlighting the importance of the emphasis and suppression mechanisms.

In summary, this paper makes the following contributions: First, we propose a new neural network module called MMTM to fuse knowledge from intermediate features of unimodal CNNs. Second, we design different network architectures for three different multimodal fusion applications: gesture recognition using multiple visual modalities, audio-visual speech enhancement, and action recognition with RGB and body joints. We demonstrate through experiment on these tasks that MMTM improves the performance beyond the late fusion approach.

2. Related Work

In late fusion, the prediction of each unimodal stream are fused to make the final prediction. Fusion can be via element-wise summation, a weighted average [15], a bilinear product [21], or a more sophisticated rank minimization [22] method. Another approach to late fusion utilizes attention to pick the best expert for each input signal [16]. The gated multimodal units [23] extends this method by enabling gating at intermediate feature levels. More recently, Hu *et al*. [24] propose a dense multimodal intermediate fusion network for hierarchical joint feature learning. Similar to [23], the dense fusion operator in [24] assumes identical spatial dimensions for different streams. Despite the similarity of these approaches to our work, their applicability is limited to layers where the multimodal features' spatial dimensions are the same, or at the very end of the network where spatial dimensions are already aggregated. The squeeze operation proposed in this work allows the fusion of modalities with different spatial dimensions at any level of the feature hierarchy.

In a related multimodal learning topic, called cross-modal learning, information from multiple modalities are used to improve the learning performance within any individual modality. It is assumed that data from all the modalities are present during training but the performance is tested on only one modality [25]. MTUT [12] uses spatiotemporal semantic alignment loss to improve the performance of each stream in gesture recognition. We believe cross-modal learning approaches are orthogonal to our work since the improved unimodal networks learned by these methods can initialize weights of the CNN streams in our model.

Multimodal Action Recognition in Videos Video [1, 14, 26] and skeleton [11, 27, 28] modalities have been extensively used for the action recognition task. Each of these approaches have their own drawbacks. With the lack of explicit human body model, video based action recognition methods deal poorly with background clutter and non-action movements [11]. On the other hand, by solely relying on body pose most of contextual and global cues present in the video will be lost. Recent methods develop architectures to fuse these modalities to further improve the performance of action recognition. In [28], an end-to-end trainable multitask network for joint pose estimation and action recognition is proposed. PoseMap [11] utilizes a two stream network to process spatiotemporal pose heatmaps and skeleton separately, and uses late fusion for the final prediction. A bilinear pooling block that separately pools input features in modality and time directions is employed in [29].

Audio-Visual Speech Enhancement (AVSE) Work in AVSE is strongly motivated by the cocktail party effect, which refers to humans' ability to selectively attend to auditory signals of interest within a noisy environment. Experiments in neuroscience have demonstrated that cross-modal integration of audio-visual signals may improve the perceptual quality of the targeted acoustic signal [30–32]. Inspired by the results from biological research, recent studies focus on augmenting audio only speech enhancement methods with visual information, such as lip movement. State-of-the-art results have been achieved by recent AVSE models that use deep neural networks [6, 7, 33, 34]. The predominant approach taken for AV fusion is late fusion [13],

where the audio and visual information is processed separately then integrated at a singular point via channel-wise concatenation.

Hand Gesture Recognition Interpreting hand gestures via machine learning algorithms is significantly important in human-computer interaction. We review the 3D convolutional neural network based gesture recognition algorithms [35–39] that are designed for processing time series data among other branches [40–43]. In [35], a novel 3D CNN is proposed to integrate depth and image gradient values to recognize dynamic hand gestures. Molchanov *et al*. [36] employ a multistream 3D CNN to fuse streams of data from multiple sensors including short-range radar, color, and depth sensors for recognition. A real-time method is presented in [37] to simultaneously detect and classify gestures in videos. Camgoz *et al*. [38] present a late fusion approach for fusing the scores of unimodal 3D CNN streams. Miao *et al*. propose ResC3D [39], a 3D CNN architecture that combines multimodal data using an attention model. MFFs [44] develops a data level fusion method for RGB and optical flow. FOANet [45] proposes a sparse fusion technique for hand gesture recognition. FOANet decomposes each input modality (RGB, depth, and 2 types of optical flow) into separate focus channels (global, right hand, left hand) and processes each of these 12 focus channels in an independent unimodal network. Finally, it learns a sparsely connected late fusion network to avoid overfitting. Unlike our method, FOANet relies on the output of a detector to find the focus areas in the video.

Squeeze and Excitation (SE) Network [19] Our proposed method can be seen as a generalization to the SE module, which is proposed for unimodal deep neural networks. The SE modules uses self excitation to adaptively recalibrate channel-wise feature responses. Our work adopts the SE module for multimodal feature recalibrations.

3. Multimodal Transfer Module

In this section, we discuss the simplest case of fusion between two disjoint CNN streams, CNN_1 and CNN_2. Let $\mathbf{A} \in \mathbb{R}^{N_1 \times \cdots \times N_K \times C}$ and $\mathbf{B} \in \mathbb{R}^{M_1 \times \cdots \times M_L \times C'}$ represent the features at a given layer of CNN_1 and CNN_2, respectively. Here, N_i and M_i represent the spatial dimensions[1], and C and C' represent the number of channels of the corresponding features in CNN_1 and CNN_2 respectively. MMTM receives features $\mathbf{A}$ and $\mathbf{B}$ as input, learns a global multimodal embedding from them, and uses this embedding to recalibrate the input features. This is done in a two-step multimodal squeeze and excitation process described below.

