Fourier Test-Time Adaptation with Multi-level Consistency for Robust Classification

Yuhao Huang[1,2,3], Xin Yang[1,2,3], Xiaoqiong Huang[1,2,3], Xinrui Zhou[1,2,3], Haozhe Chi[4], Haoran Dou[5], Xindi Hu[6], Jian Wang[7], Xuedong Deng[8], and Dong Ni[1,2,3(✉)]

[1] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China
nidong@szu.edu.cn
[2] Medical Ultrasound Image Computing (MUSIC) Lab, Shenzhen University, Shenzhen, China
[3] Marshall Laboratory of Biomedical Engineering, Shenzhen University, Shenzhen, China
[4] ZJU-UIUC Institute, Zhejiang University, Hangzhou, China
[5] Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), University of Leeds, Leeds, UK
[6] Shenzhen RayShape Medical Technology Co. Ltd., Shenzhen, China
[7] School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China
[8] The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou, China

Abstract. Deep classifiers may encounter significant performance degradation when processing unseen testing data from varying centers, vendors, and protocols. Ensuring the robustness of deep models against these domain shifts is crucial for their widespread clinical application. In this study, we propose a novel approach called Fourier Test-time Adaptation (FTTA), which employs a dual-adaptation design to integrate input and model tuning, thereby jointly improving the model robustness. The main idea of FTTA is to build a reliable multi-level consistency measurement of paired inputs for achieving self-correction of prediction. Our contribution is two-fold. First, we encourage consistency in global features and local attention maps between the two transformed images of the same input. Here, the transformation refers to *Fourier*-based input adaptation, which can transfer one unseen image into source style to reduce the domain gap. Furthermore, we leverage style-interpolated images to enhance the global and local features with learnable parameters, which can smooth the consistency measurement and accelerate convergence.

Y. Huang and X. Yang—Contribute equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43898-1_22.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 221–231, 2023.
https://doi.org/10.1007/978-3-031-43898-1_22

Second, we introduce a regularization technique that utilizes style interpolation consistency in the frequency space to encourage self-consistency in the logit space of the model output. This regularization provides strong self-supervised signals for robustness enhancement. FTTA was extensively validated on three large classification datasets with different modalities and organs. Experimental results show that FTTA is general and outperforms other strong state-of-the-art methods.

Keywords: Classifier robustness · Testing-time adaptation · Consistency

1 Introduction

Domain shift (see Fig. 1) may cause deep classifiers to struggle in making plausible predictions during testing [15]. This risk seriously limits the reliable deployment of these deep models in real-world scenarios, especially for clinical analysis. Collecting data from the target domain to retrain from scratch or fine-tune the trained model is the potential solution to handle the domain shift risks. However, obtaining adequate testing images with manual annotations is laborious and impracticable in clinical practice. Thus, different solutions have been proposed to conquer the problem and improve the model robustness.

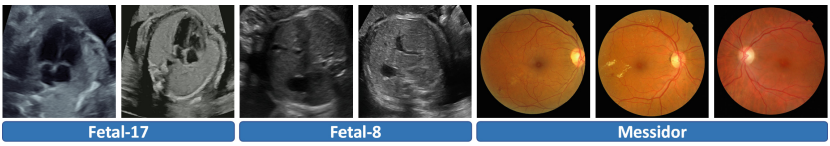

Fig. 1. From left to right: 1) four-chamber views of heart from Vendor A&B, 2) abdomen planes from Vendor C&D, 3) fundus images with diabetic retinopathy of grade 3 from Center E-G. Appearance and distribution differences can be seen in each group.

Unsupervised Domain Adaptation (UDA) refers to training the model with labeled source data and adapting it with target data without annotation [6,18, 22]. Recently, *Fourier* domain adaptation was proposed in [25,26], with the core idea of achieving domain transfer by replacing the low-frequency spectrum of source data with that of the target one. Although effective, they require obtaining sufficient target data in advance, which is challenging for clinical practice.

