Mitigating Missing Feature Channels at Inference Stage: Test-Time Adaptation through Self-Training with Data Imputation

Yan Huang, Yongyi Su, Xulei Yang, Xin Lin, and Xun Xu*

***Abstract*—The robustness of deep learning model can be compromised by out-of-distribution (OOD) testing data. Test-time adaptation (TTA) emerges as an efficient method to mitigate the distribution gap by tuning model weights at inference stage. TTA are mainly demonstrated on robustifying model on additive visual corruptions or adversarial attacks. In this work, we specify an overlooked type of OOD where feature channels could be missing in testing data, potentially due to sensor fault. We reveal that self-training and data imputation can effectively improve the model's generalization to data with missing feature channel. To address the uncertainty associated with imputed samples, we further propose to fuse the pseudo predictions between imputed and weakly augmented testing samples to produce more reliable pseudo labels. We evaluate the effectiveness on multiple image classification benchmarks with synthesized and realistic missing feature channels, and our proposed method outperforms state-of-the-art TTA methods on all benchmarks.**

***Index Terms*—Test-Time Adaptation, Missing Feature Channel, Self-Training, Data Imputation**

I. Introduction

THE success of deep neural networks hinges on the i.i.d. assumption between training and testing datasets. Despite achieving unprecedented performance on numerous learning tasks, the robustness of deep learning models is vulnerable to the violation of such assumptions. When testing data is not drawn from an identical distribution, also referred to as covariate shift [1], the generalization of deep learning models could be severely compromised. Traditional efforts into mitigating the covariate shift arises from unsupervised domain adaptation [2] which updates model weights upon seeing both source and target domain data. Nevertheless, with the advent of large-scale foundation models [3], training a model simultaneously on large-scale source data and downstream target data becomes impossible due to data privacy and prohibitive computation cost. More importantly, the target distribution is often unknown until the inference stage begins. These challenges have motivated research into test-time adaptation (TTA) at inference stage [4], [5], [6].

TTA aims to adjust model weights to adapt to the out-of-distribution testing data upon observing a stream of testing data [4]. Despite successfully demonstrating the effectiveness on adapting to out-of-distribution testing data, we found that an important scenario is overlooked, where testing data is subject to missing feature channels, potentially due to sensor fault or specific visual tasks [7][8]. Particularly in underwater visual tasks, spectrally-selective light attenuation and backscattering always result in attenuated red spectrum, i.e. missing R channel in RGB images. This phenomenon significantly diminishes the performance of underwater visual tasks. The power of TTA may not be fully unleashed without considering the specific challenges of partially visible feature channels.

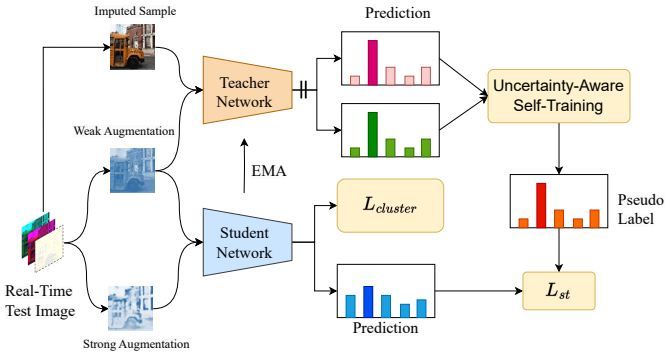

Fig. 1: Illustration of the self-training framework. Predictions on imputed sample and weak augmentation are fused to produce more reliable pseudo labels for self-training.