[1]In general, it is possible to have more than two (e.g. time dimension in 3D convolutions could be treated as a spatial dimension) or no (e.g. fully connected layers) spatial dimensions.

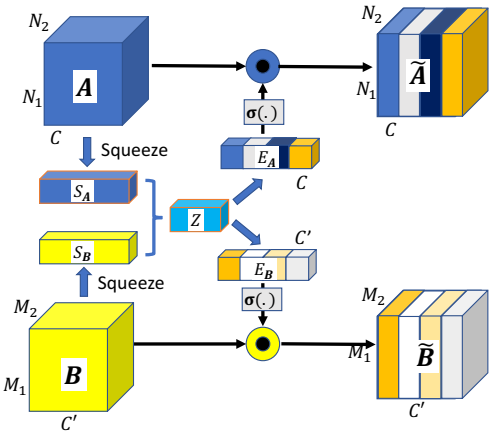

Figure 2. Architecture of MMTM for two modalities. $\mathbf{A}$ and $\mathbf{B}$, that represent the features at a given layer of two unimodal CNNs, are the inputs to the module. For better visualization we limit the number of their spatial dimensions to 2. MMTM uses squeeze operations to generate global feature descriptor from each tensor. Both tensors are map into a joint representation Z by using concatenation and fully-connected layer. The excitation signals $E_{\mathbf{A}}$ and $E_{\mathbf{B}}$ are generated using the joint representation. Finally the excitation signals are used to gate the channel-wise features in each modality.

Squeeze The information in the output features of convolution layers are limited by the size of their receptive fields and lacks global context. As suggested by [19], we first squeeze the spatial information into the channel descriptors via a global average pooling over spatial dimensions of the input features:

$$S_{\mathbf{A}}(c) = \frac{1}{\prod_{i=1}^{K} N_i} \sum_{n_1,\ldots,n_K} \mathbf{A}(n_1,\ldots,n_K,c) \quad (1)$$

$$S_{\mathbf{B}}(c) = \frac{1}{\prod_{i=1}^{L} M_i} \sum_{m_1,\ldots,m_L} \mathbf{B}(m_1,\ldots,m_L,c). \quad (2)$$

Importantly, the squeeze operation enables fusion between modalities with features of arbitrary spatial dimension. Note that while we use simple average pooling, more sophisticated pooling methods could be used at this step.

Multimodal Excitation The function of this unit is to generate the excitation signals, $E_{\mathbf{A}} \in \mathbb{R}^{C}$ and $E_{\mathbf{B}} \in \mathbb{R}^{C'}$, which can be used to recalibrate the input features, $\mathbf{A}$ and $\mathbf{B}$, by a simple gating mechanism:

$$\tilde{\mathbf{A}} = 2 \times \sigma(E_{\mathbf{A}}) \odot \mathbf{A}$$
$$\tilde{\mathbf{B}} = 2 \times \sigma(E_{\mathbf{B}}) \odot \mathbf{B},$$

where $\sigma(.)$ is the sigmoid function and $\odot$ is the channel-wise product operation. This allows the suppression or excitation of different filters in each stream. Note that the MMTM weights are regularized in order to control the proximity of $E_{\mathbf{A}}$ and $E_{\mathbf{B}}$ to zero. Specifically, increasing

the regularization weight of $E_{\mathbf{A}}$ pushes the gating signal $2 \times \sigma(E_{\mathbf{A}})$ closer to the identity vector, limiting the effect of gating on feature $\mathbf{A}$.

The gating signals must apply different calibration weights to different modalities based on the same input representation. We achieve this by first predicting a joint representation $Z \in \mathbb{R}^{C_Z}$ from the squeezed signals

$$Z = \mathbf{W}[S_{\mathbf{A}}, S_{\mathbf{B}}] + b, \tag{3}$$

and then predicting excitation signals for each modality through two independent fully-connected layers

$$E_{\mathbf{A}} = \mathbf{W}_{\mathbf{A}} Z + b_{\mathbf{A}}, \quad E_{\mathbf{B}} = \mathbf{W}_{\mathbf{B}} Z + b_{\mathbf{B}}. \tag{4}$$

Here, $[\cdot, \cdot]$ represents the concatenation operation, $\mathbf{W} \in \mathbb{R}^{C_Z \times (C+C')}, \mathbf{W}_{\mathbf{A}} \in \mathbb{R}^{C \times C_Z}, \mathbf{W}_{\mathbf{B}} \in \mathbb{R}^{C' \times C_Z}$ are the weights, and $b \in \mathbb{R}^{C_Z}, b_{\mathbf{A}} \in \mathbb{R}^{C}, b_{\mathbf{B}} \in \mathbb{R}^{C'}$ are the biases of the fully connected layers. As suggested in [19], we use $C_Z = (C + C')/4$ to limit the model capacity and increase the generalization power. For fusing more than two modalities, we simply generalize this approach by concatenating squeezed features from all the modalities in Equation 3 and predict excitation signals for each modality with an independent fully-connected layer like in Equation 4.

Learning the joint representation in this way allows the features of one modality to recalibrate the features of another modality. For instance, in gesture recognition when a gesture is blurry in RGB camera and more apparent in depth modality, MMTM cross-modal recalibration affords more efficient processing in the RGB stream. Figure 2 summarizes the overall architecture of the proposed MMTM.

