Two-Level Test-Time Adaptation in Multimodal Learning

Jixiang Lei [1] **Franz Pernkopf** [1 2]

Abstract

Test-time adaptation (TTA) aims to modulate parameters of the pre-trained source model utilizing samples from the target domain without accessing the source data. Although recent studies have revealed the high potential of TTA in different computer vision tasks, most TTA methods are constrained to the uni-modal adaptation tasks, while the reliability bias caused by uni-modal data corruption is not sufficiently discussed in multimodal tasks. Although some most recent methods suppressed the cross-modal information discrepancy (i.e. reliability bias) via modulating a modality-sharing module, the domain adaptation for the modality-specific module was neglected. In this paper, we propose a two-level test-time adaptation method (namely 2LTTA) considering both intra-modal distribution shift and cross-modal reliability bias in multimodal learning. 2LTTA modulates all normalization layers and self-attention modules of the encoder corresponding to the corrupted modality and the modality-sharing block. Additionally, we adopted a two-level objective function considering both intra-modal distribution shift and cross-modal reliability bias in the modality fusion block. Shannon entropy with sample reweighting was utilized to reduce the intra-modal distribution shift caused by data corruption. A diversity-promoting loss was employed to reduce the cross-modal information discrepancy. Our experiments demonstrate the superiority of 2LTTA over baseline methods on various data sets.

[1]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, Austria [2]Christian Doppler Laboratory for Dependable Intelligent Systems in Harsh Environments, Graz, Austria. Correspondence to: Jixiang Lei <jixiang.lei@student.tugraz.at>, Franz Pernkopf <pernkopf@tugraz.at>.

Published at ICML 2024 Workshop on Foundation Models in the Wild. Copyright 2024 by the author(s).

1. Introduction

Multimodal pre-trained models (Li et al., 2023; Radford et al., 2021) have been widely applied in vision, natural language and multimodal tasks. Although remarkable performance was achieved with such a paradigm, the distribution shift between source and target domain was neglected. To approach domain shift problems in research and industrial applications, many unsupervised domain adaptation (UDA) methods have been proposed. Test-time adaptation (TTA) aims to adapt the pre-trained source model to the target domain utilizing unlabeled target samples, independent of the source dataset. Some TTA approaches even outperform the traditional UDA methods (Wang et al., 2020; Niu et al., 2022). In our paper, we focus on TTA for multimodal learning (MML) tasks. Recent TTA approaches mainly deal with intra-modal domain adaptation via feature alignment between source and target domain. To the best of our knowledge, MM-TTA (Shin et al., 2022) is the first paper that applies TTA to MML tasks for joint 2D-3D semantic segmentation, utilizing both intra-modality pseudo-label generation and cross-modality pseudo-label refinement. Like most TTA works, MM-TTA only modified the batch normalization (BN) layers of the pre-trained source model in test time. However, for multimodal tasks, the intra-modality domain shift may give rise to enlarged information discrepancy in the downstream fusion layers (Yang et al., 2024). Yang et al. (2024) proposed a reliable fusion and robust adaptation (READ) method to modulate only the fusion layer in the attention module of the fusion block, while the intra-modality feature extractors were frozen. In contrast, most TTA methods optimize the pre-trained CNN model by manipulating BN layers only. However, in vision Transformers (ViTs) (Dosovitskiy et al., 2020), there exists no BN layers, but layer normalization (LN) instead. Different from BN, LN re-estimates the mean and standard deviation of the input across the dimensions of the input (Ba et al., 2016). As a result, the dimension of affine transformation parameters for LN are also consistent with that of the input sample. Specifically for TTA in ViT, Kojima et al. (2022) propose to re-estimate statistics and modulate affine parameters of LN. Additionally, according to prior work (Lee et al., 2022), model fine-tuning works best considering only the first few convolutional layers. In Lee et al.

(2022), surgical fine-tuning was evaluated within another TTA method named MEMO (Zhang et al., 2022) using CIFAR-10-C (Krizhevsky, 2009; Hendrycks & Dietterich, 2019) with 15 representative corruptions of severity level 5. In this paper, we propose a novel parameter updating strategy to further improve the model performance for MM tasks. For the intra-modality level, we propose to modulate attention modules of the Transformer encoder (Dosovitskiy et al., 2020), corresponding to a feature extractor (Liang et al., 2020) in CNN-based model, for the corrupted modality. For the cross-modality level, attention modules including fusion layer and projection layer of the fusion block are modified. Besides, LN modulation is implemented for both the Transformer encoder of the corrupted modality and the fusion block. The parameters of the Transformer encoder for the uncorrupted modality are fixed.

In our experiments, we evaluate 2LTTA on different audio-visual datasets using a ViT-based architecture called Contrastive Audio-Visual Masked Auto-Encoder (CAV-MAE) (Gong et al., 2022). To conclude, our contributions are summarized below:

- We propose a two-level test-time adaptation approach by specifically defining Shannon entropy as objective for the Transformer encoder of the corrupted modality and a diversity-promoting loss as objective for the modality fusion block.
- We adopt a novel fine-tuning strategy that covers shallow Attention modules of the Transformer encoder and LN layers of the pre-trained CAV-MAE.
- The proposed approach significantly improved the test-time adaptation performance for Transformer-based models on various data sets.

2. Two-Level Test-Time Adaptation

2.1. Problem Definition

Most TTA-related methods are evaluated on uni-modal learning scenarios with corruption. In multimodal learning scenarios, the correlation between uni-modality corruption and information discrepancy between modalities is often not considered, which may lead to severe degradation in modality fusion. Specifically, once some modalities are contaminated with distribution shifts, the information discrepancy between modalities is enlarged, leading to multimodal reliability bias (Yang et al., 2024). However, most existing objective functions are intended for uni-modal learning scenarios. For that reason, an objective function needs to be carefully designed to reduce both intra-modality domain shift and cross-modality reliability bias.

In terms of parameter updating strategy, T3A (Iwasawa & Matsuo, 2021) is the only TTA method that adapted classifier layers only. However, according to TTA benchmarks in Yu et al. (2023), the prediction accuracy of T3A is much lower than those freezing the FC layers and adapt FE instead. Furthermore, it is shown in Lee et al. (2022) that tuning only shallow convolutional layers outperforms tuning all layers. Nevertheless, in most MML scenarios, where ViT-based models are deployed, surgical fine-tuning can not be directly applied since ViTs have no convolutional layers. Considering the very different parameter update approaches of previous works, we seek to obtain an optimal parameter fine-tuning strategy for ViT-based MML models, in order to achieve a reliable and satisfactory prediction performance.

