Everything to the Synthetic: Diffusion-driven Test-time Adaptation via Synthetic-Domain Alignment

Supplementary Material

A. Implementation Details

A.1. Baselines.

We choose DDA [17] as our primary competitor since it is the best-performing publicly available diffusion-driven TTA method. Same as DDA, we include DiffPure [47] and MEMO [72] as baselines. We also compare SDA against the recent SOTA GDA [66] using their paper results. For data stream sensitivity comparison, we compare SDA with 10 additional traditional TTA methods, including TENT [67], ROID [39], NOTE [18], CoTTA [69], TRIBE [61], BN [41], UniMIX [65], RoTTA [71], LAME [4] and UniTTA [12]. The results are evaluated across various TTA benchmarks, including ImageNet-C [24], ImageNet-W [33], CIFAR-10-C [24] and PASCAL VOC-C [13].

A.2. Settings.

All experiments are conducted with 8 A100 GPUs. For ImageNet variants, we explore ResNet [23], ConvNeXt [37], and Swin [36] as source models. DiT [49] and ADM [10] are adopted as conditional and unconditional diffusion models, respectively. For CIFAR-10-C [24], we use ResNet as the source model. EDM [28] and I-DDPM [46] are adopted as conditional and unconditional diffusion models, respectively. For PASCAL VOC-C [13], we use DeepLabv3 [6] as the source segmenter. Dataset Diffusion [42] and FLUX schnell [30] are adopted as conditional and unconditional diffusion models, respectively. For classification tasks via MLLMs, we use LLaVA 1.5-7b [35] as the source model. For each task, we generate 50K images with balanced class labels. For different source models and target domains, the synthetic data only needs to be generated once. The detailed fine-tuning settings of classifiers and segmenters are summarized in Tab. 17. For MLLM (LLaVA) fine-tuning, we follow the default configurations in [35]. Fig. 6 shows the task format for fine-tuning and evaluating MLLMs.

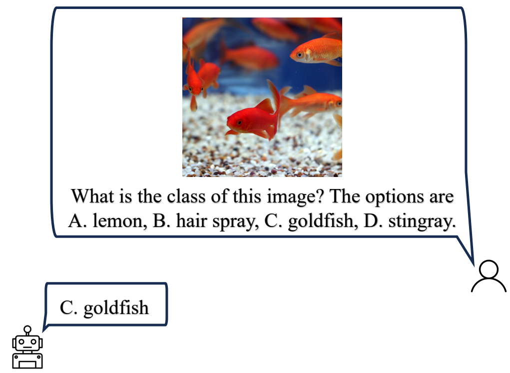

Figure 6. Task format for fine-tuning and evaluating MLLMs. Given an image, we ask an MLLM to choose the correct image class from four provided options.

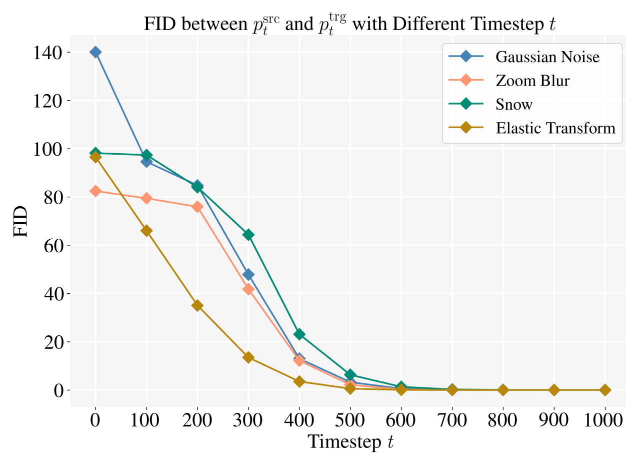

Figure 7. Fréchet Inception Distance (FID) [25] between p_t^{src} and p_t^{trg} with different timestep t. We conduct experiments on four typical adaptation types from ImageNet-C.

B. Selection of Timestep for TTA

As aforementioned in Eq. 4, the success of diffusion-driven data adaptation relies on the selection of a suitable minimum t^* that satisfies $p_{t^*}^{\text{src}} \approx p_{t^*}^{\text{trg}}$. In Fig. 7, we leverage FID [25] to measure the domain divergence of p_t^{src} and p_t^{trg} with different timestep t. The results indicate that for a 1000-step diffusion scheduler and adaptation tasks from the standard benchmark ImageNet-C [24], diffusion-driven data adaptation typically requires a t^* larger than 500. We empirically demonstrate that applying such t^* to diffusion-driven TTA methods leads to significant misalignment between the source and synthetic domains, as shown in Tab. 2. In our experiments, we set the same $t^* = 500$ as our baseline DDA [17]. Here $t^* = 500$ refers to using half sampling steps as the whole diffusion scheduler, *e.g.*, for a 100-step scheduler, the actual sampling step for adaptation is 50.

