This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Jiayi Guo[1,2] Junhao Zhao[1] Chaoqun Du[1] Yulin Wang[1] Chunjiang Ge[1] Zanlin Ni[1]
Shiji Song[1] Humphrey Shi[2*] Gao Huang[1*]
[1]Tsinghua University [2]SHI Labs @ Georgia Tech

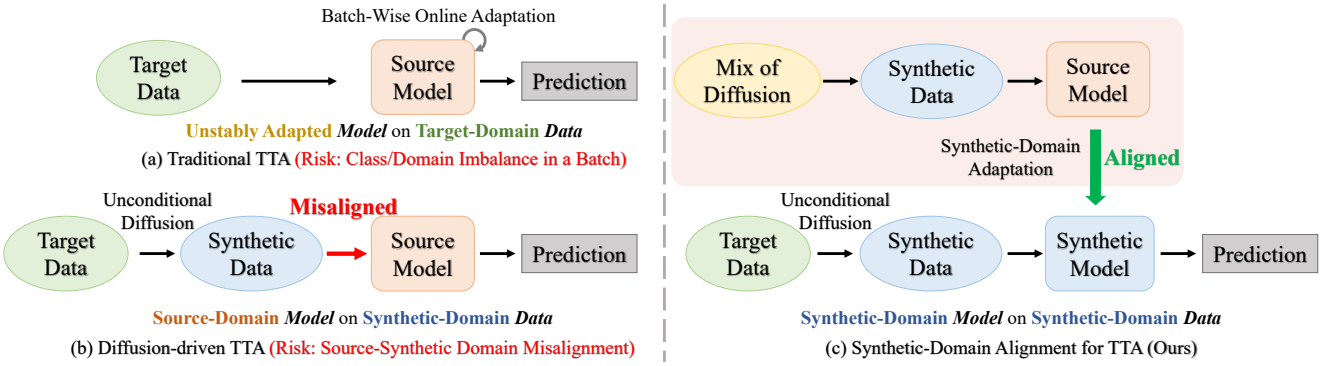


Figure 1. **Comparison of different test-time adaptation (TTA) frameworks.** (a) Traditional TTA methods continuously adapt source model weights to fit target data batches. However, their performance is sensitive to the amount and order of target data streams, *e.g.*, adapting the model with batches containing data from only a single category can lead to overfitting. (b) Diffusion-driven TTA methods project the target data back to the synthetic domain of diffusion models, which still maintains domain misalignment with the source domain. (c) We propose the Synthetic-domain Alignment (SDA) framework for TTA, which simultaneously aligns the domains of the source model and target data with the same synthetic domain for superior performance.

Abstract

Test-time adaptation (TTA) aims to improve the performance of source-domain pre-trained models on previously unseen, shifted target domains. Traditional TTA methods primarily adapt model weights based on target data streams, making model performance sensitive to the amount and order of target data. The recently proposed diffusion-driven TTA methods mitigate this by adapting model inputs instead of weights, where an unconditional diffusion model, trained on the source domain, transforms target-domain data into a synthetic domain that is expected to approximate the source domain. However, in this paper, we reveal that although the synthetic data in diffusion-driven TTA seems indistinguishable from the source data, it is unaligned with, or even markedly different from the latter for deep networks. To address this issue, we propose a ***S****ynthetic-****D****omain* ***A****lignment (SDA) framework. Our key insight is to fine-tune the source model with synthetic data to ensure better alignment. Specifically, we first employ a conditional diffusion model to generate labeled samples, creating a synthetic dataset. Subsequently, we use the aforementioned unconditional diffusion model to add noise to and denoise each sample before fine-tuning. This Mix of Diffusion (MoD) process mitigates the potential domain misalignment between the conditional and unconditional models. Extensive experiments across classifiers, segmenters, and multimodal large language models (MLLMs,* e.g.*, LLaVA) demonstrate that SDA achieves superior domain alignment and consistently outperforms existing diffusion-driven TTA methods. Our code is available at https://github.com/SHI-Labs/Diffusion-Driven-Test-Time-Adaptation-via-Synthetic-Domain-Alignment.*

1. Introduction

Test-Time Adaptation (TTA) [15, 17, 31, 46, 61, 65, 66, 68, 71] is an emerging research field that tackles domain misalignment when source models are evaluated on shifted target data. Unlike traditional domain adaptation (DA) [16, 37, 57] and source-free adaptation (SFA) [28, 30, 33], TTA addresses more practical scenarios where neither source data nor complete target data are accessible. Instead, the adapta-

*Corresponding authors.

	Model Adaptation Direction	Data Adaptation Direction
Traditional TTA	Expected: Source→Target Risk: Imbalanced Target Streams	N/A
Diffusion-driven TTA	N/A	Expected: Target→Source Actual: Target→Synthetic
SDA (Ours)	Source→Synthetic	Target→Synthetic

Table 1. **Adaptation directions of different TTA methods.**

tion relies solely on streaming batches of target data.