Domain Generalization (DG) aims to generalize models to the unseen domain not presented during training. Adversarial learning-based DG is one of the most popular choices that require multi-domain information for learning domain-invariant representations [11,12]. Recently, Liu et al. [14] proposed to construct a continuous frequency space to enhance the connection between different domains. Atwany et al. [1] imposed a regularization to reduce gradient variance from different domains for diabetic retinopathy classification. One drawback is that they require multiple types of source data for extracting rich

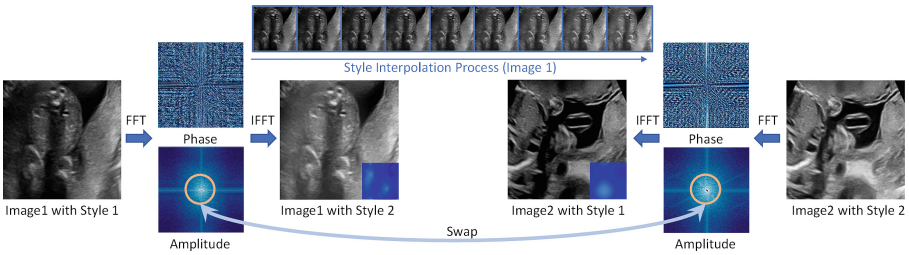

Fig. 2. Illustration of the amplitude swapping between two images with different styles. Pseudo-color images shown in the right-down corner indicate the differences between images before and after amplitude swapping.

features. Other alternatives proposed using only one source domain to perform DG [4,27]. However, they still heavily rely on simulating new domains via various data augmentations, which can be challenging to control.

Test-Time Adaptation (TTA) adapts the target data or pre-trained models during testing [8,9]. Test-time Training (TTT) [21] and TTT++ [15] proposed to minimize a self-supervised auxiliary loss. Wang et al. [23] proposed the TENT framework that focused on minimizing the entropy of its predictions by modulating features via normalization statistics and transformation parameters estimation. Instead of batch input like the above-mentioned methods, Single Image TTA (SITA) [10] was proposed with the definition that having access to only one given test image once. Recently, different mechanisms were developed to optimize the TTA including distribution calibration [16], dynamic learning rate [24], and normalizing flow [17]. Most recently, Gao et al. [5] proposed projecting the test image back to the source via the source-trained diffusion models. Although effective, these methods often suffer from the problems of unstable parameter estimation, inaccurate proxy tasks/pseudo labels, difficult training, etc. Thus, a simple yet flexible approach is highly desired to fully mine and combine information from test data for online adaptation.

In this study, we propose a novel framework called Fourier TTA (FTTA) to enhance the model robustness. We believe that this is the first exploration of dual-adaptation design in TTA that jointly updates input and model for online refinement. Here, one assumption is that a well-adapted model will get consistent outputs for different transformations of the same image. Our contribution is two-fold. First, we align the high-level features and attention regions of transformed paired images for complementary consistency at global and local dimensions. We adopt the *Fourier*-based input adaptation as the transformation strategy, which can reduce the distances between unseen testing images and the source domain, thus facilitating the model learning. We further propose to smooth the hard consistency via the weighted integration of features, thus reducing the adaptation difficulties of the model. Second, we employ self-consistency of frequency-based style interpolation to regularize the output logits. It can provide direct and effective hints to improve model robustness. Validated on three classification datasets, we demonstrate that FTTA is general in improving classification robustness, and achieves state-of-the-art results compared to other strong TTA methods.