In this work, we propose to tackle missing feature channels at inference stage by adapting a pre-trained model on-the-fly. Specifically, we are inspired by the success of self-training [9] for test-time adaptation [4] and use a teacher model to generate pseudo labels which are further used to supervise the updating of student network. As direct self-training is susceptible to incorrect pseudo labels [10], we further introduce an efficient data imputation module to directly compensate the negative impact of missing feature channels. This module fits a parametric distribution on source domain data and testing samples with missing feature channels are completed by sampling from the prior distribution. Such a light-weight data imputation mechanism does not require access to the source domain data at inference stage and poses no risk of privacy leakage. As the data imputation alone infers the missing feature channel from a probabilistic perspective, there is guarantee on improving pseudo label quality. Hence, we further propose a uncertainty based pseudo label fusion module to synergistically fuse the predictions from imputed samples and weakly augmented samples. Finally, as self-training is still prone to incorrect

X. Xu and X. Yang are with the Institute for Infocomm Research (I^2R), A*STAR, Singapore. Y. Huang and Y. Su are with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China. X. Lin is with the school of Artificial Intelligence, Guangzhou University. This work is supported by the National Natural Science Foundation of China under Grant 62106078 and by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021). * Correspondence to X. Xu <xu_xun@i2r.a-star.edu.sg>.

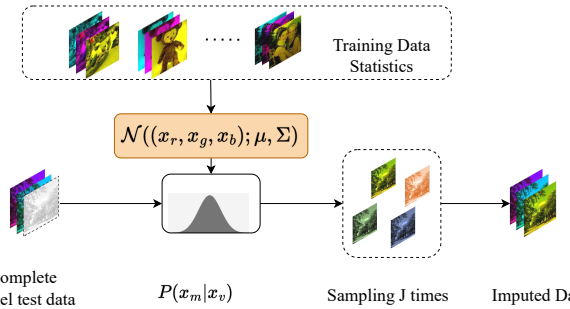

Fig. 2: Illustration of the statistical missing feature channel imputation procedure.

pseudo labels, we introduce an additional deep clustering regularization to increase the robustness of self-training for TTA.

The proposed missing feature channel scenario differs from the traditional task of learning and testing on missing modality data where explicitly fusion of multi-modal data, e.g. text, image, depth, etc, is introduced to the network [11], [12], [13], [14]. In these works, mitigating the impact of missing modality is often achieved at training stage through adversarial training [11], distribution alignment [13], searching fusion policy [12], prompt learning [14], etc. These works are not applicable for tackling missing feature channel at inference stage only.

We summarize the contributions of this work as follows. **i)** We identified an overlooked challenge of generalizing pre-trained models to testing data with missing feature channels, revealing that direct testing on image data with missing feature channels could significantly harm the performance. **ii)** We propose to adapt pre-trained model weights on testing data stream at inference stage through self-training with data imputation and deep clustering regularization. The proposed method does not require access to source domain data at adaptation stage. **iii)** We created a benchmark of test-time adaptation for missing feature channels from four image classification datasets. Our method outperforms generic state-of-the-art test-time adaptation approaches.

II. Methodology

We denote a training dataset $\mathcal{D}_{tr} = \{x_i \in \mathbb{R}^C, y_i \in \mathcal{K}\}$ where each sample consists of C feature channels and the labels space spans K classes, $\mathcal{K} = \{1, \cdots K\}$. A source domain model $f(x; \Theta)$ is pretrained on $\mathcal{D}_{tr}$. The target domain, $\mathcal{D}_{te} = \{x_j \in \mathbb{R}^C\}$, consists of testing data with missing feature channels, e.g. due to sensor failure. We specify the missing channel with a binary vector $v \in \{0, 1\}^C$ with $v_c = 0$ indicating the c-th channel is missing. We aim to update model weights Θ such that the performance on missing channel testing data is improved.

A. Statistical Missing Channel Imputation

Imputing the c-th missing channel can be formulated as building a conditional generative model $p(x_c|\{x_k|k \neq c\})$. Without a proper knowledge on the prior distribution, recent diffusion models have been employed to impute missing values [15], [16]. However, the high computation cost of the generative process of diffusion model renders the method impractical for efficient TTA. In this work, we opt for a light-weight parametric model, a multivariate Gaussian distribution $p(x_1, \cdots x_C) = \mathcal{N}(x_1, \cdots x_C|\mu, \Sigma)$, to model the feature distribution. When feature channels $\mathcal{M} = \{c|\forall c, v_c = 0\}$ are missing, we can easily obtain the conditional distribution for the missing channel $p(\{x_c|c \in \mathcal{M}\}|\{x_k|k \in \mathcal{V}\})$ where $\mathcal{V} = \{c|\forall c, v_c = 1\}$. Specifically, we rearrange the mean and covariance by grouping the missing and visible channels as below.