Traditional TTA methods (Fig. 1a) [15, 31, 50, 56, 61, 66, 68, 71] typically employ a *source-to-target* model adaptation framework. These approaches continuously update the source model weights by processing target data batches. Without annotations, the adaptation process relies either on batch-wise updates of model statistics [31, 56, 66, 68], or unsupervised or self-supervised auxiliary tasks [15, 50, 61]. However, small or imbalanced batches may poorly represent the target domain, making these approaches sensitive to the amount and order of the data stream [12, 17, 66]. For instance, adapting the model with batches containing data from only a single category can lead to overfitting.

Recently, the impressive generation capabilities of diffusion models [22, 25, 29, 48, 51, 53, 67] have sparked the development of diffusion-driven TTA methods (Fig. 1b) [17, 46, 65], leveraging a *target-to-source* framework. These approaches employ an unconditional diffusion model, pretrained on the source domain, aiming to project each target sample to the source domain independently. This enables the source model to make predictions without modifying its weights. As a preliminary work, DiffPure [46] addresses adversarial perturbations by first applying a forward diffusion process, introducing a small amount of noise to the target data, followed by a reverse diffusion process to restore a clean image to approach the source domain. Building on this concept, more recent studies [17, 65] tackle challenging domain shifts—such as severe data corruption—by incorporating additional structural guidance from the target data, helping preserve semantics and improve performance.

In this paper, we uncover that while diffusion-driven TTA methods aim to project target data back to the source domain, the projected target data remains confined within the synthetic domain of the unconditional diffusion model. As the synthetic-domain data are ultimately processed by the source-domain model, this domain misalignment limits the final performance. To address this issue, we propose Synthetic-Domain Alignment (**SDA**) (Fig. 1c), a new category of framework for TTA tasks which *simultaneously aligns the domains of the source model and target data with the same synthetic domain of a diffusion model* (Tab. 1).

SDA distinguishes itself from existing diffusion-driven TTA methods [17, 46, 65] by introducing an additional *source-to-synthetic* model adaptation phase before testing on adapted target data (Fig. 2). Since the adapted target data aligns with the synthetic domain generated by an unconditional diffusion model, SDA aims to adapt the source model to this same synthetic domain. Specifically, SDA employs a Mix-of-Diffusion (**MoD**) technique to generate synthetic data for model adaptation. Given that the source data is inaccessible after pretraining, MoD first uses a conditional diffusion model to generate samples conditioned on domain-agnostic labels, creating a labeled synthetic dataset. Subsequently, the aforementioned unconditional diffusion model is leveraged by MoD add noise to and denoise these samples, addressing potential domain misalignment between the conditional and unconditional models. With a sufficiently large synthetic dataset, the fine-tuned model becomes highly effective at discriminating within the synthetic domain. Thus, the SDA framework transforms the cross-domain TTA task into an in-domain prediction task by aligning both the source model and target data with the same synthetic domain.

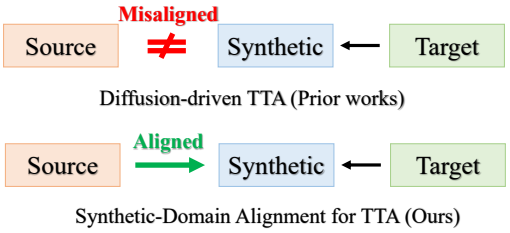
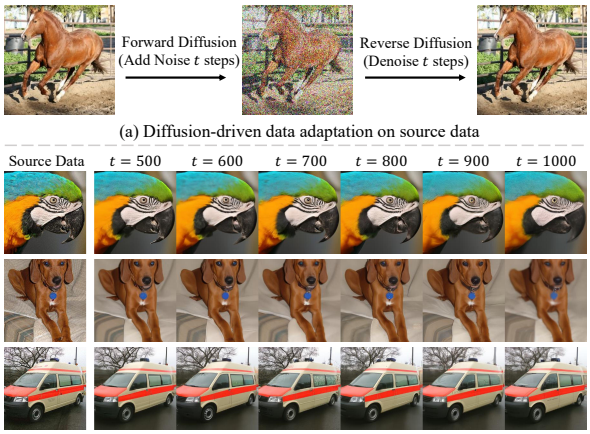


Figure 2. **Enhanced domain alignment with our framework**. Prior diffusion-driven TTA methods struggle with the domain misalignment between the source model and synthetic data, which we resolve by aligning the source model to the synthetic domain.

SDA is a general framework, not limited to specific fine-tuning techniques or diffusion-driven data adaptation methods. This flexibility allows future advancements in these areas to further broaden its applicability. Extensive experiments across classifiers, segmenters, and multimodal large language models (MLLM, e.g., LLaVA) demonstrate that SDA consistently outperforms existing methods. Moreover, the effectiveness of our approach is reinforced through visualization analysis and ablation studies.