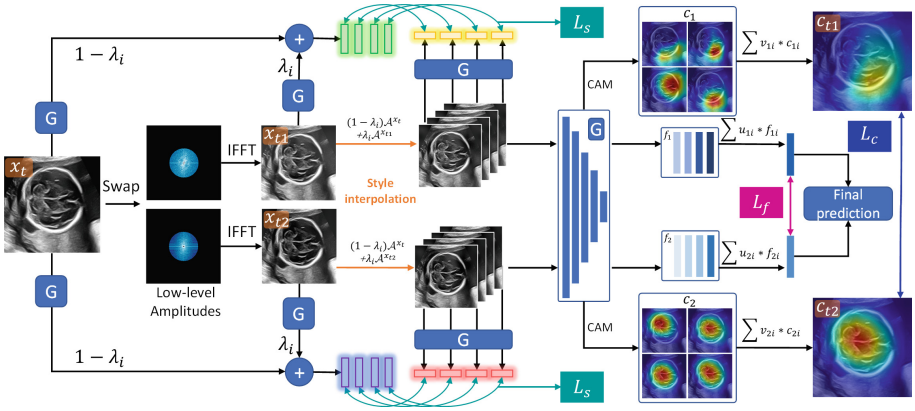

Fig. 3. Pipeline of our proposed FTTA framework.

2 Methodology

Figure 3 shows the pipeline of FTTA. Given a trained classifier G, FTTA first conducts *Fourier*-based input adaptation to transfer each unseen testing image x_t into two source-like images (x_{t1} and x_{t2}). Then, using linear style interpolation, two groups of images will be obtained for subsequent smooth consistency measurement at global features (L_f) and local visual attention (L_c). Furthermore, regularization in the logit space can be computed following the style interpolation consistency in the frequency space (L_s). Finally, FTTA updates once based on the multi-consistency losses to output the final average prediction.

Fourier-Based Input Adaptation for Domain Transfer. Transferring unseen images to the known domain plays an important role in handling domain shift risks. In this study, instead of learning on multiple domains, we only have access to one single domain of data during training. Therefore, we need to utilize the limited information and find an effective way to realize the fast transfer from the unseen domain to the source domain. Inspired by [20,25,26], we adopt the Fast Fourier Transform (FFT) based strategy to transfer the domain information and achieve input adaptation during testing. Specifically, we transfer the domain information from one image to another by low-frequency amplitude ($\mathcal{A}$) swapping while keeping the phase components (see Fig. 2). This is because in *Fourier* space, the low-frequency $\mathcal{A}$ encodes the style information, and semantic contents are preserved in $\mathcal{P}$ [25]. Domain transfer via amplitude swapping between image x_s to x_t can be defined as:

$$\mathcal{A}^{x_{t'}} = ((1-\lambda)\mathcal{A}^{x_t} + \lambda\mathcal{A}^{x_s}) \circ \mathcal{M} + \mathcal{A}^{x_t} \circ (1-\mathcal{M}), \tag{1}$$

where $\mathcal{M}$ is the circular low-pass filtering with radius r to obtain the radial-symmetrical amplitude [26]. λ aims to control the degree of style interpolation [14], and it can make the transfer process continues (see Fig. 2). After inverse FFT (IFFT, $\mathcal{F}^{-1}$), we can obtain an image x'_t by $\mathcal{F}^{-1}(\mathcal{A}^{x'_t}, \mathcal{P}^{x_t})$.

Since one low-level amplitude represents one style, we have n style choices. n is the number of training data. The chosen styles for input adaptation should be representative of the source domain while having significant differences from each other. Hence, we use the validation set to select the styles by first turning the whole validation data into the n styles and calculating n accuracy. Then, styles for achieving *top-k* performance are considered representative, and L2 distances between the C_K^2 pairs are computed to reflect the differences.

Smooth Consistency for Global and Local Constraints. Building a reliable consistency measurement of paired inputs is the key to achieving TTA. In this study, we propose global and local alignments to provide a comprehensive consistency signal for tuning the model toward robustness. For *global consistency*, we compare the similarity between high-level features of paired inputs. These features encode rich semantic information and are therefore well-suited for assessing global consistency. Specifically, we utilize hard and soft feature alignments via pixel-level L2 loss and distribution-level cosine similarity loss, to accurately compute the global feature loss L_f. To ensure *local consistency*, we compute the distances between the classification activation maps (CAMs) of the paired inputs. It is because CAMs (e.g., Grad-CAM [19]) can reflect the local region the model focuses on when making predictions. Forcing CAMs of paired inputs to be close can guide the model to optimize the attention maps and predict using the correct local region for refining the prediction and improving model robustness (see Fig. 3, c_{t1} is encouraged to be closer with c_{t2} for local visual consistency). Finally, the distances between two CAMs can be computed by the combination of L2 and JS-divergence losses.