2.2. Methodology

Two-Level Objective Function. Inspired by TENT (Wang et al., 2020), we introduce the Shannon entropy $E_{ent}(\boldsymbol{x};\tilde{\boldsymbol{\Theta}})$ based on the output embeddings of the Transformer encoder as objective function at intra-modality level. It is defined as:

$$E_{ent}(\boldsymbol{x};\tilde{\boldsymbol{\Theta}}) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{\tilde{y}\in\tilde{\mathcal{C}}} p(\tilde{y}|\mathbf{x}_i;\tilde{\boldsymbol{\Theta}})\log p(\tilde{y}|\mathbf{x}_i;\tilde{\boldsymbol{\Theta}}), \tag{1}$$

where $p(\tilde{y}|\mathbf{x}_i;\tilde{\boldsymbol{\Theta}})$ denotes the softmax of the output embedding of sample $\mathbf{x}_i$ over the Transformer encoder of the corrupted modality and $\tilde{\mathcal{C}}$ is the corresponding output space (i.e. the number of elements in the output embedding), while $\tilde{y}\in\tilde{\mathcal{C}}$. B is the size of the mini-batch. Following EATA (Niu et al., 2022), we re-weight the entropy loss of a mini-batch by a pre-defined weighting function. The optimization objective for adaptation of the Transformer encoder at intra-modality level can be written as:

$$\min_{\tilde{\boldsymbol{\Theta}}_i, i\in\hat{S}} \quad S_{ent}(\boldsymbol{x})E_{ent}(\boldsymbol{x};\tilde{\boldsymbol{\Theta}}), \tag{2}$$

where S_{ent} is the weighting function (Niu et al., 2022) specified as:

$$S_{ent}(\boldsymbol{x}) = \frac{1}{\exp\left(E_{ent}(\boldsymbol{x};\tilde{\boldsymbol{\Theta}}) - 0.4\cdot E_{max}\right)}, \tag{3}$$

$E_{max} = \ln\tilde{C}$. For cross-modality level, we introduce a diversity-promoting loss to reduce the influence of the information discrepancy caused by modality corruption. The Shannon entropy, as described in Eqn. 1, aims to encourage the adapted model to shape a discriminative output close to one-hot encoding. However, in practice, the ideal target outputs should be similar to one-hot encoding but differ from each other (i.e. diverse over the samples) (Liang et al., 2020) (Krause et al., 2010). For this purpose, we propose to add a regularization term to the objective function to account for the diversity over the data samples as follows:

$$E_{div}(\boldsymbol{x};\boldsymbol{\Theta}) = \sum_{y\in\mathcal{C}}\hat{p}_y\log\hat{p}_y, \tag{4}$$

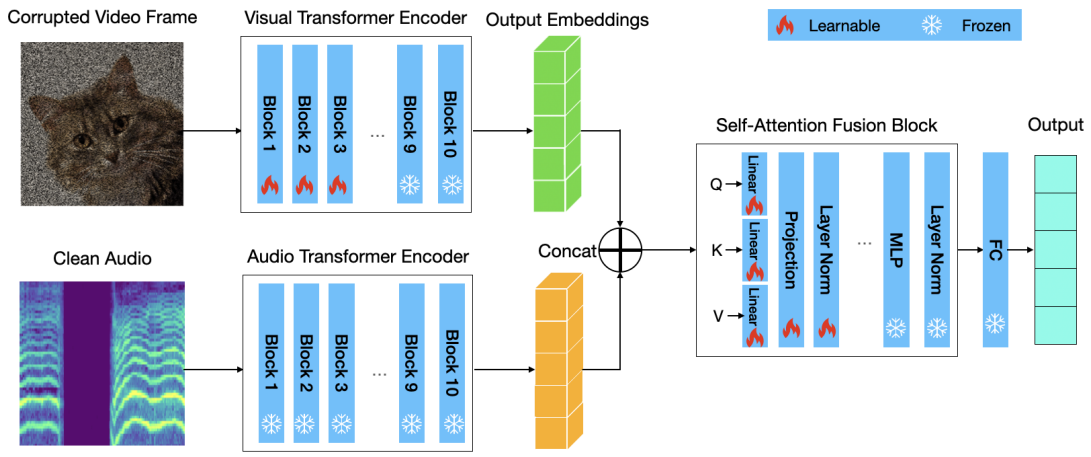

Figure 1. Pipeline of 2LTTA framework. The visual modality is corrupted while the audio modality is clean. All transformation parameters in LN are trainable as proposed by TENT (Wang et al., 2020). Inspired by Lee et al. (2022), shallow Attention modules in the Transformer encoder of the corrupted modality are learnable while deep Attention modules and deep FC layers are frozen. In the self-attention fusion block, the Attention-based linear fusion layers are learnable. A two-level objective function is adopted in order to improve prediction accuracy of the adapted model during test time.

where $\hat{p}_y$ is the average of softmax output $\hat{p}_y = \frac{1}{B}\sum_{i=1}^{B} p(y|\mathbf{x}_i;\mathbf{\Theta})$ of all test samples in each mini-batch of size B. $\mathcal{C}$ is the output space of the fusion block (i.e. the set of classes). E_{div} is termed as diversity-promoting objective.

Parameter Updating Strategy. Inspired by surgical fine-tuning in Lee et al. (2022) we propose fine-tuning of shallow Attention modules in addition to LN layers in the Transformer encoder of the corrupted modality (see Figure 1). Similarly, the fusion layer, projection layer and all LN in the fusion block are also learnable. Similar to Lee et al. (2022), we optimize parameters only with respect to a subset $\hat{S}$ of layers as described in the following:

$$\min_{\mathbf{\Theta}_i, i\in\hat{S}} E(\boldsymbol{x};\mathbf{\Theta}), \tag{5}$$

where $\mathbf{\Theta} = \{\mathbf{\Theta}_1, \mathbf{\Theta}_2, ..., \mathbf{\Theta}_i, ..., \mathbf{\Theta}_n\}$, $\mathbf{\Theta}_i$ denotes parameters in the ith layer and $E(\cdot;\cdot)$ is the objective function. This means that only those parameters belonging to the surgery subset $\hat{S}$, i.e. $\mathbf{\Theta}_i, i \in \hat{S}$ are trainable, while the non-surgery parameters $\mathbf{\Theta}_j, j \notin \hat{S}$, are fixed to the pre-trained source model. The objective function of 2LTTA is written as follows:

$$\min_{\mathbf{\Theta}_i, i\in\hat{S}; \tilde{\mathbf{\Theta}}_j, j\in\tilde{S}} S_{ent}(\boldsymbol{x})E_{ent}(\boldsymbol{x};\tilde{\mathbf{\Theta}}) + \alpha E_{div}(\boldsymbol{x};\mathbf{\Theta}), \tag{6}$$

where $\alpha > 0$ is the diversity-promoting hyper-parameter. Specifically, $\hat{S}$ denotes the subset of trainable parameters in the whole network, while $\tilde{S}$ is termed as the subset of trainable parameters in the Transformer encoder of the corrupted modality only. Results on 2LTTA are reported in Section 3.

