Towards Good Practices for Missing Modality Robust Action Recognition

Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, Changick Kim

Korea Advanced Institue of Science and Technology (KAIST)
{smwoo95, suminlee94, yeonju29, madin, changick}@kaist.ac.kr

Abstract

Standard multi-modal models assume the use of the same modalities in training and inference stages. However, in practice, the environment in which multi-modal models operate may not satisfy such assumption. As such, their performances degrade drastically if any modality is missing in the inference stage. We ask: *how can we train a model that is robust to missing modalities?* This paper seeks a set of good practices for multi-modal action recognition, with a particular interest in circumstances where some modalities are not available at an inference time. First, we study how to effectively regularize the model during training (*e.g.*, data augmentation). Second, we investigate on fusion methods for robustness to missing modalities: we find that transformer-based fusion shows better robustness for missing modality than summation or concatenation. Third, we propose a simple modular network, ActionMAE, which learns missing modality predictive coding by randomly dropping modality features and tries to reconstruct them with the remaining modality features. Coupling these good practices, we build a model that is not only effective in multi-modal action recognition but also robust to modality missing. Our model achieves the state-of-the-arts on multiple benchmarks and maintains competitive performances even in missing modality scenarios.

Introduction

This study aims to answer the underlying question about multi-modal learning for action recognition in practical situations: *How can we train a model that is robust to missing modalities?* Typical multi-modal models assume complete modalities in both training and inference phases (Bruce, Liu, and Chan 2021; Bruce et al. 2022). In reality, however, the multi-modal system may be unable to access particular modalities during the inference phase, despite being able to access all modalities reliably during the training phase (see Figure 1). There are a number of potential causes for such circumstances, including malfunctioning sensors, high data acquisition costs, inaccessibility due to security or privacy concerns, self-deficiencies, *etc*. This situation affects the reliability, accuracy, and safety of the model in real-world applications. Considering the autonomous driving situation, an error caused by insufficient use of the sensors due to factors such as inclement weather would be life-threatening. As such, the missing modality problem is critical when multi-modal models are employed in practice.

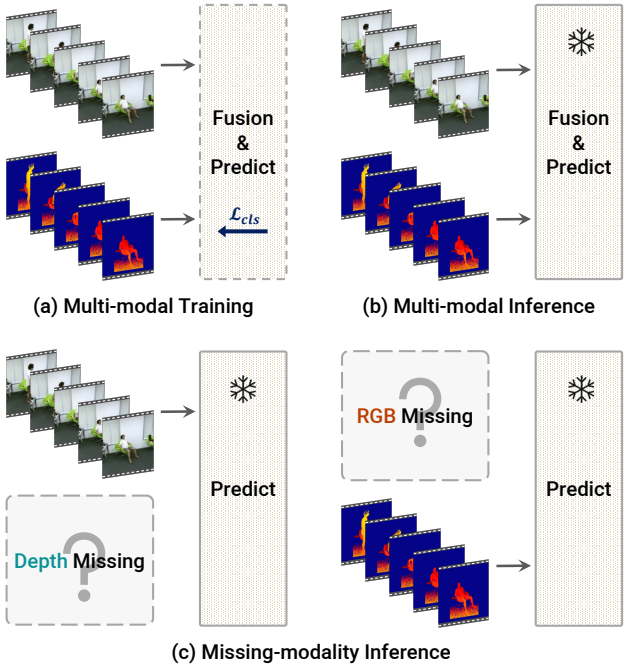

Figure 1: Action recognition with missing modality. Standard multi-modal action recognition assumes that the modalities used in the training stage are complete at inference time: (a) → (b). This paper addresses the action recognition problem in situations where such assumption is not established, *i.e.*, when modalities are incomplete at inference time: (a) → (c). Our goal is to maintain performance in the absence of any input modality. ❄ indicates the weight-frozen.

To address the missing modality scenarios, we require a robust model that operates reliably even when certain modalities are unavailable in multi-modal systems. However, conventional multi-modal models are learnt under the premise that all modalities exist, so that they cannot achieve sufficient performances when directly applied to missing modality settings. Table 1 shows how much the conventional multi-modal models suffer from missing modality. Our goal is to achieve

Dataset	Train modal*	Test modal	Acc.†	△ (%p)
NTU-RGB+D 60	R+D+I	R+D+I	93.3%	-
	R+D+I	R+I	59.4%	−**33.9**
	R+D+I	R	15.6%	−**77.7**
NW-UCLA	R+D	R+D	91.9%	-
	R+D	R	53.6%	−**38.3**
	R+D	D	70.3%	−**21.6**

*__R__: RGB, __D__: depth, __I__: infrared. † Results based on sum fusion.

Table 1: Multi-modal models suffer severely in missing modality scenarios. We observe significant performance drop when the modality is removed one by one from the complete modality setup in inference stage.