$$\mu = [\mu_m, \mu_v]^\top, \quad \Sigma = \begin{bmatrix} \Sigma_{mm} & \Sigma_{mv} \\ \Sigma_{vm} & \Sigma_{vv} \end{bmatrix} \tag{1}$$

The conditional probability follows a multivariate Gaussian distribution as follows.

$$\begin{aligned} \mu_{m|v} &= \mu_m + \Sigma_{mo}\Sigma_{vv}^{-1}(x_v - \mu_v) \\ \Sigma_{m|v} &= \Sigma_{mm} - \Sigma_{mv}\Sigma_{vv}^{-1}\Sigma_{vm} \end{aligned} \tag{2}$$

where v and m are the visible and missing channels.

The missing channels can be imputed by sampling from the conditional distribution $p(\{x_c|c \in \mathcal{M}\}|\{x_k|k \in \mathcal{V}\})$. Nonetheless, we empirically observe that direct imputing with the conditional mean μ_m is sufficient to improve model's robustness.

B. Uncertainty-Aware Test-Time Self-Training

We employ a self-training (ST) framework for adapting pre-trained model to target domain with missing feature channels. ST has been demonstrated to be effective for test-time adaptation to different types of corruptions [4]. Precisely, we adopt a teacher-student architecture [17] as illustrated in Fig. 1. To maintain high stability of pseudo labeled, the teacher network $\tilde{\Theta}$ is the exponential moving average of the student network Θ. We denote the probablistic prediction of teacher model given input x_i as $\phi(x_i; \Theta) = \sigma(f(x_i; \tilde{\Theta})) \in [0, 1]^K$ where σ refers to the softmax function. The existing self-training strategy adopted in TTA [4] often applies a simple weak augmentation $\mathcal{W}$, which consists of RandomCrop and RandomHorizontalFlip, to the input sample for teacher network. However, we argue that directly predicting on the input with weak augmentation is susceptible to the missing feature channels. Therefore, we further introduce the statistical data imputation introduced in the previous section as an additional data augmentation, denoted as $\mathcal{I}(x)$. The two predictions are fused by their respective confidence to produce the final pseudo label for self-training as follows where $H(q)$ measures the entropy for distribution q .

$$\tilde{q}_i = \frac{e^{-H(\phi(\mathcal{I}(x_i)))}\phi(\mathcal{I}(x_i)) + e^{-H(\phi(\mathcal{W}(x_i)))}\phi(\mathcal{W}(x_i))}{e^{-H(\phi(\mathcal{I}(x_i)))} + e^{-H(\phi(\mathcal{W}(x_i)))}} \tag{3}$$

With the fused predictions by the teacher model, we define the self-training loss as the confidence weighted cross-entropy loss as in Eq. 4. The per-sample weight is defined as $w(q_i) = e^{-\beta H(q_i)}$, the strong augmentation $\mathcal{S}(\cdot)$ can be realised as RandAugment [18] and the student probabilistic prediction writes $p_i = \phi(\mathcal{S}(x_i); \Theta)$.

$$\mathcal{L}_{st} = -\frac{1}{N}\sum_{i=1}^{N} w(\tilde{q}_i)\tilde{q}_i^\top \log p_i \tag{4}$$

C. Clustering Regularization for Self-Training

Self-training is prone to incorrect pseudo labels and applying self-training alone for TTA could suffer if the pseudo labels are less accurate [4]. Inspired by the practice of using clustering to reveal the cluster structure of target data in unsupervised domain adaptation[19], we use deep clustering

[20] to learn a discriminative model from the test data with incomplete channels, for more reliable self-training.