3. Experiments

3.1. Experiment Setup

Datasets and Networks. Our study focused on the evaluation of the proposed 2LTTA in the context of image recognition tasks. Empirical studies on the two widely-used multimodal datasets Kinetics (Kay et al., 2017) and VGGSound (Chen et al., 2020), were conducted. Following Yang et al. (2024), 15 types of corruptions for video modality and 6 for audio modality were introduced. For each corruption, 5 different levels of severity were defined. As a result, the two benchmarks named Kinetics-C and VGGSound-C (Yang et al., 2024) are generated with either corrupted audio or corrupted video. We define the clean benchmarks Kinetics and VGGSound as source domain and the corrupted benchmarks Kinetics-C and VGGSound-C as target domain. In our experiments we mainly focus on the validation with target data of high severity level, in order to check in which extent the pre-trained source model can be improved at worst corruption case. The ViT-based CAV-MAE (Gong et al., 2022) is the backbone of all pre-trained models utilized in our paper. The CAV-MAE model consists of 10 modality-specific blocks (i.e. Transformer encoder) and 1 modality-sharing block (i.e. modality fusion block). As a result, we define an objective function for the Transformer encoder (i.e. intra-modality level) and fusion block (i.e. cross-modality level), respectively.

Fine-Tuning Details We modulate all LN layers in the CAV-MAE model. Additionally, projection layers in the first 5 Attention modules in the Transformer encoder of the corrupted modality are also modulated. For the fusion block, we re-estimate not only fusion layers for Q, V and

Table 1. Results on Kinetics50-C benchmark with corrupted video modality (severity level 5) using CAV-MAE as backbone and comparison to state-of-the-art. The results are averaged over 5 runs. The best results are highlighted in bold.

Methods	Gaussian	Shot	Impulse	Defocus	Glas	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	Jpeg	Average
CAV-MAE (baseline)	46.8	48.0	46.9	67.5	62.2	70.8	66.7	61.6	60.3	46.7	75.2	52.1	65.7	66.5	61.9	59.9
TENT	46.3	47.0	46.3	67.2	62.5	71.0	67.6	63.1	61.1	34.9	75.4	51.6	66.8	67.2	62.7	59.4
EATA	46.8	47.6	47.1	67.2	62.7	70.6	67.2	62.3	60.9	46.7	75.2	52.4	65.9	66.8	62.5	60.1
SAR	46.7	47.4	46.8	67.0	61.9	70.4	66.4	61.8	60.6	46.0	75.2	52.1	65.7	66.4	62.0	59.8
READ	49.4	49.7	49.0	68.0	65.1	71.2	69.0	64.5	64.4	57.4	75.5	53.6	68.3	68.0	65.1	62.5
1LTTA of visual intra-modality (Ours)	49.6	51.5	49.9	66.7	64.7	70.6	67.2	62.1	63.6	48.3	75.0	55.7	67.8	71.9	67.9	62.2
1LTTA of cross-modality (Ours)	53.7	54.3	53.6	69.0	68.0	72.8	70.5	65.4	67.1	62.7	76.4	56.5	70.8	70.8	69.7	65.4
2LTTA (Ours)	**54.6**	**55.6**	**55.0**	**68.7**	**69.5**	**72.6**	**70.4**	**65.6**	**67.4**	**62.8**	**75.3**	**58.8**	**72.1**	**73.1**	**70.8**	**66.2**

K as proposed in Yang et al. (2024), but also the projection layer. In Section D.1, a sensitivity study regarding the number of fine-tuned shallow Attention modules is introduced. Furthermore, to study the contribution of model adaption at different levels, i.e. intra-modality and cross-modality, we performed TTA for the Transformer encoder of the corrupted modality (namely 1LTTA:intra-modality) using $S_{ent}(\boldsymbol{x})E_{ent}(\boldsymbol{x};\tilde{\boldsymbol{\Theta}})$ only. For the TTA adaptation with respect to both the Transformer encoder of the corrupted modality and the fusion block (namely 1LTTA: cross-modality), we used $E_{div}(\boldsymbol{x};\boldsymbol{\Theta})$ only.

Implementation Details In the experiments, we use the same hyper-parameters as READ implemented in Yang et al. (2024). In detail, we update parameters in the source model using Adam optimizer. The learning rate is 0.0001 for every mini-batch of size 64 within a single epoch. In all experiments, the hyper-parameter α in Eq. 6 is fixed as 0.5 for all settings. All evaluations are run on Ubuntu 20.04 platform with NVIDIA A100 G40 GPU. Due to space limitation, more implementation details and results are moved to the Appendix.

3.2. Comparisons with State-of-the-Art

Results on Kinetics50-C with corrupted video modality. We compared our proposed 2LTTA with state-of-the-art TTA methods in Table 1 using CAV-MAE (Gong et al., 2022) on Kinetics50-C with severity level 5. According to the results, our proposed 1LTTA and 2LTTA achieved the best performance in all 15 corruption domains. To be specific, 1LTTA with respect to the visual intra-modality significantly outperformed all other methods, but slightly underperformed READ, i.e., 62.2% vs. 62.5% average accuracy over 15 corruption types. In contrast, 1LTTA on cross-modality level performed 2.9% better than READ and achieved a prediction accuracy of 65.4% average accuracy over 15 corruption types. Considering adaption in both intra-modality and cross-modality level, 2LTTA further improved the performance up to 66.2%. It indicates that our robust two-level test-time adaptation method works best, even on datasets with severe corruption.

Results on Kinetics50-C with corrupted audio modality. We evaluated our proposed methods on Kinetics50-C with corrupted audio modality. As shown in Table 2, 2LTTA achieved the best averaged accuracy over 5 independent runs on 6 different corruption types with severity level 5. To be specific, 1LTTA applied at intra-modality level performed nearly the same as READ, while 1LTTA in the cross-modality setting and 2LTTA achieved the best result compared to state-of-the-art. It is worth mentioning that the audio modality in the Kinetics dataset contains less task-specific information for the event classification task, compared to the visual modality. In other words, adaptation of the Transformer encoder of the less informative audio modality led to less performance improvement, compared to adapting the corrupted video encoder in Table 1.

Table 2. Results on Kinetics50-C benchmark with corrupted audio modality (severity level 5) using CAV-MAE as backbone and comparison to state-of-the-art. The results are averaged over 5 runs. The best results are highlighted in bold.