2. Related Work

Test-time adaptation (TTA) is an emerging research area that addresses domain shifts by adapting either models [15, 31, 50, 56, 61, 66, 68, 71] or data [17, 46, 65] during evaluation on streaming target data batches. Early model adaptation methods update batch normalization statistics to match the target distribution [31, 56, 66, 71], while others leverage self-supervised tasks like rotation prediction [61] or image restoration [15, 50] to adjust model weights. However, these approaches rely heavily on continuous weight updates, making them sensitive to the amount, order, and diversity of target data. In contrast, diffusion-driven TTA methods [17, 46, 65] focus on data adaptation by projecting each target sample back into the source domain, achiev-



(a) Diffusion-driven data adaptation on source data

(b) Diffusion-driven data adaptation results (on source data) under different timesteps

Figure 3. **(a) Illustration of diffusion-driven data adaptation on source data and (b) Adapted images across different timesteps.** The results are obtained using DDA [17], with no noticeable visual degradation observed in the adapted images.

ing stable performance without online model updates. DiffPure [46] purifies adversarial samples with diffusion models, while DDA [17] and GDA [65] use structural guidance to preserve image content under severe corruption. Building on this, our work investigates the issue of misalignment between domains of the source model and synthetic images in diffusion-driven TTA and proposes a new TTA framework.

Synthetic data for discriminative tasks. Synthetic data, generated by models [20, 21, 42–44, 48, 52, 53] rather than collected from the real world, has shown significant potential in enhancing visual representations for various discriminative tasks [14, 62, 63]. It has been effectively applied in areas such as recognition [3, 63], object detection [49, 55], semantic segmentation [7, 54, 58], image assessment [19], autonomous driving [1], and robotics [39, 69]. In this work, we explore the potential of diffusion-generated synthetic data to enhance domain alignment in TTA.

3. Methodology

In this section, we introduce the background of diffusion-driven TTA methods and identify their source-synthetic domain misalignment issue in Sec. 3.1 and Sec. 3.2. To tackle this issue, we propose the synthetic-domain alignment (SDA) framework in Sec. 3.3 and introduce its key technique, the mix of diffusion (MoD) in Sec. 3.4.

3.1. Background

Diffusion process. Given a source data point $\boldsymbol{x}_0^{\text{src}} \sim p_0^{\text{src}}$, diffusion models ϵ_θ [25] gradually transform p_0^{src} into a Gaussian noise distribution $N(\mathbf{0}, \boldsymbol{I})$ through a T-step forward diffusion process. At each timestep $t \in \{1, 2, \cdots, T\}$, the intermediate state $\boldsymbol{x}_t^{\text{src}} \sim p_t^{\text{src}}$ is computed as:

$$\boldsymbol{x}_t^{\text{src}} = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}^{\text{src}} + \sqrt{\beta_t}\boldsymbol{\epsilon}_t, \tag{1}$$

Timestep t	500	600	700	800	900	1000
Swin-B (Source-Synthetic, Misaligned)	61.6	57.8	55.7	52.6	45.0	38.5
Swin-B (Synthetic-Synthetic, Aligned)	67.6	65.9	65.0	61.5	55.7	48.4
Δ	+6.0	+8.1	+9.3	+8.9	+10.7	+9.9
ConvNeXt-B (Source-Synthetic, Misaligned)	65.1	61.3	60.3	57.2	50.0	41.5
ConvNeXt-B (Synthetic-Synthetic, Aligned)	70.6	68.2	67.4	64.9	58.5	50.7
Δ	+5.5	+6.9	+7.1	+7.7	+8.5	+9.2

Table 2. **Source model accuracy across different timesteps of diffusion-driven data adaptation.** For suitable timesteps for TTA ($t \geq 500$) [17], model accuracy shows a monotonically decreasing trend with the growth of timestep, indicating the increase of the misalignment of source domain p_0^{src} and synthetic domain $p_{0,\text{u}}^{\text{syn}}$. By aligning the source model to the synthetic domain, our methods (rows 3 & 6) significantly help alleviate the performance degradation. Results are evaluated on the ImageNet [9] validation set.

where $\boldsymbol{\epsilon}_t$ is random Gaussian noise and $\beta_t \in (0, 1)$ represents the diffusion rate at step t. By defining $\alpha_t = 1 - \beta_t$, $\overline{\alpha_t} = \prod_{t=1}^{T} \alpha_t$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{I})$, we obtain:

$$\boldsymbol{x}_t^{\text{src}} = \sqrt{\overline{\alpha_t}}\boldsymbol{x}_0^{\text{src}} + \sqrt{1-\overline{\alpha_t}}\boldsymbol{\epsilon}. \tag{2}$$

The reverse diffusion process recovers a clean $\widehat{\boldsymbol{x}_0^{\text{src}}}$ by progressively removing noise from $\boldsymbol{x}_T^{\text{src}}$:

$$\widehat{\boldsymbol{x}_{t-1}^{\text{src}}} = \frac{1}{\sqrt{\alpha_t}}\left(\widehat{\boldsymbol{x}_t^{\text{src}}} - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha_t}}}\epsilon_\theta(\widehat{\boldsymbol{x}_t^{\text{src}}}, t)\right) + \sigma_t\boldsymbol{\epsilon}, \tag{3}$$

where σ_t is the posterior noise variance [25].