Despite global and local consistency using single paired images can provide effective self-supervised signals for TTA in most cases, they may be difficult or even fail in aligning the features with a serious gap during testing. This is because the representation ability of single-paired images is limited, and the hard consistency between them may cause learning and convergence difficulties. For example, the left-upper CAMs of $c1$ and $c2$ in Fig. 3 are with no overlap. Measuring the local consistency between them is meaningless since JS divergence will always output a constant in that case. Thus, we first generate two groups of images, each with four samples, by style interpolation using different λ. Then, we fed them into the model for obtaining two groups of features. Last, we propose learnable integration with parameters u and v to linearly integrate the global and local features. This can enhance the feature representation ability, thus smoothing the consistency evaluation to accelerate the adaptation convergence.

Style Consistency for Regularization on Logit Space. As described in the first half of Eq. 1, two low-level amplitudes (i.e., styles) can be linearly combined into a new one. We propose to use this frequency-based style consistency to regularize the model outputs in logit space, which is defined as the layer before *softmax*. Thus, it is directly related to the model prediction. A total of 8 logit

Table 1. Datasets split of each experimental group.

	Groups	Training	Validation	Testing
Fetal-17	A2B	2622	1135	4970
	B2A	3472	1498	3757
Fetal-8	C2D	3551	1529	5770
	D2C	4035	1735	5080
Messidor	E2FG	279	121	800
	F2EG	278	122	800
	G2EF	279	121	800

pairs can be obtained (see Fig. 3), and the loss can be defined as:

$$L_s = \left(\sum_{i=1}^{2} \sum_{j=1}^{4} ||(1-\lambda_j) * y_{log}(x_t) + \lambda_j * y_{log}(x_{ti}) - y_{log}(x_{ij})||_2 \right) /8, \quad (2)$$

where x_t and $x_{ti,i\in 1,2}$ are the testing image and two transformed images after input adaptation. x_{ij} represents style-interpolated images controlled by λ_j. $y_{log}(\cdot)$ outputs the logits of the model.

3 Experimental Results

Materials and Implementations. We validated the FTTA framework on three classification tasks, including one private dataset and two public datasets (see Fig. 1). Approved by the local IRB, the in-house *Fetal-17* US dataset containing 8727 standard planes with gestational age (GA) ranging from 20 to 24^{+6} weeks was collected. It contains 17 categories of planes with different parts, including limbs (4), heart (4), brain (3), abdomen (3), face (2), and spine (1). Four 10-year experienced sonographers annotated one classification tag for each image using the Pair annotation software package [13]. *Fetal-17* consists of two vendors (A&B) and we conducted bidirectional experiments (A2B and B2A) for method evaluation. The Maternal-fetal US dataset named *Fetal-8* (GA: 18–40 weeks) [2][1] contains 8 types of anatomical planes including brain (3), abdomen (1), femur (1), thorax (1), maternal cervix (1), and others (1). Specifically, 10850 images from vendors ALOKA and Voluson (C&D) were used for bidirectional validation (C2D and D2C). Another public dataset is a fundus dataset named *Messidor*, which contains 1200 images from 0–3 stage of diabetic retinopathy [3][2]. It was collected from three ophthalmologic centers (E, F&G) with each of them can treated as a source domain, allowing us to conduct three groups of experiments (E2FG, F2EG and G2EF). Dataset split information is listed in Table 1.

[1] https://zenodo.org/record/3904280#.YqIQvKhBy3A.
[2] https://www.adcis.net/en/third-party/messidor/.

Table 2. Comparisons on Fetal-17 dataset. The best results are shown in bold.