Deep clustering [20] discovers cluster structures in the testing data by iteratively updating a target distribution and minimizing the distance between target and predicted distribution. In the context of test-time adaptation for missing feature channel, we also hope to encourage the testing samples to form cluster structures. To this end, we introduce the clustering loss by minimizing the KL-Divergence between a target distribution $\hat{p}$, which can be seen as a sharpened probabilistic prediction by the student model, and the student prediction p as below.

$$KL(\hat{p}||p) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\hat{p}_{ik}\log\hat{p}_{ik}/p_{ik} \tag{5}$$

Moreover, we follow the assumption of classical clustering and consider the pseudo labels to follow a uniform distribution, thereby leading to a more class balanced predictions by the student, thus mitigating the degenerate solutions. We achieve this objective by minimizing the following KL divergence loss,

$$KL(\boldsymbol{s}\|\boldsymbol{u}) = \frac{1}{N}\sum_{k=1}^{K} s_k \log\frac{s_k}{u_k} \tag{6}$$

where u_k is the uniform distribution prior and $s_k = \frac{1}{N}\sum_{i=1}^{N}\hat{p}_{ik}$ is the empirical frequency of $\hat{p}$. The final deep clustering regularization combines the two objects as, $\mathcal{L}_{cluster} = KL(\hat{p}||p) + KL(s||u)$, where the target distribution can be obtained from a closed-form solution [20] as,

$$\hat{p}_{ik} = \frac{p_{ik}/\left(\sum_{i'=1}^{N} p_{i'k}\right)^{\frac{1}{2}}}{\sum_{k'=1}^{K} p_{ik'}/\left(\sum_{i'=1}^{N} p_{i'k'}\right)^{\frac{1}{2}}} \tag{7}$$

Augmented Online Clustering With Memory Bank: When working with small batch sizes or numerous categories, we employ a memory bank to enhance $\hat{p}$ by utilizing the stored features from previously batches. In particular, for each iteration, we use a memory bank $\mathcal{M}$ to retain the features from previous batches, the probabilistic prediction is then computed by the last fc classifier layer, which is subsequently employed to compute an enhanced distribution $\hat{p}$ in Eq. (7). We retain the most recent samples in the past 256 minibatches in $\mathcal{M}$. Note that when computing $\mathcal{L}_{cluster}$, we only use the enhanced $\hat{p}$ from the current batch.

D. Final TTA Objective

We combine the self-training loss and the clustering loss as the final optimization objective for TTA, $\mathcal{L}_{total} = \mathcal{L}_{st} + \lambda\mathcal{L}_{cluster}$. For each testing sample, we first predict the label through student model and the gradient is accumulated. Gradient descent is implemented after observing a whole minibatch of testing samples. The updated model is used for inference for the subsequent testing samples.

III. Experiments

A. Implementation Details

We conduct our experiments on 4 classic image classification datasets. **CIFAR-10** consists of 50,000 training images and 10,000 test images, with a total of 10 classes. **CIFAR-100** comprises a total of 100 classes, with each class encompassing 600 images. where 500 are used for training and the remaining 100 for testing. **ImageNet** dataset contains 1000 classes with 1.28 million diverse training images and 50000 validation images. **Underwater-ImageNet** dataset [21] consists of underwater visual images with 7 classes, comprising 6,817 distorted underwater images along with their corresponding undistorted versions, and an additional 1,813 distorted test images.

To simulate missing feature channels, we randomly pick m channels from an n-channel dataset for masking ($m < n$) and initialize the masked pixel values to zero to ensure compatibility with DNN models. We particularly focus on three-channel RGB images with 1 or 2 simulated missing channels in the first 3 classical datasets and also validate our ideas on a realistic underwater dataset. We use ResNet-50[22] pretrained on respective clean training dataset as our backbone model [23], [4]. We set $\alpha = 0.99$ for the EMA scale and $\beta = 0.5$ for the scale of uncertainty weight. The weight of clustering loss is set with $\lambda = 1.0$ for standard clustering and $\lambda = 3.0$ for memory bank augmented online clustering. We use SGD optimizer with batch size of 256, momentum of 0.9 and a weight decay of 5e-4. The learning rate is set by $\text{lr} = 0.0001$ for ImageNet and $\text{lr} = 0.005$ for others.