4. Applications

The MMTM is generic and can be easily integrated to any multimodal CNN architectures. In this section, we explore a few applications that can benefit from MMTM and describe the architecture changes necessary to support multimodal fusion. We evaluate the performance of the proposed multimodal models in the experiment section.

4.1. Hand Gesture Recognition

Hand gesture recognition is a video classification task. It is shown that complementary sensory information, such as depth and optical flow, improves the performance of the gesture recognition [12, 37, 41, 44]. There are multiple multimodal datasets available for this task [37, 41, 46, 47] and several previous fusion methods have reported their results on these datasets [36–39].

We design a gesture recognition network for fusing RGB, depth, and optical flow video streams via MMTM. To process the temporal inputs, we use I3D network architecture [48] with an inflated inception-v1 [49] backbone for all the streams. In I3D network, convolution and pooling kernels of the backbone network are expanded into 3D, enabling efficient spatial-temporal feature processing. We apply MMTM after the last 6 inception modules (the connectivity is similar to figure 1). Note that the output of 3D convolutions has a time dimension in addition to height, width, and channel dimensions. We empirically find that the best performance is achieved when the squeeze operation is applied over all the dimensions except for the channel dimension.

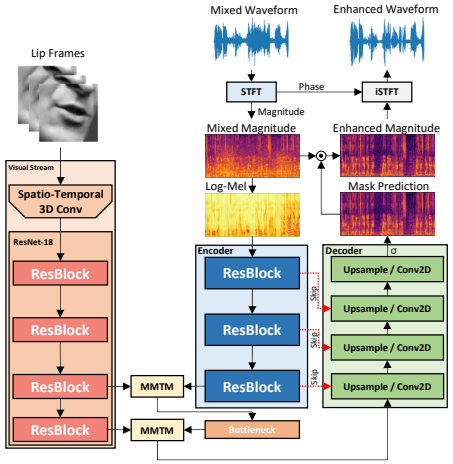

Figure 3. An overview of our AVSE architecture.

4.2. Audio-Visual Speech Enhancement

The predominant method for AV speech enhancement combines audio and visual signals via channel-wise concatenation (CWC) using the late fusion approach. As an application of MMTM, we explore AV fusion for speech enhancement tasks using MMTM instead of the CWC-based late fusion. Model details are provided below, and an overview of our AVSE architecture can be found in Figure 3.

Visual Network We use the spatio-temporal residual network proposed by [50], which consists of a 3D spatio-temporal convolution followed by a 2D ResNet-18 [51]. Processing 3D features in a 2D convolution operation is achieved by packing the temporal dimension, t, into the batch dimension. The network is randomly initialized and trained concurrently with the AVSE task.

Audio Network Our audio network is an autoencoder with skip connections; we follow the design detailed in [52]. Figure 3 (top) depicts the audio processing strategy, which follows the audio processing procedures of [6] and is detailed in Section 5.2. The network takes a log-mel mixture magnitude spectrogram, log-mel(X_{mix}), as input and outputs the predicted ideal ratio mask, M. The enhanced magnitude spectrogram, X_{enh}, is obtained via $X_{enh} = M \odot X_{mix}$, where $\odot$ denotes element-wise multiplication. The network is trained by minimizing the reconstruction loss between the enhanced magnitude, X_{enh}, and the target magnitude,

X_{spec}, where X_{spec} is obtained via short-time Fourier transform (STFT) from the target waveform. The optimization objective is given by $\mathcal{L} = ||X_{enh} - X_{spec}||_1$.

Audio-Visual Fusion via MMTM Let F_a^j denote the audio feature at layer j of the autoencoder with $F_a^j \in \mathbb{R}^{b \times t \times f \times c_a}$, where b, t, f, and c_a are the batch, temporal, frequency, and audio channel dimensions, respectively. Let F_v^i denote visual feature at layer i of the visual network's ResNet-18 with $F_v^i \in \mathbb{R}^{b \cdot t \times h \times w \times c_v}$, where h, w are spatial dimensions and b, t, c_v are the batch, temporal, and visual channel dimensions, respectively. We unpack t from the batch dimension of F_v^i via reshaping such that $F_v^i \in \mathbb{R}^{b \times t \times h \times w \times c_v}$. The MMTM takes F_a and F_v as input and carries out the fusion procedure detailed in Section 3. For AVSE, the final output is from audio tower; consequently, MMTM does not gate on visual network.

4.3. Human Action Recognition

Recent methods in human activity recognition combine video and 3D pose information to further improve the performance of action recognition [11, 28, 29]. Following the same approach, we utilize MMTM for intermediate fusion between a visual and a skeleton based network. Similar to the gesture recognition application, we use I3D for the RGB video stream and HCN, as suggested by [53], for the skeletal stream. Although HCN is not the sate-of-the-art for skeleton-based action recognition, the simplicity of its design makes it suitable for our approach.

As it is shown in Figure 4, HCN is comprised of two 2D convolution subnetworks: one branch processes the raw skeleton data, and the other branch processes the motion–the temporal gradients of the skeletal data. The two subnetworks are fused via channel-wise concatenation and followed by two convolution operations (conv5 and conv6), and finally, a fully connected layer (fc7).