C. Additional Results

C.1. Conditional and Unconditional Synthetic Domain Misalignment on Other Datasets

Beyond Tab. 10, new results on CIFAR-10-C and PASCAL VOC-C in Tab. 12 indicate the conditional and unconditional synthetic domain misalignment issue may be significant for almost all datasets. SDA mitigates misalignment and therefore improves performance in all settings.

SDA Component Domain Adaptation Direction	CIFAR-10-C	PASCAL VOC-C
DDA (Target to Uncond. Syn)	65.3	38.6
+ Cond. Generation (Source to Cond. Syn)	69.8 (+4.5)	38.8 (+0.2)
+ Uncond. Alignment (Cond. Syn to Uncond. Syn), SDA	**72.4** (+7.1)	**39.8** (+1.2)

Table 12. SDA mitigates synthetic domain misalignment and improves performance across different datasets.

C.2. Diffusion Model Selection

Besides DiT used in the main paper, we explore more diffusion models pretrained on source data (SiT) and web data (Stable Diffusion XL) with ConvNeXt-B to indicate the model insensitivity of SDA.

DDA	SDA (DiT)	SDA (SiT)	SDA (Stable Diffusion XL)
49.4	**51.9** (+2.5)	**51.5** (+2.1)	**51.9** (+2.5)

Table 13. ImageNet-C accuracy with different diffusion models.

C.3. Generalization to Other Diffusion-driven Data Adaptation Methods

We additionally integrate SDA with DiffPure in Tab. 14. The results show consistent improvements.

	ConvNeXt-B	Swin-B
DiffPure	32.7	28.9
SDA (with DiffPure)	**46.3** (+13.6)	**46.3** (+11.7)

Table 14. ImageNet-C accuracy using SDA with DiffPure.

C.4. SDA Performance on Clean ImageNet

SDA does not notably downgrade source model performance on the original clean data, as shown in Tab. 15.

	ConvNeXt-B	Swin-B
Source Model	83.7	83.4
SDA (Ours)	83.6 (-0.1)	83.2 (-0.2)

Table 15. SDA accuracy on ImageNet validation set.

C.5. Time Cost

Analysis on time cost for conditional data generation, unconditional data alignment, and fine-tuning are listed in Tab. 16. Two settings are considered: (1) Fine-tuning on 50K images, which consumes more time but achieves better performance, and (2) Fine-tuning on 1K images, which reduces time cost by 50× and maintains comparable performance (also see Tab. 11 in our paper). Note that generation and alignment don't need to be redone for different models.

Settings	Generation	Alignment	Fine-tuning	Accuracy
50K images	$\sim$ 3 hours	$\sim$ 6 hours	$\sim$ 3 minutes	32.5
1K images	$\sim$ 4 minutes	$\sim$ 8 minutes	$<$ 10 seconds	31.9

Table 16. Time cost with 8 A100 GPUs with ResNet-50.

C.6. Detailed Comparisons

We provide detailed comparisons of SDA and baselines across 15 adaptation domains of ImageNet-C in Tabs. 18 to 21 and across 12 class/domain balance/imbalance settings from the UniTTA benchmark [12] in Tab. 22.

Dataset	ImageNet		CIFAR-10	PASCAL VOC
Model	ResNet-50	Swin-T/B & ConvNeXt-T/B	ResNet-18	DeepLabv3
optimizer	SGD	AdamW	SGD	SGD
base learning rate	5e-4	2e-5	5e-2	1e-4
weight decay	1e-4	1e-8	1e-4	5e-6
optimizer momentum	0.9	$\beta_1, \beta_2 = 0.9, 0.999$	0.9	0.9
batch size	512	1024	128	32
training epochs	15	15	15	2500 (iterations)
learning rate schedule	step decay at epoch 10	cosine decay	step decay at epoch 10	polynomialLR
warmup epochs	None	5	None	None
warmup schedule	N/A	linear	N/A	N/A
conditional diffusion model	DiT-XL/2	DiT-XL/2	EDM-VP	Dataset Diffusion
conditional sampling steps	250	250	512	100
classifier-free guidance	1.0	1.0	1.0	7.5
unconditional diffusion model	ADM	ADM	I-DDPM	FLUX schnell
unconditional sampling steps	50	50	50	25

Table 17. Synthetic-domain model adaptation settings.