B. Comparisons with TTA Methods

We compare our proposed method with several test-time adaptation methods. **Test** implements inference directly with the pre-trained model. **TestBN** updates batch normalization statistics on the test data at test time. **TENT** [5] updates the BN layers' parameters by minimizing the self entropy of the target data. **SHOT** [24] only updates the feature extractor by proposing a self-supervised pseudo labeling module. **TTAC** [4] alleviates the domain shift via class-wise feature distribution alignment between source and target data. **TIPI** [25] enforces the network to be invariant under some input transformations, using a regularizer based on their derived bound. Finally, we evaluate our proposed method (**Ours**) and **DDPM** [26] replaces our statistical imputation to DDPM generation while retaining other modules.

Test-Time Adaptation Results: We present the results of test-time adaptation on missing feature channel testing data on ImageNet in Tab. I (left). When test data is contaminated with missing channels, the performance drops substantially from 76.13% $\rightarrow$ 41.39% with the average accuracy of the six type of missing channels, which indicates that the absence of feature channels poses a significant challenge to model generalization. When Test BN is adapted to TTA, we observe a further drop of performance. SHOT and TTAC demonstrates superior performance compared to Test BN and TENT, which suggests that only updating BN layer is not enough to tackle the distribution shift caused by missing feature channels. Moreover, the success of SHOT underscores the effectiveness of self-supervised or self-training methods. Distribution alignment is another effective approach to address the problem of missing feature channels, according to the result of TTAC. We also find that the new work TIPI perform poorly in our missing feature channels setting, even worse than TTAC. Finally, our proposed method combines the self-training with channel imputation and the clustering regularization, achieving the best performance. Specifically, our method based on statistical imputation slightly outperforms the DDPM-based approach. We

also reach similar conclusions on the other datasets in Tab. II, Tab. III and Tab. I (right) respectively. The improvement of our method is more significant on the more challenging CIFAR-100 and underwater dataset.

TABLE I: Accuracies(%) of six missing channel scenarios on ImageNet and Underwater dataset.

Method	Missing R	Missing G	Missing B	Missing R&G	Missing R&B	Missing G&B	Avg.	Underwater ImageNet
clean data				76.13				/
Test	56.42	33.07	61.05	23.38	36.92	37.50	41.39	57.09
TestBN	26.07	29.65	29.37	14.34	24.38	20.21	24.00	19.97
TENT[5]	66.01	68.19	68.48	50.25	59.73	58.00	61.78	59.46
SHOT[24]	70.38	71.41	72.04	60.17	65.37	63.69	67.18	60.34
TTAC[4]	69.08	70.55	71.78	54.50	62.99	56.72	64.27	58.74
TIPI[25]	62.63	63.16	65.14	33.73	54.11	46.42	54.20	59.57
DDPM [26]	71.49	72.02	72.69	61.08	66.77	65.35	68.23	60.62
Ours	**71.65**	**72.23**	**72.88**	**63.14**	**67.05**	**65.97**	**68.82**	**62.60**

TABLE II: Accuracies(%) of six missing channel scenarios on CIFAR-10.

Method	Clean Data	Missing R	Missing G	Missing B	Missing R&G	Missing R&B	Missing G&B	Avg.
Test	94.55	89.85	87.53	91.38	86.15	87.87	87.97	88.46
TestBN		90.62	91.53	90.89	87.24	89.88	89.22	89.90
TENT[5]		91.31	92.20	91.53	88.71	90.49	89.72	90.66
SHOT[24]		91.77	92.47	91.90	89.23	91.39	90.88	91.27
TTAC[4]		93.14	**94.08**	93.28	90.28	**92.38**	91.39	92.43
TIPI[25]		88.61	90.43	89.53	80.90	84.24	84.29	86.33
DDPM [26]		92.11	93.10	92.35	88.84	91.49	90.84	91.46
Ours		**93.29**	93.44	**93.49**	**90.94**	91.96	**91.85**	**92.50**

TABLE III: Accuracies(%) of six missing channel scenarios on CIFAR-100.