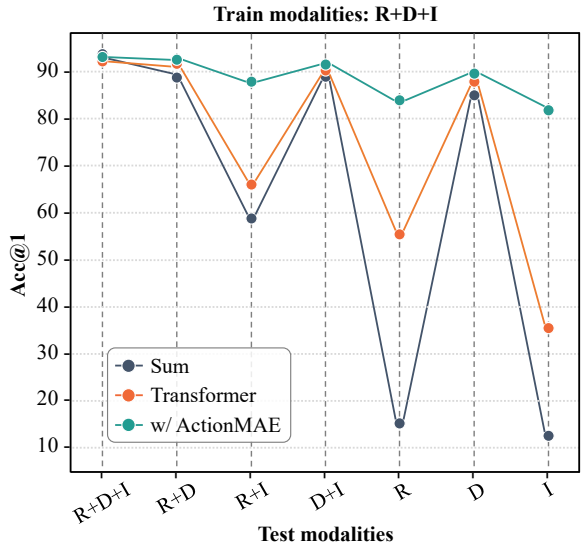

Figure 2: Towards missing modality robust action recognition. **(i)** Transformer fusion is robust than sum fusion. **(ii)** In contrast to conventional approaches, our ActionMAE well preserves the multi-modal performance even with incomplete modalities. All models are trained with R+D+I modalities and tested on NTU RGB+D 60 (Shahroudy et al. 2016).

on-par results as when all modalities are available, even if any of the input modality is missing. We achieve this by designing a set of good practices for missing modality scenarios in multi-modal action recognition (see Figure 2): **(i)** As a preliminary step, we first demonstrate a recipe of how to effectively build a multi-modal model with several training ingredients such as architecture, data augmentation, and regularization. **(ii)** We observe that the fusion method matters for robustness to missing modality. We compare the three most widely used methods in the literature among the numerous methods for multi-modal fusion: summation, concatenation, and transformer. Among them, we find that transformer-based fusion method is the most robust against missing modality. **(iii)** We propose Action Masked Auto Encoder (ActionMAE), a simple modular network that learns missing modality predictive coding. The model with ActionMAE is optimized under the two objectives. It jointly learns to reconstruct the missing modality features and to predict the correct action label during training. We can plug-and-play ActionMAE on top of standard space-time encoders. Despite its simplicity, ActionMAE achieves a win-win situation: it not only regularizes the multi-modal model to achieve better accuracy in complete modality cases, but also maintains accuracy in missing modality cases.

We report our results on four challenging action recognition benchmarks: NTU RGB+D 60 (Shahroudy et al. 2016), NTU RGB+D 120 (Liu et al. 2019), NW-UCLA (Wang et al. 2014), and UWA3D (Rahmani et al. 2016). We set new state-of-the-art results in both complete and missing modality settings. Furthermore, we observe some intriguing properties of ActionMAE from experiments, including the following: **(i)** Missing modality reconstruction yields a nontrivial and meaningful self-supervision. As such, the results with reconstructed features are superior than those with the original features extracted directly from space-time encoder. **(ii)** ActionMAE effectively regularizes the multi-modal model, hence the model equipped with ActionMAE achieves performance gain over the vanilla model when using complete modalities. **(iii)** Moreover, ActionMAE alleviates the bias of the multi-modal model toward the dominant modality (*i.e.*, the most contributing modality to the learning objective). **(iv)** The robustness of ActionMAE is agnostic to the type or the number of missing modalities. This suggests that it can preserve accuracy even in real-world environments where we do not know which modality might be missing.

Preliminary: Multi-modal Action Classifier

We first study how to design a strong multi-modal action recognition model with several design choices.

Architecture

The standard architecture of multi-modal action recognition is in a form of space-time encoder followed by a fusion unit. We seek the optimal space and time encoders for multi-modal action recognition with the fusion unit set to a simple summation by default. Table 2 compares the accuracies of the simplest combinations among the numerous candidates: R(2+1)D (Tran et al. 2018), ViT (Dosovitskiy et al. 2020), ResNet (He et al. 2016), Transformer (Vaswani et al. 2017).

Space encoder*	Time encoder	Fusion	Train → Test	
			RD → RD	RD → R
R(2+1)D			89.2%	41.2%
ViT-tiny/16†	Transformer		81.4%	35.7%
ViT-small/16†	Transformer	Sum	83.0%	36.3%
ResNet18	Transformer		91.9%	41.9%
ResNet34	Transformer		**92.6%**	**42.5%**

*Space encoders are Imagenet initialized.
† Re-implementations from (Wightman 2019).

Table 2: ResNet34 + Transformer achieves the best accuracy in both full (RD → RD) and missing modality (RD → R) cases. Experiments are conducted on NTU-RGB+D 60.

We choose ResNet34 and Transformer as our space and time encoders, respectively. The procedure of space-time encoding is formalized as follows. We sample a sequence of frames $[f_i^m]_{i=1}^T$ ($f_i^m \in \mathbb{R}^{C_0 \times H \times W}$) from a video $\mathcal{V}^m$, where

Variant	Train→Test	
	RDI → RDI	RD → R
Center crop, horizontal flip	85.9	34.8
+ Imagenet initialization	89.2 (+**3.3**)	39.5 (+**4.7**)
+ Shifted center crop	90.5 (+**1.3**)	40.3 (+**0.8**)
+ Temporal random sampling	91.0 (+**0.5**)	41.4 (+**1.1**)
+ Color jitter (RGB)	92.2 (+**1.2**)	42.0 (+**0.6**)
+ Label smoothing ($\alpha = 0.1$)	**93.3** (+**1.1**)	**42.5** (+**0.5**)
- Color jitter (Depth)	91.5 (−**1.8**)	-
- Dropout ($p = 0.1$)	92.6 (−**0.7**)	42.5

Table 3: Progressively adding regularizations helps much in both full and missing modality cases. However, applying color jittering on depth maps and using dropout rather reduced the performance in full modality case. Experiments are conducted on NTU RGB+D 60 with the sum fusion model.