Methods	Fetal-17: A2B				Fetal-17: B2A			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Upper-bound	96.33	95.79	94.54	94.61	91.81	89.28	88.89	88.67
Baseline [7]	61.25	64.57	59.46	57.83	63.51	60.60	61.60	57.50
TTT [21]	71.91	65.62	67.33	63.28	72.77	57.34	58.11	55.51
TTT++ [15]	78.65	79.08	75.21	73.20	78.60	76.15	71.36	71.02
TENT [23]	76.48	74.54	71.41	69.40	75.83	73.38	70.75	67.01
DLTTA [24]	85.51	88.30	83.87	83.83	83.20	81.39	76.77	76.94
DTTA [5]	87.00	86.93	84.48	83.87	83.39	82.17	78.73	79.09
TTTFlow [17]	86.66	85.98	85.38	84.96	84.08	**85.41**	76.99	75.70
FTTA-IA	73.56	72.46	67.49	61.47	69.07	64.17	52.81	50.64
FTTA-C1	82.27	81.66	76.72	75.32	81.55	78.56	68.43	67.17
FTTA-C2	83.06	82.42	81.81	79.92	81.95	69.03	67.51	65.78
FTTA-C3	84.77	82.14	77.04	76.30	82.09	78.57	80.24	77.15
FTTA-C*	80.70	82.52	77.24	78.18	79.24	80.64	74.42	74.00
FTTA-C	88.93	89.43	82.40	82.96	84.91	80.91	75.64	76.52
Ours	**91.02**	**89.62**	**89.74**	**89.37**	**87.41**	82.46	**81.91**	**81.15**

Table 3. Comparisons on Fetal-8 and MESSDIOR datasets.

Methods	C2D		D2C		E2FG		F2EG		G2EF	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Upper-bound	87.49	85.14	94.05	85.24	60.91	53.46	66.26	57.63	62.14	53.34
Baseline	67.68	65.93	79.92	72.09	47.62	37.09	31.13	31.30	41.25	39.41
Ours	82.13	76.89	91.87	73.19	59.26	43.98	57.43	37.26	58.02	45.46

We implemented FTTA in Pytorch, using an NVIDIA A40 GPU. All images were resized to 256×256, and normalized before input to the model. For the fetal datasets, we used a 1-channel input, whereas, for the fundus dataset, 3-channel input was utilized. During training, we augmented the data using common strategies including rotation, flipping and contrast transformation. We selected ImageNet-pretrained *ResNet-18* [7] as our classifier backbone and optimized it using the AdamW optimizer in 100 epochs. For offline training, with batch size = 196, the learning rate (lr) is initialized to 1e−3 and multiplied by 0.1 per 30 epochs. Cross-entropy loss is the basic loss for training. We selected models with the best performance on validation sets to work with FTTA. For online testing, we set the lr equal to 5e−3, and $\lambda_{j,j=1,2,3,4}$ for style interpolation was set as 0.2, 0.4, 0.6, and 0.8, respectively. We only updated the network parameters and learnable weights once based on the multi-level consistency losses function before obtaining the final predictions.

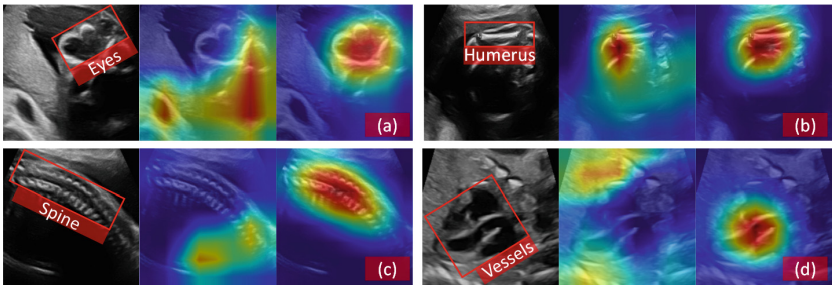

Fig. 4. Four typical cases in Fetal-17 dataset: (a) Axial orbit and lenses, (b) Humerus plane, (c) Sagittal plane of the spine, and (d) Left ventricular outflow tract view. (Color figure online)