Method	Clean Data	Missing R	Missing G	Missing B	Missing R&G	Missing R&B	Missing G&B	Avg.
Test	75.16	59.33	59.03	64.25	42.36	55.86	57.69	56.42
TestBN		64.27	67.16	66.49	55.73	55.76	60.40	61.64
TENT[5]		66.94	70.06	69.28	60.08	60.83	63.73	65.15
SHOT[24]		67.50	70.71	69.93	60.71	62.91	64.66	66.07
TTAC[4]		66.65	69.63	69.14	58.83	61.94	63.67	64.98
TIPI[25]		65.18	67.42	66.72	55.62	59.13	60.77	62.47
DDPM [26]		69.69	72.53	71.86	61.73	65.99	65.46	67.88
Ours		**70.25**	**72.63**	**72.27**	**62.83**	**66.99**	**66.26**	**68.54**

Computation Cost Measured in Wall-Clock Time: For the online stream test sample batch, all competing methods were allowed to construct a sample queue, with TTA training applied over multiple iterations. While these components add computational overhead, TTA prioritizes low inference latency. We measure the average per-sample wall-clock time, as shown in Tab. IV. Our method demonstrates a computational cost comparable to competing approaches, achieving superior performance with less wall-clock time than the previous state-of-the-art, TTAC. Compared to more efficient methods like TENT, our approach incurs 2–3 times higher cost but delivers a 2–4% accuracy gain. Additionally, our method shows a clear advantage over DDPM-based approaches.

C. Ablation Study

We evaluate the effectiveness of individual components on CIFAR-100 and ImageNet datasets in Tab. V, reporting the average accuracy across six types of missing channels. Simple imputation significantly improves robustness to missing features (56.42% $\rightarrow$ 64.74% on CIFAR-100 and 41.39% $\rightarrow$ 65.77% on ImageNet). Similarly, applying self-training with pseudo-labels generated through weak augmentation alone

TABLE IV: Impact of queue size on the accuracy (%) / per-sample wall-clock time (sec) on CIFAR-100 with missing G channel.

	w/o Queue	w/ Queue			
#Iterations	1	1	2	3	4
TENT	67.64/0.0013	67.64/0.0132	70.11/0.0321	69.91/0.0481	70.06/0.0483
SHOT	68.64/0.0026	70.30/0.0275	70.43/0.0734	70.53/0.1081	70.71/0.1203
TTAC	69.39/0.0032	68.94/0.1255	69.11/0.1730	69.39/0.1954	69.63/0.2232
DDPM	69.43/0.0524	71.87/0.2539	71.99/0.2846	72.32/0.3159	72.53/0.3471
Ours	71.89/0.0030	71.90/0.0812	72.05/0.1204	71.62/0.1416	72.63/0.1750
Inference	0.0003				

TABLE V: Ablation study on CIFAR-100 and ImageNet dataset.

ImputeST	Self Training	Clustering Regularization	Cifar100	ImageNet
-	-	-	56.42	41.39
✓	-	-	64.74	65.77
-	✓	-	65.36	66.14
✓	✓	-	66.10	66.63
✓	✓	✓	68.54	68.82

yields a notable performance boost (56.42% $\rightarrow$ 65.36% and 41.39% $\rightarrow$ 66.14%). Combining weak augmentation with data imputation further enhances the reliability of pseudo-labels, resulting in an additional 1% accuracy improvement. Finally, the introduction of clustering regularization achieves the highest performance, with accuracies of 68.54% on CIFAR-100 and 68.82% on ImageNet, surpassing all individual components.

Sensitivity to Hyperparameter: For high-dimensional datasets like ImageNet, a small learning rate ensures stable updates during test-time adaptation, preventing instability and overfitting. In contrast, for smaller datasets like CIFAR-100, a larger learning rate facilitates faster, more efficient adaptation. We conducted a sensitivity analysis of the learning rate across datasets, shown in Fig. 3 (left), which demonstrates the impact on model performance and highlights the trade-off between stability and adaptation speed. Fig.3 (right) presents a comparison of clustering loss weights λ. We ultimately selected a weight of 3, as this value strikes a balance between enabling the clustering loss to function effectively.