i indicates the frame index and m indicates the modality. In practice, we set $T = 16$, $C_0 = 3$, $H = W = 224$, and $m \in \{\mathrm{R}, \mathrm{D}, \mathrm{I}\}$. A frame of modality m is processed with a modality-specific space encoder $\mathcal{E}^m$ followed by an average pooling:

$$x_i^m = \texttt{avgpool}(\mathcal{E}^m(f_i^m)) \,. \tag{1}$$

The output feature is $x_i^m \in \mathbb{R}^C$, where $C = 512$. The space encoder operates on every frame and we obtain a sequence of features $[x_i^m]_{i=1}^T$. Before feeding them to the time encoder, which is a transformer encoder, a learnable class token `cls` is prepended to the sequence. Its representation at the last layer of the encoder serves as the final representation used by the classification layer (Devlin et al. 2018; Dosovitskiy et al. 2020; Arnab et al. 2021).

$$\mathcal{X}_{(0)}^m = [\texttt{cls}^m, x_1^m, ..., x_T^m] + \texttt{pos} \,, \tag{2}$$

where, `pos` is a fixed absolute embedding which represent the temporal positions [1]. A sequence of tokens is then fed to a transformer encoder with L layers (we set $L = 2$), where each layer l is composed of Multi-Head Self Attention (MHSA), Layer Normalization (LN), and Feed Forward Network (FFN) [2]:

$$\mathcal{Y}_{(l)}^m = \text{MHSA}(\text{LN}(\mathcal{X}_{(l)}^m)) + \mathcal{X}_{(l)}^m \,, \tag{3}$$

$$\mathcal{X}_{(l+1)}^m = \text{FFN}(\text{LN}(\mathcal{Y}_{(l)}^m)) + \mathcal{Y}_{(l)}^m \,. \tag{4}$$

The output at the final layer $\mathcal{X}_{(L)}^m \in \mathbb{R}^{(T+1)\times C}$ holds the same dimension as that of the input $\mathcal{X}_{(0)}^m$. At last, we use the final classification token $\texttt{cls}_L^m \in \mathbb{R}^C$ to classify the action classes with the fully-connected (FC) layer.

Tips and Tricks for Training

In our preliminary experiments, we noticed that our models are prone to overfit, highlighting the necessity for a bag of tricks to train our models effectively. In order to see the effect of augmentations and regularizations, we ablate several strategies in Table 3. By default, we employ center crop and horizontal flip for all modality data, and progressively add more strategies: Imagenet initialization on space encoder is highly beneficial, implying that it eases the subsequent optimization of time encoder. It has relatively large impact in the missing modality case compared to the full modality case, *i.e.*, the strong inductive bias provided by ImageNet initialization can have a particularly positive effect in the absence of depth information. Instead of the vanilla center crop, we center crop the larger region ($\times 1.2$) and then random crop ($\times 1$) within it. Instead of uniformly sampling frames from video, we randomly sample frames and arrange them in temporal order. Color jitter increases divergence of RGB distribution, which has a beneficial effect. label smoothing (Szegedy et al. 2016) adjusts the label distribution of the ground truth to encourage the model to produce less confident predictions. Generally, regularizing the model or increasing the data diversity were beneficial in both full and missing modality cases. However, some methods were detrimental in full modality case: Color jitter on depth map rather harms the consistency of 3D structural information. Dropout (Srivastava et al. 2014) makes optimization rather difficult. In missing modality case, Dropout did not have a detrimental effect, but it also did not help regularize the model. Overall, we achieve a substantial improvement of 7.4%p (7.7%p for missing modality

Fusion	Train modal	Test modal ↑	△(%p) ↑	$\lvert\bar{\triangle}\rvert\downarrow$
Sum	R+D+I	R+D+I (**93.3%**)	-	**34.8**
		R+D (89.0%)	−**4.3**	
		R+I (59.4%)	−**33.9**	
		D+I (89.2%)	−**4.1**	
		R (15.6%)	−**77.7**	
		D (85.9%)	−**7.4**	
		I (12.1%)	−**81.2**	
Concat	R+D+I	R+D+I (**93.2%**)	-	**33.5**
		R+D (91.4%)	−**1.8**	
		R+I (75.4%)	−**17.8**	
		D+I (18.6%)	−**74.6**	
		R (53.1%)	−**40.1**	
		D (88.2%)	−**5.0**	
		I (31.7%)	−**61.5**	
Transformer	R+D+I	R+D+I (**92.6%**)	-	**21.0**
		R+D (91.6%)	−**1.0**	
		R+I (66.6%)	−**26.0**	
		D+I (90.4%)	−**2.2**	
		R (56.6%)	−**36.0**	
		D (88.9%)	−**3.7**	
		I (35.4%)	−**57.2**	

Table 4: Transformer is a missing modality robust fusion unit. Experiments are conducted on NTU RGB+D 60. △ is defined as accuracy discrepancy between missing modality and full modality cases ($\triangle = \text{Acc}_{miss} - \text{Acc}_{full}$). While all three fusion methods show comparable performance when test modalities are complete, they show different aspects in missing modality inference. ↑/↓: The higher/lower the better.

[1] It is added to the tokens to maintain positional information since the subsequent self-attention operations in the transformers are permutation-invariant.

[2] We leave an original paper as a reference (Vaswani et al. 2017) for further details on transformer building blocks.