Figure 4 illustrates the complete network we are proposing. We add 3 MMTMs that receive inputs from last three inception modules of the I3D and conv5, conv6, and fc7 of HCN network. Let $\mathbf{A} \in \mathbb{R}^{t \times w \times h \times C}$ represent an I3D feature, where t represents temporal dimension and w, h are the spatial dimensions. Let $\mathbf{B} \in \mathbb{R}^{t \times n \times C'}$ represent HCN features after conv5 and conv6 layers, where t is the temporal dimension and n is the body-joints dimension. The output of the fully connected layer (fc7) in HCN network is a 1-dimensional vector with no spatial dimension. In MMTMs, we aggregate all the dimensions of the inputs $\mathbf{A}$ and $\mathbf{B}$ except the channels. The dimensions of the I3D and HCN features sent to the MMTMs ($\mathbf{A}$ and $\mathbf{B}$) do not match, but MMTM's squeezing operation makes the fusion possible.

5. Experimental Results

In this section, we evaluate the performance of the proposed method in gesture recognition, speech enhancement, and action recognition tasks. Due to the large number of experiments, we use a simple rule to decide the number of MMTMs in each architecture without an extensive architecture tuning scheme. We use MMTMs after each module in the second half of the network with minimum depth. This is 6 MMTMs for hand gesture recognition experiments, 2 in speech enhancement, and 3 in action recognition experiment. Refer to Section 5.4 for the study of the number of MMTMs in hand gesture recognition task.

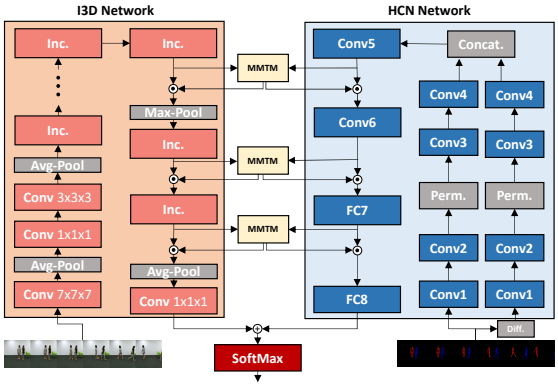

Figure 4. Proposed multimodal architecture for action recognition. Each "Inc." block represents an inception module described in [48].

5.1. Hand Gesture Recognition

In this section, we evaluate our method against state-of-the-art dynamic hand gesture methods. We conduct experiments on two recent publicly available multimodal dynamic hand gesture datasets: *EgoGesture* [41, 46] and *NVGestures* [37] datasets. Figure 5 (a), (b) shows sample frames from the different modalities of these datasets.

Implementation Details: In the design of our method, we adopt the architecture of I3D network [48] as the backbone network for each modality. The architecture details can be found in Section 4.1. We start with the publicly available ImageNet [56] + Kinetics [57] pretrained networks for all of our experiments on I3D. We optimize the objective function with the standard SGD optimizer using a momentum of 0.9. We start with the base learning rate of 10^{-2} and reduce it $10\times$ when the loss is saturated. We use a batch size of 4 containing 64-frames (32-frames for *EgoGesture*) snippets in the training stage. We employ the following spatial and temporal data augmentations during the training stage. For spatial augmentation, videos are resized to 256×256 pixels, and then randomly cropped with a 224×224 patch. The resulting video is randomly flipped horizontally. For temporal augmentation, 64 consecutive frames are picked randomly from the videos. Shorter videos are zero-padded on both sides to obtain 64 frames. During testing, we use 224×224 center crops, apply the models over the full video, and average the predictions.

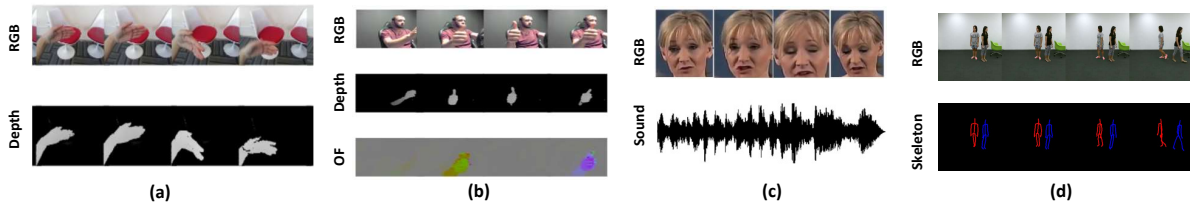

Figure 5. Sample sequences from multimodal datasets: (a) EgoGesture [41] (b) NVGesture [37] (c) VoxCeleb2 [54] (d) NTU-RGBD [55]

Method	Input Modalities	Accuracy
I3D [48]	RGB	90.33
I3D [48]	Depth	89.47
VGG16 [58]	RGB+Depth	66.5
VGG16 + LSTM [59]	RGB+Depth	81.4
C3D [60]	RGB+Depth	89.7
C3D+LSTM+RSTTM [41]	RGB+Depth	92.2
I3D late fusion [48]	RGB+Depth	92.78
Ours	RGB+Depth	**93.51**

Table 1. Accuracies of different multimodal fusion hand gesture methods on the EgoGesture dataset [41].

5.1.1 EgoGesture Dataset

EgoGesture dataset [41, 46] is a large multimodal hand gesture dataset collected for the task of egocentric gesture recognition. This dataset contains $24,161$ hand gesture clips with 83 gesture classes being performed by 50 subjects. Videos in this dataset include both static and dynamic gestures captured with an Intel RealSense SR300 device in RGB-D modalities across multiple indoor/outdoor scenes.