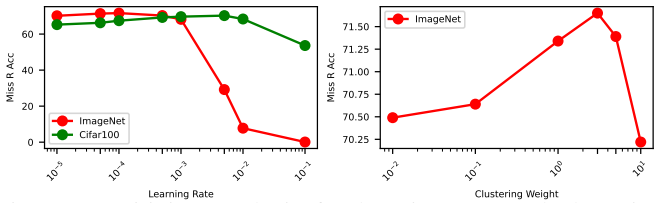

Fig. 3: Sensitivity Analysis for learning rate and clustering weight.

IV. Conclusion

In this work, we address a specific type of out-of-distribution (OOD) testing data where feature channels may be missing. To tackle this challenge, we first propose a lightweight statistical data imputation technique to reconstruct the missing channel features. Next, we employ self-training for real-time adaptation, combining predictions on test samples with weak augmentation and imputed data to generate more reliable pseudo-labels. Finally, we introduce a deep clustering regularization to enhance the self-training process. Evaluations on four RGB image classification datasets demonstrate the effectiveness of our method. Future research may consider more challenging inputs from high dimensional sensors, e.g. RGBD sensors and hyperspectral images.

REFERENCES

[1] Mei Wang and Weihong Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, 2018.

[2] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, 2015.

[3] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al., "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt," *arXiv preprint arXiv:2302.09419*, 2023.

[4] Yongyi Su, Xun Xu, and Kui Jia, "Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering," in *Advances in Neural Information Processing Systems*, 2022.

[5] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *International Conference on Learning Representations*, 2021.

[6] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *International conference on machine learning*, 2020.

[7] Kanjar De and Marius Pedersen, "Impact of colour on robustness of deep neural networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[8] Stewart Jamieson, Jonathan P. How, and Yogesh Girdhar, "Deepseecolor: Realtime adaptive color correction for autonomous underwater vehicles via deep learning methods," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[9] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, 2020.

[10] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *2020 International joint conference on neural networks (IJCNN)*, 2020.

[11] Yixin Wang, Yang Zhang, Yang Liu, Zihao Lin, Jiang Tian, Cheng Zhong, Zhongchao Shi, Jianping Fan, and Zhiqiang He, "Acn: Adversarial co-training network for brain tumor segmentation with missing modalities," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2021.

[12] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng, "Are multimodal transformers robust to missing modality?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[13] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim, "Towards good practices for missing modality robust action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[14] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee, "Multimodal prompting with missing modalities for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[15] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon, "Csdi: Conditional score-based diffusion models for probabilistic time series imputation," in *Advances in Neural Information Processing Systems*, 2021.

[16] Shuhan Zheng and Nontawat Charoenphakdee, "Diffusion models for missing value imputation in tabular data," in *NeurIPS Table Representation Learning (TRL) Workshop*, 2022.

[17] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, 2017.

[18] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020.

[19] Hui Tang, Ke Chen, and Kui Jia, "Unsupervised domain adaptation via structurally regularized deep clustering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[20] Kamran Ghasedi Dizaji, Amirhossein Herandi, Cheng Deng, Weidong Cai, and Heng Huang, "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[21] Cameron Fabbri, Md Jahidul Islam, and Junaed Sattar, "Enhancing underwater imagery using generative adversarial networks," in *2018 IEEE international conference on robotics and automation (ICRA)*, 2018.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[23] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi, "Ttt++: When does self-supervised test-time training fail or thrive?," *Advances in Neural Information Processing Systems*, 2021.

[24] Jian Liang, Dapeng Hu, and Jiashi Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *International conference on machine learning*, 2020.

[25] A. Tuan Nguyen, Thanh Nguyen-Tang, Ser-Nam Lim, and Philip H.S. Torr, "Tipi: Test time adaptation with transformation invariance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[26] Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang, "Back to the source: Diffusion-driven adaptation to test-time corruption," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.