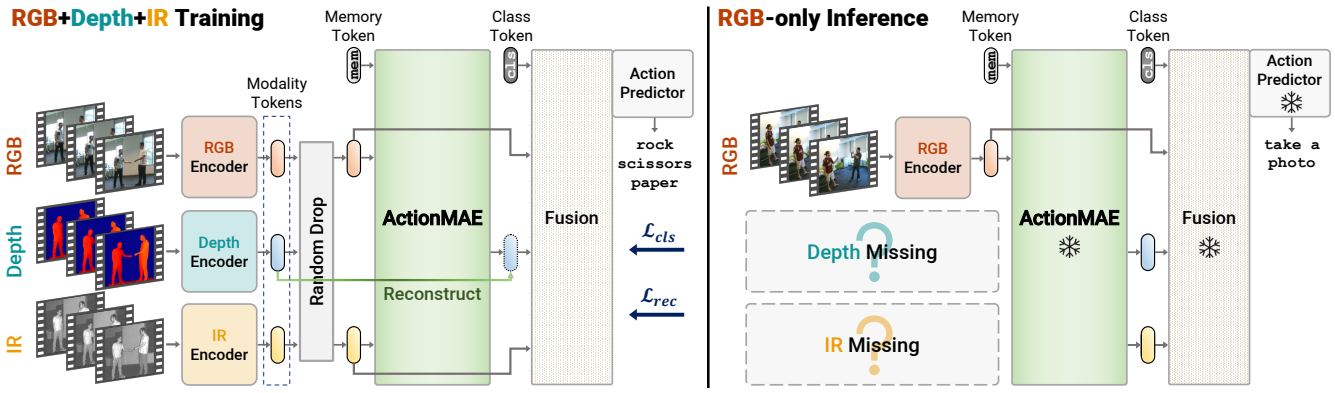

Figure 3: ActionMAE framework. **(i)** *Multi-modal training*: We first obtain modality-specific features via their respective encoding stream. Then, we randomly drop K (here, $K = 1$) input tokens from N (=3) modality tokens, where K ($0 \leq K < N$) varies at every iteration. ActionMAE learns to reconstruct missing modality tokens using the remaining $N - K$ modality tokens. During autoencoding, the memory token, which is also provided as an input, serves as global memory across varying inputs. The original and reconstructed tokens are reorganized as an input sequence to the fusion unit. The succeeding processes follow the standard fusion-and-predict procedure. The overall training objective is defined in Equation (6). **(ii)** *Missing modality inference*: We drop K (=2) tokens (here, Depth, IR) and predict the action class by reconstructing K tokens with remaining $N - K$ (=1) token (RGB). We note that ActionMAE is agnostic to the types or the number of missing modalities.

case) on NTU RGB+D 60. Note that no spatial or temporal augmentations are used in testing.

Fusion

In Table 4, we examine exemplary fusion methods in terms of multi-modal inference and missing modality inference. We take three simple yet effective fusion methods that are commonly used in various multi-modal models: summation, concatenation, and transformer.

- **Sum.** Each modality passes through its respective FC layer and then is all summed up: $\sum_m \text{FC}^m(\texttt{cls}_L^m)$.
- **Concat.** Modalities are channel-wise concatenated ($\|$) and passed through a single FC layer: $\text{FC}(\|_m \texttt{cls}_L^m)$. [3]
- **Transformer.** Extra class token is prepended to a sequence of modality tokens $[\texttt{cls}, \texttt{cls}_L^R, \texttt{cls}_L^D, \texttt{cls}_L^I]$, which is then passed through the transformer (Eqs. (3) and (4)). The class token of the final transformer layer is subsequently passed through the FC layer.

The test accuracies with complete modalities (rows makred in gray) are not much different between three fusion methods. Among them, the transformer-based fusion method is the most inferior (lags behind sum by 0.7%p). To our surprise, however, we observe the opposite tendency in a missing modality situation. The transformer-based fusion is the most robust to the missing modalities: it reduces the mean accuracy discrepancy ($|\bar{\triangle}|$) by 13.8%p compared to sum fusion. Nevertheless, we notice the bias in all three fusion methods. They specifically works well with a certain modality (*e.g.*, depth), but poorly with other modalities (*e.g.*, RGB, IR) under the situation of missing modality.

[3] In the case of concatenation, uni-modal predictions are approximated by splitting the weight matrix of FC layer into sub-matrices and dividing the bias by the number of modalities.

Method: ActionMAE

To address the missing modality scenario in action recognition problem, we propose ActionMAE that learns missing modality predictive coding as illustrated in Figure 3.

Our ActionMAE is a simple modular autoencoder that reconstructs the missing modalities based on partial observation of the remaining modalities. Like generic autoencoders, our solution is made up of an encoder that maps the observed modalities to a latent representation and a decoder that reconstructs the missing modalities from the latent representation. The encoder and decoder adopted in our ActionMAE are simple transformer encoders (Vaswani et al. 2017). By design, it can be plugged into a typical multi-modal action classifier with a fusion unit. We plug-and-play ActionMAE on top of the model explored in preliminary experiments: ResNet34 + Transformer. During training, the model equipped with ActionMAE simultaneously learns missing modality information and predicts the correct action class based on these.