Diffusion-driven data adaptation. Denote by p_0^{trg} the target data distribution, from which each target data point $\boldsymbol{x}_t^{\text{trg}}$ is sampled. Prior work [46] demonstrates that:

$$D_{\text{KL}}(p_{t+1}^{\text{src}} || p_{t+1}^{\text{trg}}) - D_{\text{KL}}(p_t^{\text{src}} || p_t^{\text{trg}}) \leq 0, \tag{4}$$

where D_{KL} is the KL divergence. Since $p_T^{\text{src}} = p_T^{\text{trg}} = N(\mathbf{0}, \boldsymbol{I})$, for any arbitrarily small value δ, there exists a minimum timestep t^* such that $D_{\text{KL}}(p_{t^*}^{\text{src}} || p_{t^*}^{\text{trg}}) < \delta$. As discussed in [17], a diffusion process with $T = 1000$ requires $t^* \geq 500$ to eliminate domain shifts. We empirically validate this choice of t^* in the supplementary materials.

The initial diffusion-driven data adaptation in DiffPure [46] is carried out as follows: Given the minor divergence between $p_{t^*}^{\text{src}}$ and $p_{t^*}^{\text{trg}}$, we recover each $\boldsymbol{x}_{t^*}^{\text{trg}} \sim p_{t^*}^{\text{trg}}$ to its corresponding $\widehat{\boldsymbol{x}_0^{\text{src}}} \sim p_0^{\text{src}}$ by executing the reverse diffusion process (Eq. 3) t^* times.

Since the reverse process is stochastic, subsequent works [17, 65] further introduce additional structure guidance to ensure the content consistency between each $\boldsymbol{x}_0^{\text{trg}}$ and its adapted $\widetilde{\boldsymbol{x}_0^{\text{src}}}$. In this work, we adopt the same data adaptation process as DDA [17]:

$$\widetilde{\boldsymbol{x}_{t-1}^{\text{src}}} = \widehat{\boldsymbol{x}_{t-1}^{\text{src}}} - \boldsymbol{w}\nabla_{\widehat{\boldsymbol{x}_t^{\text{src}}}}\left\|\phi(\boldsymbol{x}_0^{\text{trg}}) - \phi(\widetilde{\boldsymbol{x}_{0,t}^{\text{src}}})\right\|_2, \tag{5}$$

where $\widehat{\boldsymbol{x}_{t-1}^{\text{src}}}$ is computed as Eq. 3, $\boldsymbol{w}$ is the structure guidance scale, ϕ is a structure extractor [17] and $\widetilde{\boldsymbol{x}_{0,t}^{\text{src}}}$ is an esti-

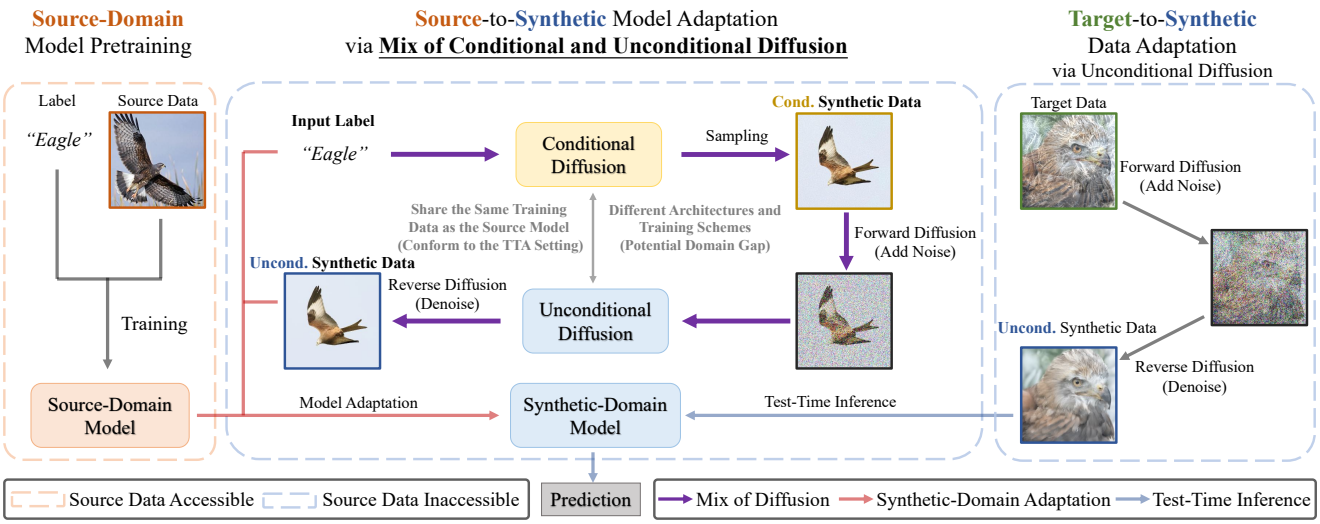


Figure 4. **Overview of the Synthetic-Domain Alignment (SDA) framework.** SDA is a novel TTA framework aligning both the domains of the source model and the target data with the synthetic domain. SDA involves three phases: (left): a *source-domain model pretraining* phase, where the source model is trained on source data prior to TTA; (middle): a *source-to-synthetic model adaptation* phase, where the source model is adapted to a synthetic-domain model using synthetic data generated via a Mix of Diffusion (MoD) technique; and (right): a *target-to-synthetic data adaptation* phase, where target data is adapted into synthetic data using an unconditional diffusion model. Finally, the adapted synthetic data is fed into the synthetic-domain model for test-time inference.

mate of $\widetilde{\boldsymbol{x}_0^{\text{src}}}$ at timestep t [25]. Diffusion-driven data adaptation is typically performed with unconditional diffusion models [10] since the target data labels are unknown.