	Gaussian	Shot	Impluse	Defocus	Glass	Motion	Zoom	Frost	Snow	Fog	Brightness	Contrast	Elastic	Pixelate	JEPG	Avg.
Source	39.1	37.7	38.8	29.0	11.1	33.4	34.7	51.1	43.4	59.8	71.3	41.2	27.1	35.9	54.0	40.5
DDA	53.8	49.2	50.3	28.5	26.2	33.4	34.9	49.4	42.8	40.9	67.9	38.0	43.1	52.7	57.1	44.5
SDA (Ours)	55.3	53.5	53.7	32.5	31.1	37.7	38.3	51.1	43.8	42.4	69.7	34.4	47.8	58.3	60.8	**47.4** **(+2.9)**

Table 18. Comparisons of SDA and baselines across 15 adaptation domains of ImageNet-C. Results are conducted with Swin-B.

	Gaussian	Shot	Impluse	Defocus	Glass	Motion	Zoom	Frost	Snow	Fog	Brightness	Contrast	Elastic	Pixelate	JEPG	Avg.
Source	40.1	39.1	38.7	25.6	11.4	33.0	31.2	49.3	43.8	41.9	70.3	45.0	22.5	41.0	57.2	39.3
DDA	55.6	51.6	51.3	24.7	26.9	31.9	32.3	48.4	42.6	34.3	66.7	39.9	42.2	54.6	59.3	44.2
SDA (Ours)	56.7	53.9	53.8	29.9	32.0	36.2	36.8	49.7	43.7	36.4	68.0	39.0	47.1	59.8	62.1	**47.0** **(+2.8)**

Table 19. Comparisons of SDA and baselines across 15 adaptation domains of ImageNet-C. Results are conducted with ConNeXt-T.

	Gaussian	Shot	Impluse	Defocus	Glass	Motion	Zoom	Frost	Snow	Fog	Brightness	Contrast	Elastic	Pixelate	JEPG	Avg.
Source	29.9	28.2	28.3	23.1	9.5	24.4	27.8	46.6	36.3	47.0	68.4	34.5	20.8	27.4	50.1	33.5
DDA	51.4	46.6	46.3	21.0	22.1	23.9	27.9	45.5	36.2	40.5	64.3	30.6	40.4	48.5	54.2	40.0
SDA (Ours)	52.4	50.2	50.1	24.4	26.5	29.1	32.4	46.2	37.3	38.8	65.3	26.1	46.1	55.3	57.2	**42.5** **(+2.5)**

Table 20. Comparisons of SDA and baselines across 15 adaptation domains of ImageNet-C. Results are conducted with Swin-T.

	Gaussian	Shot	Impluse	Defocus	Glass	Motion	Zoom	Frost	Snow	Fog	Brightness	Contrast	Elastic	Pixelate	JEPG	Avg.
Source	6.1	7.5	6.7	14.3	7.6	11.8	21.5	21.4	16.2	19.1	55.1	3.6	14.5	33.3	42.1	18.7
DDA	46.9	42.0	41.3	13.8	16.4	12.0	22.3	26.8	21.0	17.1	51.1	3.1	36.2	45.7	50.2	29.7
SDA (Ours)	43.4	43.2	42.5	18.8	21.6	16.6	27.4	30.0	22.6	18.1	53.1	3.1	41.0	52.1	53.4	**32.5** **(+2.8)**

Table 21. Comparisons of SDA and baselines across 15 adaptation domains of ImageNet-C. Results are conducted with ResNet-50.