Missing Modality Predictive Coding

We define a missing modality scenario as a case in which N modalities are accessible during training, but K modalities are missing during inference, *i.e.*, only $N - K$ modalities are accessible, where $1 \leq K < N$. We use three modalities in practice, RGB, Depth, and IR (as such, $N = 3$), but the number or type of modalities are not limited to these. From N modality tokens $[z^m]_{m=1}^N$ (equivalent to $[\texttt{cls}_L^m]_{m=1}^N$ of Eq. (4)), we randomly drop K tokens, *i.e.*, only leaving $N - K$ tokens, where $0 \leq K < N$ (if $K = 0$, we do not drop any of modality tokens). In addition to remaining $N - K$ tokens, the ActionMAE encoder takes extra learnable token, *i.e.*, memory token, that acts as global memory across varying inputs. It is worth noting that the input sequence to ActionMAE encoder continues to vary with each iteration. We first obtain a sequence of latent representations

via the encoder. Then, the K dummy tokens are inserted into the sequence at the position where the modality tokens are missing, and the resulting sequence is fed to the decoder. As the dummy tokens pass through the decoder, they are forced to mimic the dropped tokens. The original modality tokens are reorganized with the reconstructed tokens and passed to the subsequent fusion unit and prediction layer in turns. Notice that the memory token is used only for encoding-decoding process, not for prediction.

Training & Inference

Overall training of the model with ActionMAE is governed by the two loss functions: standard classification loss (*i.e.*, cross-entropy loss) $\mathcal{L}_{\text{cls}}$ and reconstruction loss $\mathcal{L}_{\text{rec}}$. The decoder aims to convert dummy tokens into original modality tokens. We define the indices of missing tokens as $j_1, ..., j_K$. The reconstruction loss computes the mean squared error (MSE) between the reconstructed $[\hat{z}^{j_1}, ..., \hat{z}^{j_K}]$ and missing tokens $[z^{j_1}, ..., z^{j_K}]$ in the feature space:

$$\mathcal{L}_{rec} = \frac{1}{K}\sum_{i=1}^{K}(z^{j_i} - \hat{z}^{j_i})^2 \,. \tag{5}$$

Note that only memory token and dummy tokens are affected by the reconstruction loss (*i.e.*, reconstruction loss is not computed for the remaining tokens) during training.

Our training objective is jointly optimized upon the two loss functions:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{rec}}\mathcal{L}_{rec} \,. \tag{6}$$

In our experiments, we empirically find that balancing hyper-parameters $\lambda_{\text{cls}} = \lambda_{\text{rec}} = 1$ works well.

At inference time, K modalities are dropped from the input stage, and remaining $N - K$ modalities are encoded by their respective modality encoders. ActionMAE takes only the memory token and $N - K$ encoded modality tokens, and reconstructs the missing tokens. We progressively drop the modalities to see whether ActionMAE is robust under all circumstances, regardless of which modality or how many modalities are lost. Experiments demonstrate that our ActionMAE is quite robust to any cases of modality missing.

Experiments

Implementation

Inputs. The input of each modality network is a 16 frame clip of the corresponding modality, and their input size is 224×224 pixels. In the spatial domain, random horizontal flip and shifted center crop are performed on all modality inputs. RGB frames are applied with color jittering, and the 1-channel depth frames are converted into 3-channels based on a JET color map with 256 scales. In the temporal domain, we randomly sampled 16 frames from the raw video, and arrange them in temporal order. At test time, each clip consists of 16 frames that are uniformly sampled, center cropped, and no data augmentation techniques are applied.

Model	Train modal	Test modal ↑	△(%p) ↑	$\lvert\bar{\triangle}\rvert\downarrow$
Sum	R+D+I	R+D+I (93.3%)	-	**34.8**
Concat		R+D+I (93.2%)	-	**33.5**
Transformer		R+D+I (92.6%)	-	**21.0**
Transformer w/ ActionMAE	R+D+I	R+D+I (93.0%)	-	**8.3**
		R+D (92.6%)	**−0.4**	
		R+I (87.6%)	**−5.4**	
		D+I (91.9%)	**−1.1**	
		R (83.4%)	**−9.6**	
		D (89.7%)	**−3.3**	
		I (81.7%)	**−11.3**	

Table 5: ActionMAE is a missing modality robust learner. Experiments are conducted on NTU RGB+D 60. ↑/↓: The higher/lower the better.

Settings. We used AdamW optimizer (Loshchilov and Hutter 2017) with an initial learning rate of 10^{-4} and weight decay of 10^{-4} for a batch size of 32. The learning rate is linearly decayed by a factor of 10 every 30 epochs. We used cross-entropy loss with label smoothing of factor 0.1 (Szegedy et al. 2016). All space encoders are initialized with imagenet pre-trained weights (Deng et al. 2009). ActionMAE features a symmetrical design with two encoder and decoder layers.

Datasets

We used NTU RGB+D 60 (NTU60) (Shahroudy et al. 2016), NTU RGB+D 120 (NTU120) (Liu et al. 2019), NW-UCLA (Wang et al. 2014), and UWA3DII (Rahmani et al. 2016) for experiments. Following the convention (Liu et al. 2019), we evaluate our model using the cross-subject protocol in both NTU60 and NTU 120. For NW-UCLA, we followed the cross-view protocol suggested in (Wang et al. 2014), using two views (V1, V2) for training and the remaining (V3) for testing. For UWA3DII, we used the top and right views for training and front and left views for testing.