3.2. Source-Synthetic Domain Misalignment

Unlike a *real* source data point $\boldsymbol{x}_0^{\text{src}}$, the recovered version $\widetilde{\boldsymbol{x}_0^{\text{src}}}$, derived from $\boldsymbol{x}_{t^*}^{\text{trg}}$ is *synthetic*. Specifically, $\boldsymbol{x}_0^{\text{src}}$ follows the source domain p_0^{src}, whereas $\widetilde{\boldsymbol{x}_0^{\text{src}}}$ follows the synthetic domain $p_{0,\text{u}}^{\text{syn}}$ of an unconditional diffusion model $\epsilon_\theta^{\text{u}}$ with parameters θ. In this section, we empirically reveal the misalignment between the source domain p_0^{src} and the synthetic domain $p_{0,\text{u}}^{\text{syn}}$, and investigate how this misalignment impacts the performance of the source model. For simplicity and clarity, we will substitute $\widetilde{\boldsymbol{x}_0^{\text{src}}}$ with $\boldsymbol{x}_{0,\text{u}}^{\text{syn}}$ throughout the following discussion.

Based on the above analysis, diffusion-driven TTA methods face two potential misalignments: (1) source-target domain misalignment arising from inherent data distribution shifts, which has been the primary focus of prior research, and (2) source-synthetic domain misalignment introduced by diffusion models, which we address as the main focus of this work, complementing existing efforts.

To precisely examine the impact of the source-synthetic domain misalignment and isolate it from the influence of the source-target domain misalignment, we evaluate the performance of the source model f on synthetic data $\boldsymbol{x}_{0,\text{u}}^{\text{syn}}$ adapted by the diffusion model $\epsilon_\theta^{\text{u}}$ from source data $\boldsymbol{x}_0^{\text{src}}$ (Fig. 3a). Specifically, we test ImageNet pretrained models [9] on the ImageNet validation set adapted by the popular diffusion-driven data adaptation method, DDA [17] across different timesteps t. As indicated in Tab. 2, with an increase in t, the accuracy of the model exhibits a monotonically decreasing trend. For ideal $t^* \geq 500$ for TTA tasks, performance degradation of more than 18.8% for ConvNeXt [36] and 21.8% for Swin [35] is observed compared to their official results [8] on source data which are 83.4% and 83.9%. This indicates a significant domain misalignment between p_0^{src} and $p_{0,\text{u}}^{\text{syn}}$. By fine-tuning the source model on the synthetic data generated by our Mix of Diffusion process (introduced in Sec. 3.4), the aligned models f' achieve performance improvements of 5.5% for ConvNeXt and 6.0% for Swin.

In Fig. 3b, we show that the diffusion synthetic data $\boldsymbol{x}_{0,\text{u}}^{\text{syn}}$ and source data $\boldsymbol{x}_0^{\text{src}}$ exhibit no noticeable visual differences across different timesteps t. This further suggests that the performance degradation is not due to the quality of diffusion-generated images but rather the implicit misalignment between the source and synthetic domains.

3.3. Synthetic-Domain Alignment Framework

Given that the diffusion-adapted data $\boldsymbol{x}_{0,\text{u}}^{\text{syn}}$ aligns more closely with the synthetic domain $p_{0,\text{u}}^{\text{syn}}$ instead of the source domain p_0^{src}, we propose simultaneously adapting the source model f to the same synthetic domain $p_{0,\text{u}}^{\text{syn}}$. By doing so, the alignment between the data and model within $p_{0,\text{u}}^{\text{syn}}$ can be effectively achieved. To this end, we introduce a novel TTA framework: **S**ynthetic-**D**omain **A**lignment (**SDA**).

In Fig. 4, we present the complete diagram of SDA, which consists of three key phases: (1) a *source-domain*

model pretraining phase, where the source model f is trained on source data $\boldsymbol{x}_0^{\text{src}}$ prior to TTA; (2) a *source-to-synthetic model adaptation* phase, where f is fine-tuned to a synthetic-domain model f'; and (3) a *target-to-synthetic data adaptation* phase, where target data $\boldsymbol{x}_0^{\text{trg}}$ is adapted into synthetic data $\boldsymbol{x}_{0,\text{u}}^{\text{syn}}$ using an unconditional diffusion model $\epsilon_\theta^{\text{u}}$ following Eq. 5. Finally, the adapted synthetic data $\boldsymbol{x}_{0,\text{u}}^{\text{syn}}$ is fed into the synthetic-domain model f' for test-time inference. Consistent with the standard TTA protocol [66], the source data is accessible only during the source model pretraining phase and remains inaccessible during both the model and data adaptation phases.

The rationale behind SDA is straightforward: by aligning the model and data domains to the same synthetic domain, the original *cross-domain* TTA task is transformed into an easier *in-domain* prediction task, thus addressing the core challenge of TTA and improving performance.