We assess the performance of our method along with various hand gesture recognition methods published. Table 1 compares unimodal test accuracies for I3D on separate modalities and test accuracies of different hand gesture methods by fusion of RGB and depth. VGG16 [58] processes each frame independently and VGG16+LSTM [59] combines this method with a recurrent architecture to leverage the temporal information. As can be seen, the 3D CNN-based methods, C3D [60], C3D+LSTM+RSTMM [41], and I3D [48] outperform the VGG16-based methods. However, among the 3D CNN architectures, our method outperforms the top performers I3D late fusion by 0.73%.

5.1.2 NVGesture Dataset

NVGestures dataset [37] was captured with multiple sensors for studying human-computer interfaces. It contains 1532 dynamic hand gestures recorded from 20 subjects inside a car simulator with artificial lighting conditions. This dataset includes 25 classes of hand gestures. The gestures were recorded with SoftKinetic DS325 device as the RGB-D sensor and DUO-3D for the infrared streams. In addition, the optical flow and infrared disparity map modalities are usually used to enhance the prediction results. Following the previous works [37, 44], we only use RGB, depth, and optical flow modalities in our experiments. The optical flow is calculated using the method presented in [61]. The RGB and optical flow modalities are well-aligned in this dataset, however, the depth map includes a larger field of view.

Method	Input Modalities	Accuracy
I3D [48]	RGB	78.42
I3D [48]	Opt. flow	83.19
I3D [48]	Depth	82.28
HOG+HOG2 [64]	RGB+Depth	36.9
I3D late fusion [48]	RGB+Depth	84.43
Ours	RGB+Depth	**86.31**
Two Stream CNNs [14]	RGB+Opt. flow	65.6
iDT [62]	RGB+Opt. flow	73.4
R3DCNN [37]	RGB+Opt. flow	79.3
MFFs [44]	RGB+Opt. flow	84.7
I3D late fusion [48]	RGB+Opt. flow	84.43
Ours	RGB+Opt. flow	**84.85**
R3DCNN [37]	RGB+Depth+Opt. flow	83.8
I3D late fusion [48]	RGB+Depth+Opt. flow	85.68
Ours	RGB+Depth+Opt. flow	**86.93**
Human [37]		88.4

Table 2. Accuracies of different multimodal fusion hand gesture methods on the NVGesture dataset [37].

Table 2 presents the results of our method in comparison with the recent state-of-the-art methods: HOG+HOG2, improved dense trajectories (iDT) [62], R3DCNN [37], two-stream CNNs [14] and MFFs [44]. We also report human labeling accuracy for comparison. The iDT [62] method is often recognized as the best performing method with hand-engineered features [63]. Similar to the previous experiment, we observe that the 3D-CNN-based methods outperform other hand gesture recognition methods, and among them, Our method provides the top performance in all the modalities. FOANet [45] method achieves 91.28% on this dataset using a sparse fusion method. However, this result is not comparable with the methods in Table 2 since FOANet relies on a separate pre-trained network to detect the hand.

5.2. Audio-Visual Speech Enhancement

In this section, we evaluate our MMTM method on audio-visual speech enhancement. Using PESQ and STOI objective measures, we demonstrate that our slow fusion MMTM method outperforms state-of-the-art late fusion, channel-wise concatenation AVSE approaches. We use *VoxCeleb2* [54], a large audio-visual dataset obtained from YouTube that contains over 1 million utterances for 6,112 celebrities. The training, validation, and test datasets are split by celebrity ID (CID) such that the sets are disjoint

Method	Fusion Method	PESQ	STOI
Target	-	4.64	1.000
Mixed	-	2.19	0.900
AVSE [6]†	CWC	2.59	0.650
AO Baseline	-	2.43	0.930
AV Baseline	CWC	2.67	0.938
Ours	MMTM	**2.73**	**0.941**

Table 3. Speech enhancement evaluations on the *VoxCeleb2* dataset [54] for 3 simultaneous speakers. CWC: Channel-wise concatenation. † for approximate reference only.

over CIDs. In addition, CHiME-1/3 [65, 66], NonStationaryNoise [67], ESC50 [68], HuCorpus [69], and private datasets are used for additive noise.

Video frames are extracted at 25 FPS and S^{3}FD [70] performs face detection. Following [50], we discard redundant visual information by cropping the mouth region via facial landmarks obtained from Facial Alignment Network [71]. Lip frames are resized to 122×122, transformed to grayscale, then normalized using the global mean and variance statistics from the training set. The audio waveform is extracted from the video following the methods of [6, 72]. We specify a window length of 40ms, hop size of 10ms, and sampling rate of 16kHz to align one video frame to four audio steps. Short-time Fourier transform (STFT) with a Hanning window function converts the waveform to spectrogram, $X_{spec} \in \mathbb{R}^{T \times F}$ with a frequency resolution of $F = 321$, representing frequencies from $0 - 8$kHz.

Training samples of batch size 4 are generated on-the-fly as lip frame and spectrograms pairs, (X_{vid}, X_{spec}). Interference spectrograms, X_{inter}, are sampled from the *VoxCeleb2* set. We progressively increase the number of interference speakers during training, beginning with one and incrementing by one every 50 epochs until we reach the max of four. A noise spectrogram, X_n, is randomly sampled from the noise datasets. The mixture spectrogram is constructed via $X_{mix} = X_{spec} + \alpha X_{inter} + \beta X_n$, where α, β are mixing coefficients that achieve a specific SNR. Training and test SNRs are sampled from 0-20dB and 2.5-17.5dB ranges, respectively. X_{mix} is transformed to a log-mel representation, $\log X_{mel} \in \mathbb{R}^{T \times F}$, where $T = 116$ and $F = 80$. We augment lip frames, X_{vid}, via random cropping ($\pm$ 5 pixels) and left-right flips. Augmented frames are resized to 112×112 and fed into the visual network.