ActionMAE

In Table 5, we examined the effect of ActionMAE by comparing the model with and without ActionMAE. In preliminary experiments, we observed that complete modality performance was rather lower when using transformer than much simpler methods such as sum or concatenation. We hypothesize that this is due to the lack of inductive bias in transformer, which makes it susceptible to overfitting. When we attach ActionMAE on a transformer-fusion model, the random drop operation yields nontrivial self-supervisory task, which effectively regularizes the model. This results in 0.4%p performance improvement under complete modality setting. Generally, ActionMAE shows good robustness to any combination of missing modalities. More importantly, the missing modality performances, which were particularly poor in accuracy, are significantly improved upon vanilla transformer baseline (*e.g.*, IR: $35.4\% \rightarrow 81.7\%$). This implies that ActionMAE also has a debiasing effect on the dominant modality (*e.g.*, depth). Overall, ActionMAE reduces the mean accuracy discrepancy by $\sim 2.5\times$ compared to the transformer-fusion baseline, and by $\sim 4\times$ compared to the sum-fusion baseline.

	Model	Train modal*	Test modal	Universal†	NTU60	NTU120	NWUCLA	UWA3D
Baseline	Uni-modal (/w Transformer-fusion)	R	R	✗	86.6%	84.2%	85.0%	73.1%
		D	D	✗	92.0%	88.2%	92.7%	81.2%
		I	I	✗	85.5%	85.4%	-	-
	Multi-modal (/w Transformer-fusion)	R+D	R+D	✗	92.3%	91.7%	90.9%	82.5%
		R+D+I	R+D+I	✗	92.6%	92.1%	-	-
Complete	3DFCNN (Sanchez-Caballero et al. 2022)	D	D	✗	78.1%	-	83.6%	66.6%
	Ren *et al*. (2021)	R+D	R+D	✗	89.7%	-	-	-
	Deep Bilinear (Hu et al. 2018)	R+D+S	R+D+S	✗	85.4%	-	-	-
	DMCL-complete (Garcia et al. 2019)	R+D+F	R+D+F	✗	87.3%	89.7%	93.9%	89.8%
Missing	Hoffman *et al*. (2016)	R+D	R	✗	-	-	83.2%	66.7%
	Garcia *et al*. (2018)	R+D	R	✗	73.1%	-	86.7%	73.2%
	ADMD (Garcia, Morerio, and Murino 2019)	R+D	R	✗	73.4%	-	93.6%	78.4%
	Luo *et al*. (2018)	R+D+F+S	R	✗	89.5%	-	-	-
			D	✗	87.5%	-	-	-
	DMCL (Garcia et al. 2019)	R+D+F	R	✗	83.6%	84.3%	93.6%	78.4%
			D	✗	80.6%	82.2%	83.3%	81.9%
Ours	ActionMAE (R+D+I)	R+D+I	R+D+I	✓	93.0%	92.3%	-	-
			R+D	✓	92.6%	91.7%	-	-
			R+I	✓	87.6%	84.8%	-	-
			D+I	✓	91.9%	91.2%	-	-
			R	✓	83.4%	83.1%	-	-
			D	✓	90.1%	86.0%	-	-
			I	✓	81.7%	81.6%	-	-
	ActionMAE (R+D)	R+D	R+D	✓	92.5%	91.5%	91.0%	79.8%
			R	✓	84.5%	84.7%	84.2%	70.4%
			D	✓	90.5%	87.0%	88.2%	77.6%

*__R__: RGB, __D__: depth, __I__: infrared, __F__: optical flow, __S__: skeleton. † preserves accuracy no matter what kind and how many modalities are missing.

Table 6: Comparative study on multiple action recognition benchmarks: NTU60 (Shahroudy et al. 2016), NTU120 (Liu et al. 2019), NWUCLA (Wang et al. 2014), and UWA3D (Rahmani et al. 2016). Here, we report top-1 accuracy. By design, our ActionMAE is robust to missing modalities of any type and number.

Comparative Study

In Table 6, we benchmark our approach against competitive baselines on multiple action recognition datasets.

***Vs.* uni-/multi-modal baselines.** We begin by comparing ActionMAE with transformer-fusion baselines. With complete modalities, ActionMAE mostly improves performance upon multi-modal baselines as randomized drop of modalities during training gives a regularization effect on the overall network analogous to dropout (Srivastava et al. 2014). While our model is also fairly robust to missing modalities in both R+D+I and R+D cases, uni-modal models outperform ActionMAE when tested with a single modality. However, in the missing modality situation evaluated with R+D, our model (92.6%) slightly improves accuracy compared to the multi-modal baseline (92.3%).

***Vs.* complete modality counterparts.** We compare our approach with recent action recognition methods. These methods do not drop any modality at the test time (*i.e*., complete modality inference), yet our method significantly outperforms all listed methods even in missing modality scenarios.

***Vs.* missing modality counterparts.** Next, we compare our approach with missing modality counterparts. Generally, as more modalities are used in the training stage, they show better performance in missing modality situations. In both R+D+I and R+D cases, the accuracy of ActionMAE decreases as more modalities are dropped, albeit shows competitive performance. When we follow the most common setup in the literature (train: R+D, test: R or D), ActionMAE sets new state-of-the-art results.