Class setting	i.i.d. and balanced (i,1)		non-i.i.d. and balanced (n,1)					non-i.i.d. and imbalanced (n,u)					
Domain setting	(1,1)	(i,1)	(1,1)	(i,1)	(i,u)	(n,1)	(n,u)	(1,1)	(i,1)	(i,u)	(n,1)	(n,u)	
Corresponding setting	CoTTA	ROID	RoTTA	-	-	-	-	TRIBE	-	-	-	-	Avg.
Source	18.01	17.95	18.08	17.90	18.34	18.04	18.26	18.40	18.79	18.58	18.80	18.48	18.30
TENT [67]	29.42	8.12	1.28	0.69	0.47	0.88	0.68	2.50	0.78	0.87	2.97	1.14	4.15
ROID [39]	**39.33**	20.82	1.49	0.29	0.16	0.48	0.39	8.24	0.23	0.43	1.85	0.63	6.20
NOTE [18]	8.38	11.82	6.33	4.73	3.18	5.00	4.19	7.51	4.07	4.59	11.07	4.95	6.32
CoTTA [69]	33.13	19.33	4.87	3.20	2.67	3.78	3.67	10.30	4.80	5.50	7.89	6.29	8.78
TRIBE [61]	24.12	15.22	10.22	7.38	3.46	4.81	4.01	11.28	7.15	6.29	10.63	5.95	9.21
BN [41]	30.67	17.13	6.21	4.92	4.85	4.90	4.99	11.60	7.76	7.75	8.69	8.16	9.80
UnMIX-TNS [65]	20.36	14.45	20.26	15.58	17.33	15.43	17.19	21.33	16.72	17.66	14.96	17.62	17.40
RoTTA [71]	32.23	20.09	27.28	19.46	20.35	19.70	20.37	31.26	21.74	22.06	20.22	21.64	23.12
LAME [4]	17.45	17.74	25.52	27.79	28.23	26.48	26.87	24.30	26.56	26.46	25.62	25.61	24.88
UniTTA [12]	21.93	22.00	29.75	**33.17**	**33.58**	31.71	**31.95**	27.98	**34.32**	**33.13**	31.52	**32.42**	30.29
DDA [17]	29.89	30.32	29.88	29.94	26.33	29.58	26.28	31.67	31.28	27.29	31.3	28.18	29.33
SDA (Ours)	32.42	**32.72**	**32.34**	32.50	27.75	**32.06**	27.88	**34.36**	34.05	29.06	**34.02**	29.99	**31.60 (+2.27)**

Table 22. Data stream sensitivity comparison on ImageNet-C [24] under 12 class/domain balance/imbalance settings in the UniTTA benchmark [12]. Detailed introduction of the settings can be found in [12]. Briefly, ({i, n, 1}, {1, u}) denotes correlation and imbalance settings, where {i, n, 1} represent i.i.d., non-i.i.d. and continual, respectively, and {1, u} represent balance and imbalance, respectively. The best results are in **bold** and the second-best results are underlined.

References

[1] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *IJCV*, 2018. 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 7

[3] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *TMLR*, 2023. 3

[4] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *CVPR*, 2022. 7, 8, 1, 4

[5] Goirik Chakrabarty, Manogna Sreenivas, and Soma Biswas. Santa: Source anchoring network and target alignment for continual test time adaptation. *TMLR*, 2023. 7, 8

[6] Liang-Chieh Chen. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 7, 1

[7] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*, 2019. 3

[8] MMPreTrain Contributors. Openmmlab's pre-training toolbox and benchmark. `https://github.com/open-mmlab/mmpretrain`, 2023. 4

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3, 4

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 4, 6, 1

[11] Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *CVPR*, 2023. 7, 8

[12] Chaoqun Du, Yulin Wang, Jiayi Guo, Yizeng Han, Jie Zhou, and Gao Huang. Unitta: Unified benchmark and versatile framework towards realistic test-time adaptation. *arXiv preprint arXiv:2407.20080*, 2024. 2, 7, 8, 1, 4

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 7, 1

[14] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *CVPR*, 2024. 3

[15] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. In *NeurIPS*, 2022. 1, 2

[16] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 1

[17] 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 *CVPR*, 2023. 1, 2, 3, 4, 5, 6, 8

[18] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. In *NeurIPS*, 2022. 7, 8, 1, 4

[19] Jiayi Guo, Chaoqun Du, Jiangshan Wang, Huijuan Huang, Pengfei Wan, and Gao Huang. Assessing a single image in reference-guided image synthesis. In *AAAI*, 2022. 3

[20] Jiayi Guo, Hayk Manukyan, Chenyu Yang, Chaofei Wang, Levon Khachatryan, Shant Navasardyan, Shiji Song, Humphrey Shi, and Gao Huang. Faceclip: Facial image-to-video translation via a brief text description. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3

[21] Jiayi Guo, Chaofei Wang, You Wu, Eric Zhang, Kai Wang, Xingqian Xu, Humphrey Shi, Gao Huang, and Shiji Song. Zero-shot generative model adaptation via image-specific prompt learning. In *CVPR*, 2023. 3

[22] Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *CVPR*, 2024. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 1

[24] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 5, 6, 1, 4

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 1

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3, 4

[27] Gao Huang. Dynamic neural networks: advantages and challenges. *National Science Review*, 2024. 7