Since the pretraining phase follows standard supervised learning, and the data adaptation phase aligns with DDA [17], we omit detailed explanations of these phases. Instead, we focus on how we adapt the source model f to an ideal synthetic domain model f' using a novel technique called Mix-of-Diffusion (**MoD**), introduced next.

3.4. Model Adaptation via Mix of Diffusion

As shown in Fig. 4, MoD consists of two main processes: a data generation process powered by a conditional diffusion model ϵ_η^{c} with parameters η, and a data alignment process powered by the same unconditional diffusion model $\epsilon_\theta^{\text{u}}$ used in the target-to-synthetic data adaptation phase. It is worth noting that, both ϵ_η^{c} and $\epsilon_\theta^{\text{u}}$ are also pretrained on source data $\boldsymbol{x}_0^{\text{src}}$ and have never been exposed to target data $\boldsymbol{x}_0^{\text{trg}}$, in accordance with the TTA setting.

Conditional diffusion data generation. The generation process leverages the conditional generation capability of ϵ_η^{c} to synthesize conditional synthetic data $\boldsymbol{x}_{0,\text{c}}^{\text{syn}} \sim p_{0,\text{c}}^{\text{syn}}$. In the context of TTA, the source and target domains share the same domain-agnostic label set $\{y_i\}_{i=1}^K$. Utilizing ϵ_η^{c}, we uniformly generate samples for each class y_i from Gaussian noise ($\boldsymbol{x}_{T,c}^{\text{syn}}$) through a T-step reverse diffusion process:

$$\boldsymbol{x}_{t-1,c}^{\text{syn}} = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x}_{t,c}^{\text{syn}} - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\eta^{\text{c}}(\boldsymbol{x}_{t,c}^{\text{syn}}, t, y_i)) + \sigma_t \boldsymbol{\epsilon}. \quad (6)$$

The generation capability of ϵ_η^{c} allows for the construction of a labeled synthetic-domain dataset $\{\boldsymbol{x}_{0,\text{c}}^{\text{syn}}, y\}^N$ of arbitrary size N without any manual data collection. By fine-tuning the source model f on this synthetic dataset, an adapted model f'_{c} on domain $p_{0,\text{c}}^{\text{syn}}$ can be obtained.

Unconditional diffusion data alignment. However, since the test-time adapted data $\boldsymbol{x}_{0,\text{u}}^{\text{syn}}$ is subject to domain $p_{0,\text{u}}^{\text{syn}}$, we argue that there is still potential misalignment between different synthetic domains $p_{0,\text{c}}^{\text{syn}}$ and $p_{0,\text{u}}^{\text{syn}}$. This is mainly because of the differences in architectures and training schemes between ϵ_η^{c} and $\epsilon_\theta^{\text{u}}$. We empirically validate this argument in Tab. 10, which indicates that the model f'_{c} adapted to domain $p_{0,\text{c}}^{\text{syn}}$ performs worse than the model f'_{u} adapted to domain $p_{0,\text{u}}^{\text{syn}}$ during testing.

To obtain the f'_{u} on domain $p_{0,\text{u}}^{\text{syn}}$, we mirror the target-to-synthetic data adaptation phase as a conditional synthetic data ($\boldsymbol{x}_{0,\text{c}}^{\text{syn}}$) to unconditional synthetic data ($\boldsymbol{x}_{0,\text{u}}^{\text{syn}}$) adaptation process. In specific, we use the same t^* in Sec. 3.1 with Eq. 2 to obtain $\boldsymbol{x}_{t^*,\text{c}}^{\text{syn}}$. According to the analysis in Sec. 3.1, $\boldsymbol{x}_{t^*,\text{c}}^{\text{syn}}$ will be indistinguishable to its counterpart $\boldsymbol{x}_{t^*,\text{u}}^{\text{syn}}$ on domain $p_{t^*,\text{u}}^{\text{syn}}$. Therefore, we have:

$$\boldsymbol{x}_{t^*,\text{u}}^{\text{syn}} \approx \boldsymbol{x}_{t^*,\text{c}}^{\text{syn}} = \sqrt{\overline{\alpha}_t}\boldsymbol{x}_{0,c}^{\text{syn}} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}. \quad (7)$$

Then, following Eq. 3 and Eq. 5, the noisy $\boldsymbol{x}_{t^*,\text{u}}^{\text{syn}}$ can be gradually denoised to $\boldsymbol{x}_{0,\text{u}}^{\text{syn}}$. Finally, the expected synthetic-domain model f'_{u} is obtained by fine-tuning the source-domain model f'_{u} on dataset $\{\boldsymbol{x}_{0,\text{u}}^{\text{syn}}, y\}^N$.

Ensembling. As noted in previous diffusion-driven TTA methods [17], while diffusion models generally perform well for data adaptation, they may occasionally produce data points that are less recognizable than the original target data. To address this, prior approaches use an ensemble of model predictions on $\boldsymbol{x}_0^{\text{trg}}$ and $\boldsymbol{x}_{0,\text{u}}^{\text{syn}}$ as the final output. Following this protocol, the final prediction in SDA is:

$$\hat{y} = \arg\max_y (q(y|\boldsymbol{x}_0^{\text{trg}}) + q'(y|\boldsymbol{x}_{0,\text{u}}^{\text{syn}})), \quad (8)$$

where $q(\cdot)$ and $q'(\cdot)$ are output distributions of source model f and synthetic-domain model f'_{u}, respectively.