ActionMAE (R+D+I *vs*. R+D). In the case of R+D+I, if just one modality is missing at test time (*e.g*., R+D), there is a minimal loss of accuracy (-0.4%p). However, in the absence of two modalities (*e.g*., R), mapping from latent representation becomes more complex as the number of values to be restored increases, thus drops more (-9.6%p). On the other hand, even evaluated with the same modality (*e.g*., D), R+D performs better than R+D+I since the mapping is simpler.

Ablative Study

From Table 7, we observed several intriguing properties of ActionMAE by ablating it.

Memory token. Memory token contributes significantly to performance improvement, especially for the missing modality cases: 9.7%p gain for R-only case and 7.3%p gain for D-only case. We understand this as memory token learns meaningful correlation between varying inputs and models

case	RD	R	D
w/o mem	87.3	74.8	83.2
w/ mem	**92.5**	**84.5**	**90.5**

(a) **Memory token.**

$\mathcal{L}_{rec}$:$\mathcal{L}_{cls}$	RD	R	D
1:10	91.9	83.0	90.0
1:1	**92.5**	**84.5**	**90.5**
10:1	91.0	81.6	88.7

(b) **Loss coefficients.**

case	RD	R	D
w/o pos	87.2	76.3	84.7
w/ pos	**92.5**	**84.5**	**90.5**

(c) **Position embedding.**

case	RD	R	D
No PT	**92.5**	**84.5**	**90.5**
PT (S)	89.1	82.0	86.4
PT (S+T)	88.5	81.9	86.6

(d) **Pre-training.**

Table 7: Ablation experiments on NTU RGB+D 60 with ActionMAE (R+D).

pair-wise relationships between itself and modality tokens.

Loss hyperparameters. We investigated a good balance between reconstruction $\mathcal{L}_{rec}$ and classification loss $\mathcal{L}_{cls}$. From an exploration-exploitation perspective, when the overall objective is leaned to classification (*i.e.*, $\mathcal{L}_{rec} : \mathcal{L}_{cls} = 1 : 10$), ActionMAE underutilizes the features extracted from the space-time encoder, resulting in poor 'exploitation' of learned information. On the other hand, if the objective is leaned to reconstruction (*i.e.*, 10:1), ActionMAE loses the chance of 'exploration' to generate more class-discriminative features. We found a ratio of 1:1 properly balances the two losses.

Position embedding. We observed a significant performance improvements with position embedding (pos). From the results, we hypothesize that position embedding serves as a reference for determining which modalities needs to be reconstructed in ActionMAE, *i.e.*, position embedding provides guidance on what to reconstruct.

Pre-training. We examined whether pre-training the space-time encoder would ease the optimization or rather negatively impact the training of ActionMAE. Here, PT (S) indicates space encoder pre-training, and PT (S+T) indicates pre-training both space and time encoders. Contrary to our expectations, pre-training any encoders ahead of ActionMAE diminishes accuracy. Even, the performance degraded when more layers were pre-trained. This suggests that ActionMAE can be effectively learned through self-supervisory reconstruction of diverse inputs. However, if the space-time encoder is pre-trained, the input will remain almost constant; hence, ActionMAE will likely to find a trivial solution. Therefore, we only performed imagenet initialization on the space encoder and trained all networks from scratch.

Related Work

Multi-modality generally yields reliable results since different modalities provide complementary information (Huang et al. 2021). For example, RGB provides rich appearance information, depth provides 3D geometrical structure, and IR is robust to illumination variation. The missing modality setup assumes certain modalities available during training are unavailable at test time. As the typical multi-modal models suffer in such scenarios, there was a large body of studies to address the missing modality problem (Vapnik and Vashist 2009; Pechyony and Vapnik 2010; Vapnik, Izmailov et al. 2015; Lopez-Paz et al. 2015; Ma et al. 2022; Zhao, Li, and Jin 2021; Alayrac et al. 2020; Ma et al. 2021)

In the context of action recognition (Tran et al. 2015; Wang et al. 2016; Carreira and Zisserman 2017; Feichtenhofer et al. 2019), there are handful of studies towards this end (Hoffman, Gupta, and Darrell 2016; Luo et al. 2018; Garcia, Morerio, and Murino 2018, 2019; Garcia et al. 2019; Stroud et al. 2020), all of which follow transfer learning scheme via knowledge distillation (Hinton et al. 2015). We rather take inspiration from recent successes of masked autoencoders (Devlin et al. 2018; He et al. 2022; Feichtenhofer et al. 2022; Bachmann et al. 2022), whose core idea is to remove a portion of the data and learn to predict the removed data, and show that masked autoencoding works better than the knowledge distillation approaches. In addition to accuracy, there are several advantages of ActionMAE over the knowledge distillation approaches: **(i)** A separate teacher network training, which is a prerequisite for knowledge distillation, is unnecessary. Therefore, the training can be completed in an end-to-end manner with only a single step. **(ii)** We do not need method-specific architecture. ActionMAE can be plugged into any type of pre-existing space-time encoders. **(iii)** By design, ActionMAE is universal to the type or the number of missing modalities.

Discussion

In this paper, we have answered the following questions regarding the missing modality scenarios in multi-modal action recognition problem: **(i)** *How to train a strong multi-modal model?* We sought the good practices from three perspectives (see Tables 2 and 3): architecture, data augmentation, and regularization. **(ii)** *Which fusion should we choose for missing modality action recognition?* Of the three popular choices for fusion (*e.g.*, sum, concatenation, transformer), transformer worked the best (see Table 4), albeit still far from enough. **(iii)** *Is there more effective way to solve missing modality problem?* We showed that ActionMAE is an effective strategy for missing modality action recognition via extensive experiments (see Tables 5 to 7). In addition, it relieves bias against dominant modality, as well as effectively regularizes multi-modal model. We hope that our findings will provide insights into broader missing modality scenarios.