[28] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 6, 1

[29] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *CVPR*, 2020. 1

[30] Black Forest Labs. Flux. `https://blackforestlabs.ai/`, 2024. 2, 7, 1

[31] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *CVPR*, 2020. 1

[32] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. In *ICLR Workshops*, 2017. 1, 2

[33] Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *CVPR*, 2023. 6, 1

[34] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020. 1

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 7, 1

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 4, 6, 1

[37] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 4, 6, 1

[38] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018. 1

[39] Robert A Marsden, Mario Döbler, and Bin Yang. Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. In *WACV*, 2024. 7, 8, 1, 4

[40] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Lens: Localization enhanced by nerf synthesis. In *CoRL*, 2022. 3

[41] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020. 7, 8, 1, 4

[42] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. In *NeurIPS*, 2023. 7, 1

[43] Zanlin Ni, Yulin Wang, Renping Zhou, Jiayi Guo, Jinyi Hu, Zhiyuan Liu, Shiji Song, Yuan Yao, and Gao Huang. Revisiting non-autoregressive transformers for efficient image synthesis. In *CVPR*, 2024. 3

[44] Zanlin Ni, Yulin Wang, Renping Zhou, Yizeng Han, Jiayi Guo, Zhiyuan Liu, Yuan Yao, and Gao Huang. Enat: Rethinking spatial-temporal interactions in token-based image synthesis. In *NeurIPS*, 2024.

[45] Zanlin Ni, Yulin Wang, Renping Zhou, Rui Lu, Jiayi Guo, Jinyi Hu, Zhiyuan Liu, Yuan Yao, and Gao Huang. Adanat: Exploring adaptive policy for token-based image generation. In *ECCV*, 2024. 3

[46] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 6, 1

[47] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *ICML*, 2022. 1, 2, 3, 5, 6

[48] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, 2022. 7, 8

[49] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2, 3, 6, 1

[50] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. Learning deep object detectors from 3d models. In *ICCV*, 2015. 3

[51] Mihir Prabhudesai, Tsung-Wei Ke, Alexander Cong Li, Deepak Pathak, and Katerina Fragkiadaki. Diffusion-tta: Test-time adaptation of discriminative models via generative feedback. In *NeurIPS*, 2023. 2

[52] Yifan Pu, Zhuofan Xia, Jiayi Guo, Dongchen Han, Qixiu Li, Duo Li, Yuhui Yuan, Ji Li, Yizeng Han, Shiji Song, et al. Efficient diffusion transformer with step-wise dynamic attention mediators. In *ECCV*, 2024. 2

[53] Yifan Pu, Yiming Zhao, Zhicong Tang, Ruihong Yin, Haoxing Ye, Yuhui Yuan, Dong Chen, Jianmin Bao, Sirui Zhang, Yanbin Wang, et al. Art: Anonymous region transformer for variable multi-layer transparent image generation. *arXiv preprint arXiv:2502.18364*, 2025. 3

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3

[55] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 3

[56] Artem Rozantsev, Vincent Lepetit, and Pascal Fua. On rendering synthetic images for training an object detector. *Computer Vision and Image Understanding*, 2015. 3

[57] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020. 2

[58] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 1

[59] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *CVPR*, 2018. 3

[60] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 5, 7

[61] Yongyi Su, Xun Xu, and Kui Jia. Towards real-world test-time adaptation: Tri-net self-training with balanced normalization. In *AAAI*, 2024. 7, 8, 1, 4

[62] 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 *ICML*, 2020. 1, 2

[63] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *CVPR*, 2024. 3

[64] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. In *NeurIPS*, 2024. 3

[65] Devavrat Tomar, Guillaume Vray, Jean-Philippe Thiran, and Behzad Bozorgtabar. Un-mixing test-time normalization statistics: Combatting label temporal correlation. *arXiv preprint arXiv:2401.08328*, 2024. 8, 1, 4

[66] Yun-Yun Tsai, Fu-Chen Chen, Albert YC Chen, Junfeng Yang, Che-Chun Su, Min Sun, and Cheng-Hao Kuo. Gda: Generalized diffusion for robust test-time adaptation. *CVPR*, 2024. 1, 2, 3, 5, 6

[67] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 1, 2, 5, 7, 8, 4

[68] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024. 2

[69] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *CVPR*, 2022. 1, 2, 7, 8, 4

[70] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Tsung-Yi Lin, Alberto Rodriguez, and Phillip Isola. Nerf-supervision: Learning dense object descriptors from neural radiance fields. In *ICRA*, 2022. 3

[71] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *CVPR*, 2023. 7, 8, 1, 4

[72] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *NeurIPS*, 2022. 1, 2, 5, 6