Quantitative and Qualitative Analysis. We evaluated the classification performance using four metrics including Accuracy (*Acc*, %), Precision (*Pre*, %), Recall (*Rec*, %), and F1-score (*F1*, %). Table 2 compares the FTTA (*Ours*) with seven competitors including the *Baseline* without any adaptation and six state-of-the-art TTA methods. *Upper-bound* represents the performance when training and testing on the target domain. It can be seen from *Upper-bound* and *Baseline* that all the metrics have serious drops due to the domain shift. *Ours* achieves significant improvements on *Baseline*, and outperforms all the strong competitors in terms of all the evaluation metrics, except for the *Pre* in Group B2A. It is also noted that the results of *Ours* are approaching the *Upper-bound*, with only 5.31% and 4.40% gaps in *Acc*.

We also perform ablation studies on the *Fetal-17* dataset in the last 7 rows of Table 2. *FTTA-IA* denotes that without model updating, only input adaptation is conducted. Four experiments are performed to analyze the contribution of three consistency measurements (*–C1*, *–C2*, and *–C3* for global features, local CAM, and style regularization, respectively), and also the combination of them (*–C*). They are all equipped with the input adaptation for fair comparisons. *FTTA-C** indicates replacing the Fourier-based input adaptation with 90° rotation to augment the test image for consistency evaluation. Different from *FTTA-C*, *Ours* integrates learnable weight groups to smooth consistency measurement. Experiences show that the naive Fourier input adaptation in *FTTA-IA* can boost the performance of *Baseline*. The three consistency variants improve the classification performance respectively, and combining them together can further enhance the model robustness. Then, the comparison between *FTTA-C* and *Ours* validates the effectiveness of the consistency smooth strategy.

Table 3 reports the results of FTTA on two public datasets. We only perform methods including *Upper-bound*, *Baseline*, and *Ours* with evaluation metrics *Acc* and *F1*. Huge domain gaps can be observed by comparing *Upper-bound* and *Baseline*. All five experimental groups prove that our proposed FTTA can boost the classification performance over *baseline*, and significantly narrow the gaps between *upper-bound*. Note that *MESSDIOR* is a challenging dataset, with all

the groups having low Upper-bounds. Even for the multi-source DG method, *Messidor* only achieves 66.70% accuracy [1]. For the worst group (F2EG), *Acc* drops 35.13% in the testing sets. However, the proposed FTTA can perform a good adaptation and improve 26.30% and 5.96% in *Acc* and *F1*.

Figure 4 shows the CAM results obtained by *Ours*. The red boxes denote the key regions, like the *eyes* in (a), which were annotated by sonographers and indicate the region-of-interest (ROI) with discriminant information. We consider that if one model can focus on the region having a high overlap with the ROI box, it has a high possibility to be predicted correctly. The second columns visualize the misclassified results before adaptation. It can be observed via the CAMs that the focus of the model is inaccurate. Specifically, they spread dispersed on the whole image, overlap little with the ROI, or with low prediction confidence. After TTA, the CAMs can be refined and close to the ROI, with prediction corrected.

4 Conclusion

In this study, we proposed a novel and general FTTA framework to improve classification robustness. Based on Fourier-based input adaptation, FTTA is driven by the proposed multi-level consistency, including smooth global and local constraints, and also the self-consistency on logit space. Extensive experiments on three large datasets validate that FTTA is effective and efficient, achieving state-of-the-art results over strong TTA competitors. In the future, we will extend the FTTA to segmentation or object detection tasks.

Acknowledgement. This work was supported by the grant from National Natural Science Foundation of China (Nos. 62171290, 62101343), Shenzhen-Hong Kong Joint Research Program (No. SGDX20201103095613036), and Shenzhen Science and Technology Innovations Committee (No. 20200812143441001).