Limitations & future work. There still remain unresolved issues: **(i)** The uni-modal baselines perform better than the model equipped with ActionMAE when evaluated with a single modality in missing modality situations. **(ii)** Even with the same test modality (*e.g.*, R or D), the model with fewer training modalities (R+D) outperforms the model with more training modalities (R+D+I). We posit these challenges occur due to the difficulty of mapping from latent representation of ActionMAE. As the number of outputs exceeds the number of inputs, it yields a sub-optimal solution. We mitigate this problem using memory token, yet more investigation is needed to find a more effective way. Despite the limitations, our idea is simple and can be naturally extended to other modalities, such as vision, language, audio, *etc*. We leave this intriguing challenge to future work.

Acknowledgments

This work was supported by the Agency For Defense Development by the Korean Government (UD190031RD).

References

Alayrac, J.-B.; Recasens, A.; Schneider, R.; Arandjelović, R.; Ramapuram, J.; De Fauw, J.; Smaira, L.; Dieleman, S.; and Zisserman, A. 2020. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33: 25–37.

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6836–6846.

Bachmann, R.; Mizrahi, D.; Atanov, A.; and Zamir, A. 2022. MultiMAE: Multi-modal Multi-task Masked Autoencoders. *arXiv preprint arXiv:2204.01678*.

Bruce, X.; Liu, Y.; and Chan, K. C. 2021. Multimodal fusion via teacher-student network for indoor action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3199–3207.

Bruce, X.; Liu, Y.; Zhang, X.; Zhong, S.-h.; and Chan, K. C. 2022. MMNet: A Model-based Multimodal Network for Human Action Recognition in RGB-D Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Feichtenhofer, C.; Fan, H.; Li, Y.; and He, K. 2022. Masked Autoencoders As Spatiotemporal Learners. *arXiv preprint arXiv:2205.09113*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Garcia, N. C.; Bargal, S. A.; Ablavsky, V.; Morerio, P.; Murino, V.; and Sclaroff, S. 2019. Dmcl: Distillation multiple choice learning for multimodal action recognition. *arXiv preprint arXiv:1912.10982*.

Garcia, N. C.; Morerio, P.; and Murino, V. 2018. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 103–118.

Garcia, N. C.; Morerio, P.; and Murino, V. 2019. Learning with privileged information via adversarial discriminative modality distillation. *IEEE transactions on pattern analysis and machine intelligence*, 42(10): 2581–2593.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

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

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Hoffman, J.; Gupta, S.; and Darrell, T. 2016. Learning with side information through modality hallucination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 826–834.

Hu, J.-F.; Zheng, W.-S.; Pan, J.; Lai, J.; and Zhang, J. 2018. Deep bilinear learning for rgb-d action recognition. In *Proceedings of the European conference on computer vision (ECCV)*, 335–351.

Huang, Y.; Du, C.; Xue, Z.; Chen, X.; Zhao, H.; and Huang, L. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34: 10944–10956.

Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.-Y.; and Kot, A. C. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10): 2684–2701.

Lopez-Paz, D.; Bottou, L.; Schölkopf, B.; and Vapnik, V. 2015. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luo, Z.; Hsieh, J.-T.; Jiang, L.; Niebles, J. C.; and Fei-Fei, L. 2018. Graph distillation for action detection with privileged modalities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 166–183.

Ma, M.; Ren, J.; Zhao, L.; Testuggine, D.; and Peng, X. 2022. Are Multimodal Transformers Robust to Missing Modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18177–18186.

Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021. SMIL: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2302–2310.

Pechyony, D.; and Vapnik, V. 2010. On the theory of learnining with privileged information. *Advances in neural information processing systems*, 23.

Rahmani, H.; Mahmood, A.; Huynh, D.; and Mian, A. 2016. Histogram of oriented principal components for cross-view action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(12): 2430–2443.

Ren, Z.; Zhang, Q.; Gao, X.; Hao, P.; and Cheng, J. 2021. Multi-modality learning for human action recognition. *Multimedia Tools and Applications*, 80(11): 16185–16203.

Sanchez-Caballero, A.; de López-Diz, S.; Fuentes-Jimenez, D.; Losada-Gutiérrez, C.; Marrón-Romera, M.; Casillas-Perez, D.; and Sarker, M. I. 2022. 3dfcnn: Real-time action recognition using 3d deep neural networks with raw depth information. *Multimedia Tools and Applications*, 1–25.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1010–1019.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Stroud, J.; Ross, D.; Sun, C.; Deng, J.; and Sukthankar, R. 2020. D3d: Distilled 3d networks for video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 625–634.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 6450–6459.

Vapnik, V.; Izmailov, R.; et al. 2015. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16(1): 2023–2049.

Vapnik, V.; and Vashist, A. 2009. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6): 544–557.

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

Wang, J.; Nie, X.; Xia, Y.; Wu, Y.; and Zhu, S.-C. 2014. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2649–2656.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Gool, L. V. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, 20–36. Springer.

Wightman, R. 2019. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models. Accessed: 2022-08-16.

Zhao, J.; Li, R.; and Jin, Q. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2608–2618.