Objective evaluation results are shown in Table 3. We evaluate enhanced speech using the perceptual quality of speech quality (PESQ) [73] and the short-time objective intelligibility (STOI) [74]. The audio only (AO) model is trained without the visual network and establishes an AO speech enhancement baseline. The AV baseline model establishes a baseline for predominant AVSE approaches that perform late fusion via CWC of AV features. We closely aligned the fusion mechanism in our AV baseline model architecture to that of [6], and we matched the sample generation and training procedure as best we could given the information available. We report on [6] for reference only.

Our AVSE model outperforms the AO and AV baselines on both objective measures PESQ and STOI. We outperform the AO baseline by 0.3 PESQ and 0.01 in STOI, demonstrating that visual information improves speech enhancement performance. Further, we outperform the AV baseline with CWC fusion by 0.06 PESQ, indicating that MMTM via slow fusion affords the greatest performance improvement. Our model generalizes to speakers unseen during training since CID is disjoint across train/test sets.

Method	Input Modalities	Accuracy
HCN_{ours}	Pose	77.96
I3D [48]	RGB	89.25
DSSCA - SSLM [75]	RGB+Pose	74.86
Bilinear Learning [29]	RGB+Pose	83.0
2D/3D Multitask [28]	RGB+Pose	85.5
PoseMap [11]	RGB+Pose	91.71
Late Fusion (I3D + HCN_{ours})	RGB+Pose	91.56
Ours	RGB+Pose	**91.99**

Table 4. Accuracies of different multimodal fusion action recognition methods on the NTU-RGBD dataset [55].

5.3. Action Recognition

NTU-RGBD dataset [55] is a well-known large scale multimodal dataset. It contains $56,880$ samples captured from 40 subjects performing 60 classes of activities at 80 view-points. Each action clip includes up to two people on the RGB video as well as 25 body joints on 3D coordinate space. We followed the cross-subject evaluation [55] that splits the 40 subjects into training and testing sets. To have a fair comparison with previous works, we only use RGB and pose (skeleton) modalities. The architecture details can be found in Section 4.3. We followed section 5.1 for training settings as well as RGB data preparation and augmentation.

Table 5 shows the result of our method in comparison with the recent state-of-the-art methods on NTU-RGBD dataset. The first part of the table shows our unimodal baselines with I3D on RGB and HCN [53] on skeletons. We use 3D skeletons and follow the 32 frame subsampling method from the original paper. For simplicity in the fusion mechanism, we implemented multi-person slow fusion method [53]. Consequently, our reported accuracy on HCN is lower than the result in [53]. The second part shows state-of-the-art methods specifically design for action recognition by integrating RGB and skeleton. Our proposed fusion method outperforms all the recent action recognition algorithms. To our knowledge this is a new state-of-the-art result for RGB+Pose on the NTU-RGBD dataset [55].

Next, we use the recently released code of [17] to compare several general purpose multimodal fusion algorithms on this dataset. We implement and train the proposed method within this framework. To have an identical set-

ting with other methods, we use inflated Resnet-50 [76] for video processing and the implementation of HCN [53] provided in this framework for skeleton processing. Table 4 illustrates the performance of these unimodal networks as well as different state-of-the-art multimodal fusion methods. MFAS [17] is an architecture search algorithm that leverages a sequential architecture exploration method to find an optimal fusion architecture. In addition to the two stream CNN [14], which is a late fusion algorithm, we also report the results of two intermediate fusion algorithms Gated Multimodal Units (GMU) [23] and CentralNet [18]. Our method outperforms the state-of-the-art MFAS method without an extensive model search on this dataset. We believe this performance could be further improved by a comprehensive architecture tuning.

Method	Input Modalities	Accuracy
HCN [53]	Pose	85.24
Infalated Resnet-50 [76]	RGB	83.91
Two Stream [14]	RGB+Pose	88.60
GMU [23]	RGB+Pose	85.80
CentralNet [18]	RGB+Pose	89.36
MFAS [17]	RGB+Pose	90.04
Ours	RGB+Pose	**90.11**

Table 5. Comparison of state-of-the-art multimodal fusion algorithms on the NTU-RGBD dataset [55]. All methods use HCN and Infalated Resnet-50 backbone unimodal architectures.

5.4. Analysis of the Network

To understand the effects of some of our model choices, we explore the performance of some variations of our model on the NVGesture dataset [37]. In particular, we compare our fusion method with different architectures in the transfer layer. We also explore using a different number of transfer layers when all the implementation details are the same as RGB+Depth gesture recognition network described in Section 5.1.2.

Since the spatial dimensions are aligned in this problem, we can directly concatenate the convolutional features without squeezing them in the MMTM. In order to keep the spatial dimensions of these features across the module, we also need to change all the fully connected layers in MMTM to convolution layers with kernel size 1. This ensures that the number of parameters remains the same. We refer to this approach as convolutional MMTM. In addition, we also use a variation of the convolutional MMTM that utilizes a sum operation instead of the gating operation. This approach is closely related to residual learning [51] and has been proposed for multimodal fusion with aligned spatial dimensions [77]. Finally, we evaluate the performance of the original Squeeze and Excitation (SE) approach in which each unimodal stream uses self excitation to recalibate its own channel-wise features. The scores of these unimodal networks are fused by late fusion at the end.