4. Experiments

In this section, we first evaluate SDA on ImageNet classifiers with standard TTA benchmarks in Sec. 4.1. Next, we assess SDA's scalability across different dataset sizes, tasks, and model architectures in Sec. 4.2. In Sec. 4.3, we demonstrate SDA's advantages through Grad-CAM visualizations [59] and data stream sensitivity tests. Finally, ablation studies in Sec. 4.4 validate the design choices in our SDA framework.

Model	Source	MEMO	DiffPure	GDA	DDA	SDA (Ours)
ResNet-50	18.7	24.7	16.8	31.8	29.7	**32.5** (+2.8)
Swin-T	33.5	29.5	24.8	42.2	40.0	**42.5** (+2.5)
ConvNeXt-T	39.3	37.8	28.8	44.8	44.2	**47.0** (+2.8)
Swin-B	40.5	37.0	28.9	-	44.5	**47.4** (+2.9)
ConvNeXt-B	45.6	45.8	32.7	-	49.4	**51.9** (+2.5)

Table 3. **Comparison results on ImageNet-C**. We compare SDA with source models, MEMO [71], DiffPure [46], GDA [65] and DDA [17]. Results are the average accuracy across 15 adaptation domains at severity level 5. SDA shows consistent performance improvements compared to baselines.

	Gaussian	Shot	Impluse	Defocus	Glass	Motion	Zoom	Frost	Snow	Fog	Brightness	Contrast	Elastic	Pixelate	JEPG	Avg.
Source	48.0	47.1	48.3	32.4	14.4	39.3	39.2	55.7	50.1	51.5	75.1	52.0	29.5	48.1	60.3	46.1
DiffPure	42.6	42.0	40.9	16.8	22.2	19.1	23.0	39.0	29.1	9.1	62.5	3.0	34.3	46.8	60.7	32.7
DDA	59.9	55.7	55.9	30.7	30.5	38.1	39.2	53.2	48.0	41.9	71.0	48.1	46.9	59.5	62.0	49.4
SDA (Ours)	60.3	57.6	58.1	35.4	35.3	42.9	42.4	55.5	49.9	44.6	72.6	45.4	50.2	63.0	64.5	**51.9** (+2.5)

Table 4. **Detailed comparisons of SDA and baselines across 15 adaptation domains of ImageNet-C.** SDA shows the best average accuracy. The results are tested with ConvNeXt-B. Comparisons using other models are deferred to the supplementary materials.

4.1. Main Results on ImageNet Classifiers

Settings. We choose DDA [17] as our primary competitor since it is the best open-sourced method. SDA is also compared with DiffPure [46], MEMO [71], and the recent SOTA GDA [65] using their reported results. Source model performance is reported as "Source". DiT [48] and ADM [10] are adopted to generate and align 50K synthetic data for 15-epoch finetuning. For different source models, the synthetic data only needs to be generated once. Our results are tested on standard TTA benchmarks, ImageNet-C [24] (severity level 5) and ImageNet-W [32] using various models including ResNet [23], Swin [35] and ConvNeXt [36]. More implementation details are provided in the supplementary materials.

Comparison results on ImageNet-C. We begin by evaluating the performance of SDA on ImageNet-C. As reported in Tab. 3, our proposed SDA consistently outperforms all baseline methods across different model architectures and sizes. We emphasize the performance improvement over DDA, as we adopted DDA for target data adaptation in SDA. Compared to DDA, our SDA improves accuracy by 2.5%-2.9%. This significant improvement indicates the misalignment between the source and synthetic domains, validating the effectiveness of our synthetic-domain alignment framework. Moreover, compared to the recent SOTA GDA, SDA also achieves an improvement of 2.2% with ConvNeXt-T. Notably, SDA focuses on synthetic domain alignment, an orthogonal research direction to existing diffusion-driven methods on better adapting the target data. Therefore, the performance of SDA could potentially be further enhanced with the release of more advanced codebases like GDA. Compared to the model adaptation method, MEMO, three diffusion-driven methods (SDA, DDA, and GDA) all demonstrate superior performance, highlighting the effectiveness of diffusion models in assisting TTA tasks. DiffPure presents worse results since it is primarily designed for adversarial attacks. Without the structural guidance introduced in DDA and GDA, DiffPure may not effectively recover images with severe domain shifts. In Tab. 4, we provide a detailed comparison of the results of SDA and baselines. SDA surpasses DiffPure in all 15 adaptation domains and outperforms DDA in 14 out of 15 domains, further affirming the superiority of SDA.

Model	Source	DiffPure	DDA	SDA (Ours)
ResNet-50	37.7	29.1	52.8	**54.7** (+1.9)
Swin-T	66.5	52.7	65.9	**67.3** (+1.4)
ConvNeXt-T	67.6	55.8	67.9	**69.4** (+1.5)
Swin-B	69.1	55.5	68.3	**70.6** (+2.3)
ConvNeXt-B	70.1	57.7	70.3	**72.3** (+2.0)

Table 5. **Comparison results on ImageNet-W**. SDA shows consistent performance improvements compared to baselines.