Methods	Gaussian noise	Traffic	Crowd	Rain	Thunder	Wind	Average
CAV-MAE (baseline)	73.7	65.5	67.9	70.3	67.9	70.3	69.3
TENT	73.9	67.4	69.2	70.4	66.5	70.5	69.6
EATA	73.7	66.1	68.5	70.3	67.9	70.1	69.4
SAR	73.7	65.4	68.2	69.9	67.2	70.2	69.1
READ	74.1	69.0	69.7	71.1	71.8	70.7	71.1
1LTTA of audio intra-modality (Ours)	73.9	68.7	69.7	70.7	72.6	70.2	71.0
1LTTA of cross-modality (Ours)	74.7	70.0	71.4	71.8	73.4	71.5	72.1
2LTTA (Ours)	**75.0**	**70.7**	**71.9**	**71.9**	**73.7**	**71.7**	**72.5**

4. Conclusion

In this paper, we propose a two-level test-time adaptation method for MML tasks, to improve the transferability of pre-trained multimodal models to a potentially shifted target domain via adaptation in test time. We further fine-tune LN layers and shallow Attention modules in order to adjust the feature extractor better to the target domain with distribution shift. Furthermore, we introduced a diversity-promoting loss for adaptation at the cross-modality level in addition to the Shannon entropy loss for the intra-modality level. The experimental results indicate that 1LTTA and 2LTTA outperform most state-of-the-art methods on corruption datasets Kinetics50-C and VGGSound-C with corruption of the visual and audio modality, respectively. It is worth pointing out that our method can work on a wide range of ViT-based MML tasks.

References

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Chen, D., Wang, D., Darrell, T., and Ebrahimi, S. Contrastive test-time adaptation, 2022.

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. URL https://api.semanticscholar.org/CorpusID:57246310.

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

Gong, Y., Rouditchenko, A., Liu, A. H., Harwath, D., Karlinsky, L., Kuehne, H., and Glass, J. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations, 2019.

Iwasawa, Y. and Matsuo, Y. Test-time classifier adjustment module for model-agnostic domain generalization. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 2427–2440. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/1415fe9fea0fa1e45dddcff5682239a0-Paper.pdf.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Kojima, T., Matsuo, Y., and Iwasawa, Y. Robustifying vision transformer without retraining from scratch by test-time class-conditional feature alignment. *arXiv preprint arXiv:2206.13951*, 2022.

Krause, A., Perona, P., and Gomes, R. Discriminative clustering by regularized information maximization. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/42998cf32d552343bc8e460416382dca-Paper.pdf.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Kurmi, V. K., Subramanian, V. K., and Namboodiri, V. P. Domain impression: A source data free domain adaptation method, 2021.

Lee, Y., Chen, A. S., Tajwar, F., Kumar, A., Yao, H., Liang, P., and Finn, C. Surgical fine-tuning improves adaptation to distribution shifts. *arXiv preprint arXiv:2210.11466*, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Li, R., Jiao, Q., Cao, W., Wong, H.-S., and Wu, S. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Li, Y., Wang, N., Shi, J., Liu, J., and Hou, X. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

Liang, J., Hu, D., and Feng, J. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pp. 6028–6039. PMLR, 2020.

Litrico, M., Bue, A. D., and Morerio, P. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation, 2023.

Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., and Tan, M. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022.

Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., and Tan, M. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation, 2019.

Peng, X., Wei, Y., Deng, A., Wang, D., and Hu, D. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8238–8247, 2022.

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

Saito, K., Kim, D., Sclaroff, S., Darrell, T., and Saenko, K. Semi-supervised domain adaptation via minimax entropy, 2019.

Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., and Bethge, M. Improving robustness against common corruptions by covariate shift adaptation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11539–11551. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/85690f81aadc1749175c187784afc9ee-Paper.pdf.

Shin, I., Tsai, Y.-H., Zhuang, B., Schulter, S., Liu, B., Garg, S., Kweon, I. S., and Yoon, K.-J. Mm-tta: Multi-modal test-time adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16928–16937, 2022.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., and Hardt, M. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation, 2017.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Wang, Q., Fink, O., Gool, L. V., and Dai, D. Continual test-time domain adaptation, 2022.

Yang, M., Li, Y., Zhang, C., Hu, P., and Peng, X. Test-time adaptation against multi-modal reliability bias. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=TPZRq4FALB.

Yeh, H.-W., Yang, B., Yuen, P. C., and Harada, T. Sofa: Source-data-free feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 474–483, January 2021.

Yu, Y., Sheng, L., He, R., and Liang, J. Benchmarking test-time adaptation against distribution shifts in image classification, 2023.

Zhang, M., Levine, S., and Finn, C. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

A. Related Work

Related work on TTA, parameter update strategy and multimodal objective function is discussed. A summary of different adaptation methods is shown in Table 3.

Table 3. Domain adaptation approaches differ by their accessibility to source/target data and the definition of the loss function. Parameter update strategies consider either feature extractor (FE), batch normalization (BN) or layer normalization (LN). Sometimes other parameters are also considered for tuning. $\mathcal{X}^s$ and $\mathcal{X}^t$ denote source and target data, while $\mathcal{Y}^s$ and $\mathcal{Y}^t$ are the source and target labels.

Setting	Applicable Model	Source Data	Target Data	Train Loss	Test Loss	Parameter Update
Fine-tuning	CNN, ViT	–	$\mathcal{X}^t, \mathcal{Y}^t$	$\mathcal{L}(\mathcal{X}^t, \mathcal{Y}^t)$	–	FE
UDA	CNN	$\mathcal{X}^s, \mathcal{Y}^s$	$\mathcal{X}^t$	$\mathcal{L}(\mathcal{X}^s, \mathcal{Y}^s) + \mathcal{L}(\mathcal{X}^s, \mathcal{X}^t)$	–	FE
Test-time training (TTT) (Sun et al., 2020)	CNN	$\mathcal{X}^s, \mathcal{Y}^s$	$\mathcal{X}^t$	$\mathcal{L}(\mathcal{X}^s, \mathcal{Y}^s) + \mathcal{L}(\mathcal{X}^s)$	$\mathcal{L}(\mathcal{X}^t)$	FE
TENT (TTA)	CNN	–	$\mathcal{X}^t$	–	$\mathcal{L}(\mathcal{X}^t)$	BN
CFA	ViT	$\mathcal{X}^s$	$\mathcal{X}^t$	–	$\mathcal{L}(\mathcal{X}^s, \mathcal{X}^t)$	LN
L2TTA	ViT	–	$\mathcal{X}^t$	–	$\mathcal{L}(\mathcal{X}^t)$	LN + shallow Attention modules