Method	Accuracy	#FLOPS	#Parameters
Early Fusion	78.84	$247M$	$12.3M$
Late Fusion	84.43	$405M$	$24.6M$
Convlutional MMTM	84.43	$25.24G$	$31.6M$
Convlutional MMTM (with sum op.)	84.65	$25.24G$	$31.6M$
SE [19] + Late Fusion	85.06	$472M$	$31.6M$
MMTM	**86.31**	$472M$	$31.6M$

Table 6. Comparison of different MMTM architectures on the NVGesture dataset.

Table 6 compares the accuracy of these variations, as well as their FLOPS and number of parameters with the late fusion and MMTM. Surprisingly, the convolutional MMTM variations do not show any noticeable improvement over the late fusion method. This result highlights the importance of extracting information with global receptive field information in the squeeze unit. We also note that not using the squeeze blocks increase the number of FLOPS by about 5 times. Finally, the result of self excitation approach with no intermediate fusion clearly shows that the most of performance gain in MMTM is due to the slow fusion of the modalities rather than pure squeeze and excitation method.

As we mentioned in Section 4.1, we use MMTM after the last 6 inception modules. In the last study, we evaluate the performance of the RGB+Depth gesture recognition network with MMTM applied to a different number of inception modules. Figure 6 shows how the performance changes with respect to the number of MMTMs. This experiment indicates that the best performance is achieved when the output of half of the last inception modules (6 out of 12) are fused by MMTM. This suggests that mid-level and high-level features benefit more than low-level features from this approach.

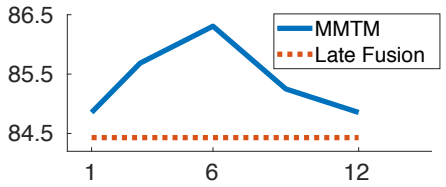

Figure 6. Accuracy vs. #MMTMs on the NVGesture dataset.

6. Conclusion

We present a simple neural network fusion module for leveraging the knowledge from multiple modalities in convolutional neural networks. The proposed module can be added at different levels of the feature hierarchy, allowing slow modality fusion. A wide range of experiments on applications with different types of modalities show applicability of the proposed module to gesture recognition, audio-visual speech enhancement, and human action recognition.

References

[1] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1, 2

[2] Xiaodong Yang, Pavlo Molchanov, and Jan Kautz. Multilayer and multimodal fusion of deep neural networks for video classification. In *International conference on multimedia*, 2016. 1

[3] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 1

[4] Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *ICASSP*, 2013. 1

[5] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. Emotion recognition using multimodal deep learning. In *NIPS*, 2016. 1

[6] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*, 2018. 1, 2, 4, 7

[7] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 1, 2

[8] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 2017. 1

[9] Charles E Schroeder and John Foxe. Multisensory contributions to low-level,unisensory processing. *Current opinion in neurobiology*, 2005. 1

[10] Emiliano Macaluso. Multisensory processing in sensory-specific cortical areas. *The neuroscientist*, 2006. 1

[11] Mengyuan Liu and Junsong Yuan. Recognizing human actions as the evolution of pose estimation maps. In *CVPR*, 2018. 1, 2, 5, 7

[12] Mahdi Abavisani, Hamid Reza Vaezi Joze, and Vishal M Patel. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In *CVPR*, 2019. 2, 4

[13] Aggelos K Katsaggelos, Sara Bahaadini, and Rafael Molina. Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE*, 103(9), 2015. 1, 2

[14] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1, 2, 6, 8

[15] Pradeep Natarajan, Shuang Wu, Shiv Vitaladevuni, Xiaodan Zhuang, Stavros Tsakalidis, Unsang Park, Rohit Prasad, and Premkumar Natarajan. Multimodal feature fusion for robust event detection in web videos. In *CVPR*, 2012. 1, 2

[16] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, Geoffrey E Hinton, et al. Adaptive mixtures of local experts. *Neural computation*, 1991. 1, 2

[17] Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. MFAS: Multimodal fusion architecture search. *CVPR*, 2019. 1, 7, 8

[18] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *ECCV*, 2018. 1, 8

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 1, 3, 4, 8

[20] Fan Li, Natalia Neverova, Christian Wolf, and Graham Taylor. Modout: Learning multi-modal architectures by stochastic regularization. In *International Conference on Automatic Face & Gesture Recognition*, 2017. 2

[21] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*, 2017. 2

[22] Guangnan Ye, Dong Liu, I-Hong Jhuo, and Shih-Fu Chang. Robust late fusion with rank minimization. In *CVPR*, 2012. 2

[23] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *ICLR Workshops*, 2017. 2, 8

[24] Di Hu, Chengze Wang, Feiping Nie, and Xuelong Li. Dense multimodal fusion for hierarchically joint representation. In *ICASSP*, 2019. 2

[25] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011. 2

[26] Mehran Khodabandeh, Hamid Reza Vaezi Joze, Ilya Zharkov, and Vivek Pradeep. DIY human action dataset generation. In *CVPR Workshops*, 2018. 2

[27] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. *TPAMI*, 2019. 2

[28] Diogo C Luvizon, David Picard, and Hedi Tabia. 2D/3D pose estimation and action recognition using multitask deep learning. In *CVPR*, 2018. 2, 5, 7

[29] Jian-Fang Hu, Wei-Shi Zheng, Jiahui Pan, Jianhuang Lai, and Jianguo Zhang. Deep bilinear learning for RGB-D action recognition. In *ECCV*, September 2018. 2, 5, 7

[30] Asif A Ghazanfar and Nikos K Logothetis. Neuroperception: Facial expressions linked to monkey calls. *Nature*, 2003. 2

[31] Sarah Partan and Peter Marler. Communication goes multimodal. *Science*, 1999.