References

1. Atwany, M., Yaqub, M.: DRGen: domain generalization in diabetic retinopathy classification. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13432, pp. 635–644. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_61
2. Burgos-Artizzu, X.P., et al.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Sci. Rep. **10**(1), 1–12 (2020)
3. Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., et al.: Feedback on a publicly distributed image database: the messidor database. Image Anal. Stereol. **33**(3), 231–234 (2014)
4. Fan, X., Wang, Q., Ke, J., Yang, F., Gong, B., Zhou, M.: Adversarially adaptive normalization for single domain generalization. In: Proceedings of the IEEE/CVF CVPR, pp. 8208–8217 (2021)
5. Gao, J., Zhang, J., Liu, X., Darrell, T., Shelhamer, E., Wang, D.: Back to the source: diffusion-driven adaptation to test-time corruption. In: Proceedings of the IEEE/CVF CVPR, pp. 11786–11796 (2023)

6. Geng, B., Tao, D., Xu, C.: DAML: domain adaptation metric learning. IEEE Trans. Image Process. **20**(10), 2980–2989 (2011)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE CVPR, pp. 770–778 (2016)
8. Huang, X., et al.: Test-time bi-directional adaptation between image and model for robust segmentation. Comput. Methods Programs Biomed. **233**, 107477 (2023)
9. Huang, Y., et al.: Online reflective learning for robust medical image segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13438, pp. 652–662. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1_62
10. Khurana, A., Paul, S., Rai, P., Biswas, S., Aggarwal, G.: SITA: single image test-time adaptation. arXiv preprint arXiv:2112.02355 (2021)
11. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: Proceedings of the IEEE CVPR, pp. 5400–5409 (2018)
12. Li, Y., et al.: Deep domain generalization via conditional invariant adversarial networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11219, pp. 647–663. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01267-0_38
13. Liang, J., et al.: Sketch guided and progressive growing GAN for realistic and editable ultrasound image synthesis. Med. Image Anal. **79**, 102461 (2022)
14. Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.A.: FedDG: federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: Proceedings of the IEEE/CVF CVPR, pp. 1013–1023 (2021)
15. Liu, Y., Kothari, P., Van Delft, B., Bellot-Gurlet, B., Mordan, T., Alahi, A.: TTT++: when does self-supervised test-time training fail or thrive? In: Advances in Neural Information Processing Systems, vol. 34, pp. 21808–21820 (2021)
16. Ma, W., Chen, C., Zheng, S., Qin, J., Zhang, H., Dou, Q.: Test-time adaptation with calibration of medical image classification nets for label distribution shift. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13433, pp. 313–323. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_30
17. Osowiechi, D., Hakim, G.A.V., Noori, M., Cheraghalikhani, M., Ben Ayed, I., Desrosiers, C.: TTTFlow: unsupervised test-time training with normalizing flow. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2126–2134 (2023)
18. Ren, J., Hacihaliloglu, I., Singer, E.A., Foran, D.J., Qi, X.: Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 201–209. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_23
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE ICCV, pp. 618–626 (2017)
20. Sharifzadeh, M., Tehrani, A.K., Benali, H., Rivaz, H.: Ultrasound domain adaptation using frequency domain analysis. In: 2021 IEEE IUS, pp. 1–4. IEEE (2021)
21. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International Conference on Machine Learning, pp. 9229–9248. PMLR (2020)
22. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE CVPR, pp. 7167–7176 (2017)

23. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: TENT: fully test-time adaptation by entropy minimization. In: ICLR (2021)
24. Yang, H., et al.: DLTTA: dynamic learning rate for test-time adaptation on cross-domain medical images. IEEE Trans. Med. Imaging **41**(12), 3575–3586 (2022)
25. Yang, Y., Soatto, S.: FDA: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF CVPR, pp. 4085–4095 (2020)
26. Zakazov, I., Shaposhnikov, V., Bespalov, I., Dylov, D.V.: Feather-light Fourier domain adaptation in magnetic resonance imaging. In: Kamnitsas, K., et al. (eds.) DART 2022. LNCS, vol. 13542, pp. 88–97. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16852-9_9
27. Zhao, L., Liu, T., Peng, X., Metaxas, D.: Maximum-entropy adversarial data augmentation for improved generalization and robustness. In: Advances in Neural Information Processing Systems, vol. 33, pp. 14435–14447 (2020)