Test Time Adaptation (TTA). Unlike traditional UDA, TTA requires only the pre-trained source model and unlabeled target data $\mathcal{X}^t$ for adaptation. Some previous works (Yeh et al., 2021; Kurmi et al., 2021; Li et al., 2020) achieve domain alignment using generative models without access to source data. In Kurmi et al. (2021) a labeled source dataset is reconstructed by feature alignment of generated source data and unlabeled target data while maintaining the prediction accuracy of the pre-trained classifier. However, this approach requires multiple auxiliary networks, which makes the training inefficient. The prediction results are also not competitive compared to other works. Another popular direction is to optimize the pre-trained model without domain alignment. Test-time entropy minimization (TENT) (Wang et al., 2020) optimizes the affine parameters and modulates the normalization statistics in BN layers batch-by-batch. SHOT (Liang et al., 2020) utilizes both an entropy and pseudo-label-based cross-entropy loss for adaptation. Besides, a diversity regularizer is added to the objective function to encourage the target output to be diverse. Specifically for TTA using ViTs, Kojima et al. (2022) proposed a new test-time adaptation method called class-conditional feature alignment (CFA[1]), which minimizes both the class conditional distribution differences and the whole distribution differences of the hidden representation between the source and target.

Parameter Update Strategy. Similar to traditional UDA, there are different parameter update strategies for TTA. AdaBN (Li et al., 2016) and PredBN+ (Schneider et al., 2020) require no fine-tuning for adaptation. Instead, the normalization statistics of the pre-trained source model is modulated using target data batch-by-batch while freezing all other parameters. TENT (Wang et al., 2020), EATA (Niu et al., 2022), SAR (Niu et al., 2023) and CFA (Kojima et al., 2022) require both estimating normalization statistics μ and σ^2 (or higher central momentums for CFA) and fine-tuning the affine parameters via entropy minimization of the test data for adaptation. Similar to TTT (Sun et al., 2020) and ADDA (Tzeng et al., 2017), SHOT (Liang et al., 2020) proposed to freeze the domain-invariant classifier of the pre-trained source model and updates the parameters of the domain-specific feature extractor (FE). In contrast, T3A (Iwasawa & Matsuo, 2021) only adjusts the trained linear classifier (the last layer of the deep neural network) based on pseudo-prototype representation. Furthermore, some other works (Zhang et al., 2022; Wang et al., 2022; Chen et al., 2022; Litrico et al., 2023) propose to adapt the parameter of both feature extractor and classifier. In the work of Yu et al. (2023), the most popular TTA methods were compared and the prediction accuracy for three corruption datasets (CIFAR-10-C (Krizhevsky, 2009; Hendrycks & Dietterich, 2019), CIFAR-100-C (Krizhevsky, 2009; Hendrycks & Dietterich, 2019), ImageNet-C (Hendrycks & Dietterich, 2019; Deng et al., 2009)) and two natural shift datasets (Office-Home (Venkateswara et al., 2017), DomainNet126 (Peng et al., 2019) (Saito et al., 2019)) were evaluated. T3A (Iwasawa & Matsuo, 2021) adjusting only the linear classifier performed worst, while EATA (Niu et al., 2022) and SAR (Niu et al., 2023) optimizing only BN layer outperform other methods on corruption datasets CIFAR10-C, CIFAR100-C and ImageNet-C.

Multimodal Objective Function. To the best of our knowledge, MM-TTA and READ are the only works specifying TTA for MML tasks. MM-TTA adopted the traditional cross-entropy loss objective based on pseudo-labeling. However, when the data corruption in target domain is severe, the performance improvement with pseudo-labeling is limited (Yu et al., 2023). Compared to cross-entropy, Shannon entropy is more noise-resistant and the model performance utilizing Shannon entropy is superior to cross-entropy. However, according to Yang et al. (2024), the Shannon entropy only works for

[1]Although CFA doesnot require the complete source domain data set, the mean and higher order central moments of overall distribution on source data need to be calculated and stored in memory (Kojima et al., 2022).

adapting uni-modality tasks, while in multimodal tasks, the Shannon entropy is not efficient any more. For that reason, a confidence-aware objective function is proposed in READ to hinder the model from overfitting noise.

B. Details about the Benchmarks

Our experiments are based on the two widely-used multimodal datasets Kinetics (Kay et al., 2017) and VGGSound (Chen et al., 2020) with corrupted visual or audio modality (Yang et al., 2024). To be specific,

- Kinetics50 is a subset of the Kinetics (Kay et al., 2017) dataset. It mainly contains videos with human motion-related action classes, sampled from Kinetics400. Each video clip lasts around 10 seconds and is labeled with a single action class. All the videos are collected from YouTube. Following Peng et al. (2022), 50 classes are randomly selected out of Kinetics400 to construct Kinetcis50, with 29204 training pairs and 2466 test pairs (Yang et al., 2024). According to the characteristic of this dataset, the visual modality contains more information compared to its audio modality.
- VGGSound is an audio-visual correspondent dataset consisting of short audio clips extracted from videos uploaded to YouTube (Chen et al., 2020). It covers every day audio events consisting of 309 classes. Each video clip has a fixed duration of 10 seconds. Different from Kinetics50, the audio modality of VGGSound contains more information compared to its visual modality. Following Yang et al. (2024), 14046 testing visual-audio pairs are utilized for TTA.

The visual and audio modality of both datasets are extracted from the original videos following the method proposed in Gong et al. (2022). To comprehensively study the distribution shift of each modality, different corruption types are introduced into the visual and audio modalities. Following Yang et al. (2024), 15 corruption types are applied for visual modality. Each corruption contains 5 severity levels for extensive validations. The 15 corruptions include "gaussian noise", "shot noise", "impulse noise", "defocus blur", "glass blur", "motion blur", "zoom blur", "snow", "frost", "fog", "brightness", "contrast", "elastic transform", "pexelate" and "jpeg compression". Similarly, the audio modality is corrupted by 6 different corruptions, namely "gaussian noise", "traffic noise", "crowd noise", "rain", "thunder" and "wind". For each audio corruption, 5 different severity levels are included. The corrupted benchmarks are called Kinetics50-C and VGGSound-C. The visualization of an example of corrupted video frames and audio spectrograms are shown in Figure 2 and Figure 3, respectively.