[32] Candy Rowe. Sound improves visual discrimination learning in avian predators. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 2002. 2

[33] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. My lips are concealed: Audio-visual speech enhancement through obstructions. *arXiv preprint arXiv:1907.04975*, 2019. 2

[34] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2018. 2

[35] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz. Hand gesture recognition with 3D convolutional neural networks. In *CVPR workshops*, 2015. 3

[36] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Kari Pulli. Multi-sensor system for driver's hand-gesture recog-

nition. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015. 3, 4

[37] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network. In *CVPR*, 2016. 3, 4, 5, 6, 8

[38] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Using convolutional 3D neural networks for user-independent continuous gesture recognition. In *International Conference on Pattern Recognition*. IEEE, 2016. 3

[39] Qiguang Miao, Yunan Li, Wanli Ouyang, Zhenxin Ma, Xin Xu, Weikang Shi, Xiaochun Cao, Zhipeng Liu, Xiujuan Chai, Zhuang Liu, et al. Multimodal gesture recognition based on the ResC3D network. In *ICCV Workshops*, 2017. 3, 4

[40] Liang Zhang, Guangming Zhu, Peiyi Shen, Juan Song, Syed Afaq Shah, and Mohammed Bennamoun. Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition. In *CVPR*, 2017. 3

[41] Congqi Cao, Yifan Zhang, Yi Wu, Hanqing Lu, and Jian Cheng. Egocentric gesture recognition using recurrent 3D convolutional neural networks with spatiotemporal transformer modules. In *CVPR*, 2017. 4, 5, 6

[42] Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *CVPR*, 2017.

[43] Guangming Zhu, Liang Zhang, Peiyi Shen, and Juan Song. Multimodal gesture recognition using 3-D convolution and convolutional LSTM. *IEEE Access*, 2017. 3

[44] Okan Kopuklu, Neslihan Kose, and Gerhard Rigoll. Motion fused frames: Data level fusion strategy for hand gesture recognition. In *CVPR Workshops*, 2018. 3, 4, 6

[45] Pradyumna Narayana, Ross Beveridge, and Bruce A. Draper. Gesture recognition: Focus on the hands. In *CVPR*, 2018. 3, 6

[46] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia*, 2018. 4, 5, 6

[47] Hamid Reza Vaezi Joze and Oscar Koller. MS-ASL: A large-scale data set and benchmark for understanding american sign language. *BMVC*, 2019. 4

[48] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 4, 5, 6, 7

[49] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 4

[50] Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*, 2017. 4, 7

[51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 8

[52] Daniel Michelsanti, Zheng-Hua Tan, Sigurdur Sigurdsson, and Jesper Jensen. On training targets and objective functions for deep-learning-based audio-visual speech enhancement. In *ICASSP*, 2019. 4

[53] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *IJCAI*, 2018. 5, 7, 8

[54] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 6, 7

[55] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *CVPR*, 2016. 6, 7, 8

[56] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[57] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[58] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 6

[59] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 6

[60] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, 2015. 6

[61] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, 2003. 6

[62] Heng Wang, Dan Oneata, Jakob Verbeek, and Cordelia Schmid. A robust and efficient video representation for action recognition. *IJCV*, 2016. 6

[63] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 6

[64] Eshed Ohn-Bar and Mohan Manubhai Trivedi. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE transactions on intelligent transportation systems*, 2014. 6

[65] Jon Barker, Emmanuel Vincent, Ning Ma, Heidi Christensen, and Phil Green. The PASCAL CHiME speech separation and recognition challenge. *Computer Speech and Language*, 2013. 7

[66] Emmanuel Vincent Jon Barker, Ricard Marxer and Shinji Watanabe. The third CHIME speech separation and recognition challenge: Analysis and outcomes. *Computer Speech and Language*, 2017. 7

[67] Zhiyao Duan, Gautham J Mysore, and Paris Smaragdis. Online plca for real-time semi-supervised source separation. In *International Conference on Latent Variable Analysis and Signal Separation*, 2012. 7

[68] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, 2015. 7

[69] Guoning Hu and DeLiang Wang. A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010. 7

[70] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3FD: Single shot scale-invariant face detector. In *CVPR*, 2017. 7

[71] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *ICCV*, 2017. 7

[72] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. *arXiv preprint arXiv:1711.08789*, 2017. 7

[73] ITU-T Recommendation. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*, 2001. 7

[74] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011. 7

[75] Amir Shahroudy, Tian-Tsong Ng, Yihong Gong, and Gang Wang. Deep multimodal feature analysis for action recognition in RGB+D videos. *TPAMI*, 2017. 7

[76] Fabien Baradel, Christian Wolf, Julien Mille, and Graham W Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *CVPR*, 2018. 8

[77] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In *ACCV*, 2016. 8