C. CAV-MAE Architecture

CAV-MAE is utilized as pre-trained model for multimodal learning. The CAV-MAE encoder consists of 11 Transformer (Attention) blocks for the modality-specific feature extraction, followed by one Transformer for cross-modal fusion. For the video streams, 10 frames are sampled from each video clip and then a single frame is randomly selected and fed into the Transformer encoder of the visual modality. For the audio streams, the original 10-second waveform audio file is firstly converted into a 2 dimensional spectrogram and then fed into the Transformer encoder of the audio modality. More details about the CAV-MAE architecture are provided in Gong et al. (2022).

D. Additional Experimental Results

D.1. Ablation Study

Surgical fine-tuning for Attention modules in the Transformer encoder. The ablation study of surgical fine-tuning for Attention modules in the Transformer encoder of the corrupted modality is shown on Kinetics50-C, where all LN layers are trainable in all the experiments and the so called "surgical" adaptation only involves projection layers of shallow Attention modules in the Transformer encoder. The results are shown in Table 4. First of all, we observe that fine-tuning of all Attention modules underperformed surgical fine-tuning. The results indicate that surgical fine-tuning works best with 5 learnable Attention modules for Kinetics50-C, obtaining 66.5% average accuracy.
Sensitivity study on loss hyper-parameter α. We further investigate the influence of the diversity-promoting loss using hyper-parameter α. As described in Table 5, our 2LTTA achieved the best performance with $\alpha = 0.5$ for Kinetics50-C with corrupted video modality.

D.2. Results on VGGSound-C with Corrupted Audio Modality

Similarly, we implemented our approaches on VGGSound-C with corrupted audio modality. Different from Kinetics, the audio modality in VGGSound contains more task-specific information for the event classification task than the visual

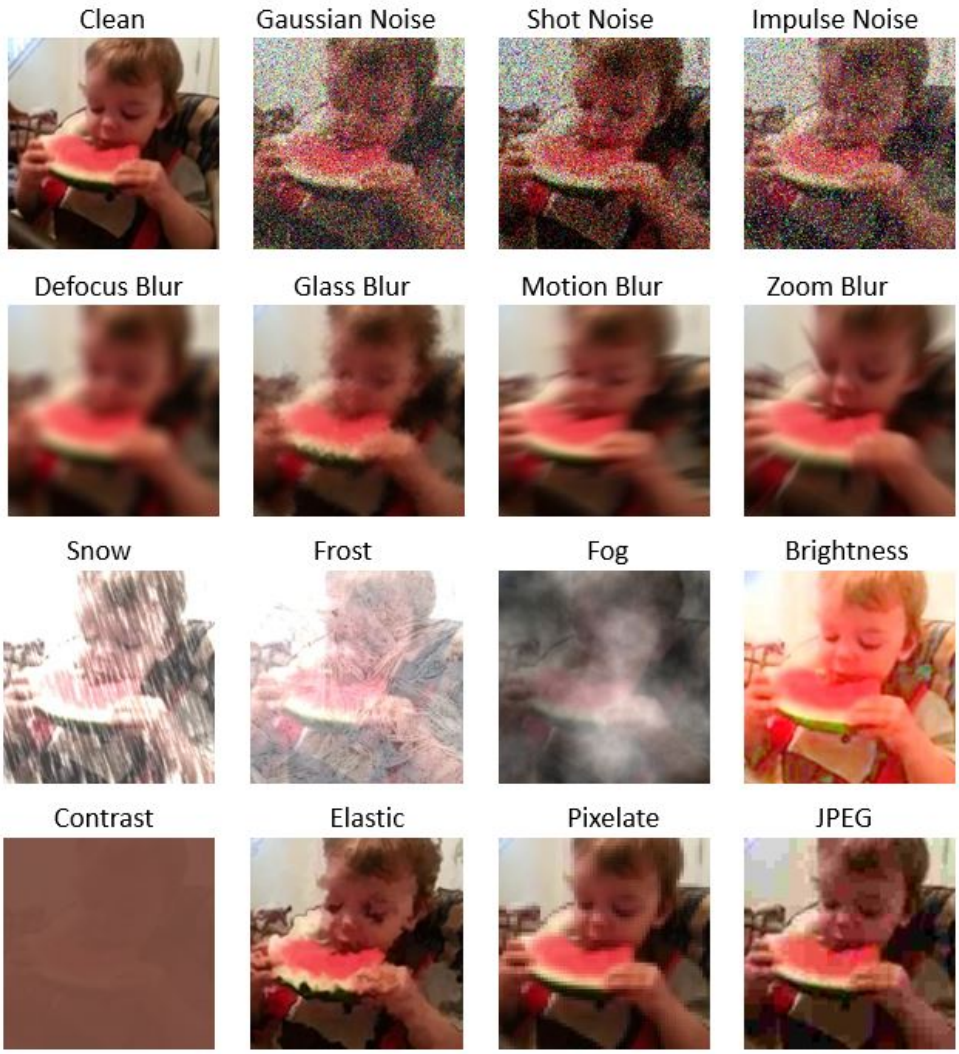

Figure 2. Visualization of a clean video frame from the Kinetis50-C benchmark and the 15 corresponding visual corruption types with severity level 5.

modality. This is reflected in the results as shown in Table 6. The performance improvement of adapting the pre-trained model on VGGSound with corrupted audio modality is remarkable. The best average result of 36.5% was achieved by 1LTTA at cross-modality level. However, 2LTTA is 1% lower than 1LTTA (cross-modality), but still significantly outperforms other methods. It indicates that the Shannon entropy objective does not always work for all dataset with different corruption types.

D.3. Averaged Performance Across All Severity Levels.

To evaluate the robustness of our method against mixed severity levels, the averaged performance across all severity levels is compared with that of READ (Yang et al., 2024) and the source model, respectively. The results with all corruption types for the visual and audio modality are summarized. As shown in Figure 4, for TTA with corrupted visual modality, 2LTTA outperforms READ and the source model in noise and digital corruptions, while the averaged performance on blur and weather corruptions is almost indistinguishable compared to READ. According to Figure 5, the averaged performance accross severity levels of audio modality using 2LTTA is not superior to READ although it performs better than READ on corruption types with severity level 5 (see the main paper).

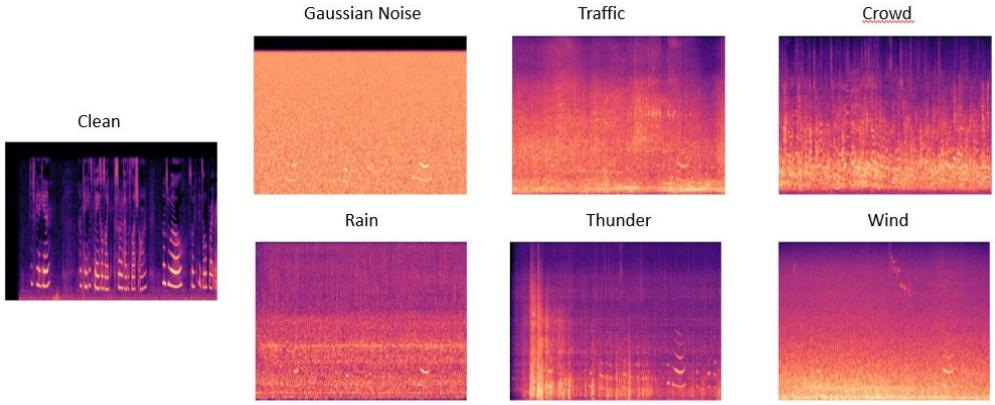

Figure 3. Visualization of a clean audio from the Kinetics50-C benchmark and the 6 corresponding audio corruption types with severity level 5.

Table 4. Accuracy of surgical fine-tuning on Kinetics50-C using pre-trained model CAV-MAE. Surgical fine-tuning is performed with 2LTTA. The surgical fine-tuning is restricted to the Attention module of the Transformer encoder for the corrupted modality only. The results are averaged over 5 runs. The best results are highlighted in bold.

Number of Attention modules in the Transformer encoder for adaptation	Gaussian	Shot	Impulse	Defocus	Glas	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	Jpeg	Average
0	53.4	54.0	53.3	69.1	68.3	72.8	70.2	65.5	66.7	61.8	76.2	56.8	70.9	71.3	69.6	65.3
1	53.8	54.6	53.8	69.3	68.9	72.9	70.8	65.9	67.3	63.2	76.3	57.4	71.7	71.7	70.1	65.9
3	54.6	55.6	55.0	68.7	69.5	72.6	70.4	65.6	67.4	62.8	75.3	58.8	72.1	73.1	70.8	66.2
5	**54.7**	**56.1**	**55.5**	**69.2**	**70.1**	**72.8**	**70.6**	**65.0**	**67.2**	**62.3**	**75.1**	**58.7**	**72.7**	**73.9**	**71.1**	**66.3**
7	54.4	54.7	54.8	68.6	69.5	72.5	70.0	63.5	65.0	59.8	74.6	57.99	73.1	73.7	70.3	65.5
ALL	52.5	52.8	51.7	63.3	65.1	67.4	65.1	60.0	61.7	55.7	70.2	54.0	70.1	69.1	64.3	61.5

D.4. Averaged Performance Across All Corruption Types.

To evaluate the robustness of our method against mixed distribution shifts, the averaged performance across all corruption types is compared with that of READ and the source model, respectively. The results with severity level varying from 1 to 5 for the visual and audio modality are summarized in Figure 6 and 7. 2LTTA performs the best over all severity levels of the visual modality. The more severe the data corruption, the more significant our method is superior to the other methods, i.e. 2LTTA is more robust against corruption severity. However, for the audio modality in Figure 7 we see that the increase of severity level did not lead to significant degradation of averaged performance across corruption types for all 3 methods. That is mainly due to the fact that the audio modality contains fewer information related to the action recognition compared to the visual modality.

D.5. Results on Severity Level 3.

In the main paper, the experimental results are based on corruption types with severity level 5 only. To evaluate the performance of our model on other severity levels, results on severity level 3 are reported in Table 7 - 9 for Kinetics50-C with corrupted visual/audio modality and VGGSound with corrupted audio modality.

Table 5. Sensitivity study for diversity-promoting hyper-parameter α in Eq. 6 on Kinetics50-C benchmark with corrupted video modality (severity 5). The results are averaged over 5 runs. The best result is highlighted in bold.

α	0	0.2	0.5	1.0	3.0
Kinetics50-C with visual corruption	64.5	65.5	**65.9**	65.8	65.6

Table 6. Results on VGGSound-C benchmark with corrupted audio modality (severity level 5) using CAV-MAE as backbone and comparison to state-of-the-art. The results are averaged over 5 runs. The best results are highlighted in bold.

Methods	Gaussian noise	Traffic	Crowd	Rain	Thunder	Wind	Average
CAV-MAE (baseline)	37.0	25.5	16.8	21.6	27.3	25.5	25.6
TENT	10.6	2.6	1.8	2.8	5.3	4.1	4.5
EATA	39.2	26.1	22.9	26.0	31.7	30.4	29.4
SAR	37.4	9.5	11.0	12.1	26.8	23.7	20.1
READ	40.4	28.9	26.6	30.9	36.7	30.6	32.4
1LTTA of audio intra-modality (Ours)	31.8	28.6	27.5	28.4	30.1	29.0	29.2
1LTTA of cross-modality (Ours)	**40.5**	**33.6**	**35.8**	**33.1**	**41.5**	**34.6**	**36.5**
2LTTA (Ours)	39.3	33.0	35.0	32.0	39.6	33.9	35.5

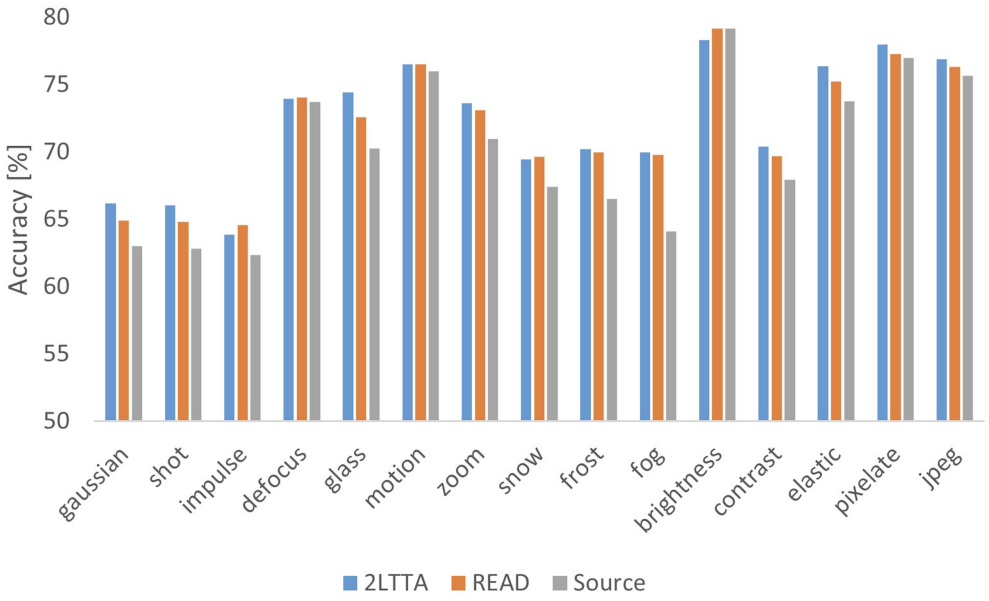

Figure 4. Averaged performance across all severity levels for 2LTTA, READ and the source model on the dataset Kinetics50-C for different corruption types in visual modality.

Table 7. Results on Kinetics50-C with corrupted video modality and severity level 3 using CAV-MAE as backbone and comparison to state-of-the-art. The results are averaged over 5 runs. The best results are highlighted in bold.

Methods	Gaussian	Shot	Impulse	Defocus	Glas	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	Jpeg	Average
CAV-MAE (baseline)	54.1	54.8	54.6	73.5	68.3	76.6	71.5	69.2	64.7	69.5	79.3	72.1	77.6	79.4	75.4	69.4
TENT	54.2	55.1	55.2	73.6	69.6	76.8	71.9	69.5	65.6	70.2	79.4	72.9	78.3	79.2	75.3	69.8
EATA	54.4	54.9	55.0	73.4	69.1	76.5	71.6	69.2	65.1	69.5	79.5	72.3	77.7	79.1	75.2	69.5
SAR	54.2	54.8	55.0	73.1	68.2	76.4	71.1	69.1	64.8	69.4	79.1	72.0	77.4	79.1	75.0	69.2
READ	56.1	56.9	56.4	73.9	70.5	76.6	72.8	70.0	68.1	70.8	79.3	73.3	78.2	79.6	75.6	70.5
1LTTA of visual intra-modality (Ours)	65.7	65.6	66.5	74.0	71.8	77.2	72.4	70.8	66.9	69.7	79.0	72.1	79.0	**79.5**	78.5	72.6
1LTTA of cross-modality (Ours)	63.7	64.2	65.0	73.6	69.6	77.2	71.8	70.9	66.7	70.0	**79.5**	72.3	78.6	79.0	78.4	72.0
2LTTA (Ours)	**66.4**	**66.2**	**67.0**	**74.3**	**72.5**	**77.2**	**72.5**	**71.6**	**68.2**	**70.5**	79.0	**72.7**	**79.5**	79.4	**78.9**	**73.1**

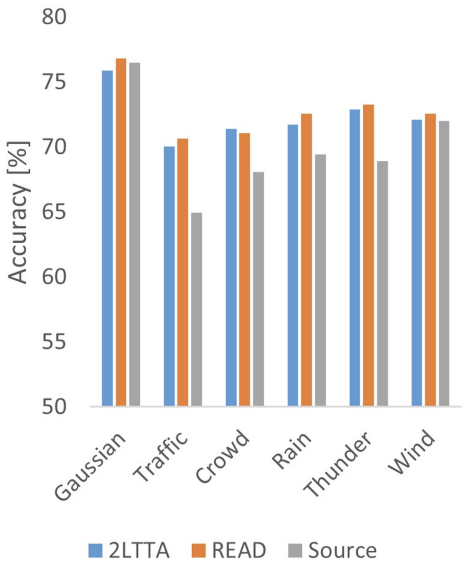

Figure 5. Averaged performance across all severity levels for 2LTTA, READ and the source model on the dataset Kinetics50-C for different corruption types in audio modality.

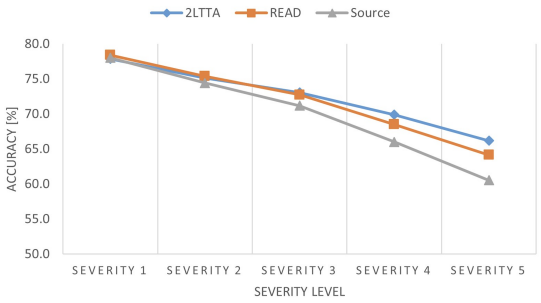

Figure 6. Averaged performance across all corruption types using 2LTTA, READ and the source model on the dataset Kinetics50-C with different severity levels in visual modality.

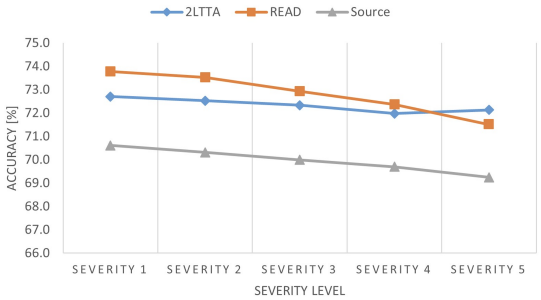

Figure 7. Averaged performance across all corruption types using 2LTTA, READ and the source model on the dataset Kinetics50-C with different severity levels in audio modality.

Table 8. Results on Kinetics50-C with corrupted audio modality and severity level 3 using CAV-MAE as backbone and comparison to state-of-the-art. The results are averaged over 5 runs. The best results are highlighted in bold.

Methods	Gaussian noise	Traffic	Crowd	Rain	Thunder	Wind	Average
CAV-MAE (baseline)	75.9	64.4	68.7	70.3	67.9	70.3	69.3
TENT	73.9	67.4	69.2	69.3	69.0	72.1	70.1
EATA	76.0	65.7	68.9	69.8	69.1	72.1	70.3
SAR	76.0	64.6	68.7	69.3	68.6	72.2	69.9
READ	76.4	69.6	70.8	72.0	72.6	72.3	72.3
1LTTA audio intra-modality (Ours)	**76.5**	64.3	67.8	69.8	69.3	71.9	69.9
1LTTA of cross-modality (Ours)	76.4	**69.8**	70.7	**72.0**	72.8	71.9	72.2
2LTTA (Ours)	76.2	69.7	**70.7**	71.9	**72.8**	**72.2**	**72.3**

Table 9. Results on VGGSound-C with corrupted audio modality and severity level 3 using CAV-MAE as backbone and comparison to state-of-the-art. The results are averaged over 5 runs. The best results are highlighted in bold.

Methods	Gaussian noise	Traffic	Crowd	Rain	Thunder	Wind	Average
CAV-MAE (baseline)	42.1	29.4	19.5	27.6	31.2	29.4	29.9
TENT	8.1	4.0	2.3	4.7	4.8	6.1	5.5
EATA	46.7	30.5	28.0	31.4	35.4	33.8	34.3
SAR	43.1	17.3	8.3	29.0	31.6	30.5	26.6
READ	47.3	32.7	29.9	33.2	38.3	33.7	35.8
1LTTA of audio intra-modality (Ours)	33.6	29.3	28.9	29.4	30.9	29.9	30.3
1LTTA of cross-modality (Ours)	**47.2**	**36.6**	**38.4**	**37.5**	**43.1**	**37.1**	**40.0**
2LTTA (Ours)	46.8	35.9	37.6	35.4	41.3	36.1	38.8