Comparison results on ImageNet-W. We extend our evaluation to ImageNet-W to assess SDA's performance under watermark-based domain shifts. As shown in Tab. 5, SDA consistently surpasses all baselines across different models. Compared to our primary baseline, DDA, SDA achieves accuracy gains ranging from 1.4% to 2.3%. Furthermore, the results in Tab. 5 reveal potential performance drops when DDA is applied to ImageNet-W with Swin-T and Swin-B models, suggesting that synthetic data may be less recognizable by the original source models. Although advancements in diffusion techniques could potentially improve outcomes, the consistent gains achieved by SDA indicate that aligning the source model with the synthetic domain offers a convenient and effective solution to enhance performance.

4.2. Scalability to Other Datasets, Tasks and Models

In addition to standard benchmarks, an effective TTA method should demonstrate its superiority across various dataset scales, task formats, and model architectures. In this section, we assess SDA from these aspects to validate its scalability. Implementation details of each experiment can be found in the supplementary materials.

Scaling to small datasets. We first evaluate the effectiveness of SDA on scenarios where both source and target domains are small-scale. Specifically, we test SDA on CIFAR-10-C [24] with ResNet-18 [23] as the source classifier. EDM [27] is used to generate synthetic data, and I-DDPM [45] is applied for data alignment. As shown in Tab. 6, SDA consistently outperforms both the source model and DDA, achieving an average accuracy improvement of 7.1% over DDA.

	Gaussian	Shot	Impluse	Defocus	Glass	Motion	Zoom	Frost	Snow	Fog	Brightness	Contrast	Elastic	Pixelate	JEPG	Avg.
Source	23.5	28.7	20.8	41.2	30.7	48.2	48.1	51.8	63.6	63.4	79.5	28.0	56.5	35.9	60.7	45.4
DDA	79.4	80.3	76.8	41.3	56.6	49.0	48.9	77.4	73.5	62.6	80.5	61.5	61.2	57.6	73.2	65.3
SDA (Ours)	82.6	83.5	79.5	59.6	61.4	65.3	64.6	76.5	78.4	71.5	88.9	53.4	75.4	64.4	80.7	**72.4** (+7.1)

Table 6. **Comparisons of SDA and baselines with ResNet-18 on CIFAR-10-C.** SDA shows the best average accuracy.

	Gaussian	Shot	Impluse	Defocus	Glass	Motion	Zoom	Frost	Snow	Fog	Brightness	Contrast	Elastic	Pixelate	JEPG	Avg.
Source	9.8	13.2	10.0	22.7	5.7	20.4	22.1	36.7	28.4	53.9	66.6	9.7	24.3	16.0	43.8	25.6
DDA	16.6	21.1	16.0	45.0	40.9	49.7	38.7	36.7	38.6	35.0	55.2	13.3	56.6	56.3	58.8	38.6
SDA (Ours)	17.0	21.1	16.2	45.5	43.0	51.1	39.9	38.7	40.2	36.3	57.9	13.1	59.1	58.1	59.7	**39.8** (+1.2)

Table 7. **Comparisons of SDA and baselines with DeepLabv3 on PASCAL VOC-C.** SDA shows the best average mIOU.

Scaling to Semantic Segmentation Tasks. We extend our evaluation to dense prediction tasks by using PASCAL-VOC-C [13] as a standard semantic segmentation benchmark with DeepLabv3 [6] as the source segmenter. Dataset Diffusion [41] generates synthetic segmentation data, and FLUX Schnell [29] is used for data alignment. As shown in Tab. 7, SDA achieves the best performance, with an average mIOU improvement of 1.2% over DDA.

Scaling to multimodal large language models (MLLMs). As an emerging research direction, MLLMs [26, 34] present advanced visual question answering [2] capability. We design a language-based classification task format to test how SDA can help MLLMs on TTA tasks on ImageNet-C: Given an image, ask an MLLM (LLaVA 1.5-7b [34] in our experiments) to choose the correct image class from four provided options. In Tab. 8, the "Source-Zero" setting tests the zero-shot results of pretrained LLaVA. The "Source" and "DDA" settings evaluate a source-data fine-tuned LLaVA while the "SDA" setting tests a synthetic-data fine-tuned LLaVA. The fine-tuning task format is the same as that at test time. Pretrained LLaVA already exhibits strong performance. While DDA improves results in some domains, it does not yield an overall gain compared to source-data fine-tuned LLaVA. In contrast, SDA aligns LLaVA with the synthetic domain, achieving the best accuracy with an improvement of 2.4%.

4.3. Analysis

Visualization. To demonstrate how synthetic data fine-tuning in SDA enhances the performance of diffusion-driven TTA methods, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) [59] to visualize the image regions that most influence classification scores across different images and models. As shown in Fig. 5, testing target images with the source model reveals distinct differences in activation maps and the occurrence of incorrect predictions compared to those from source images,

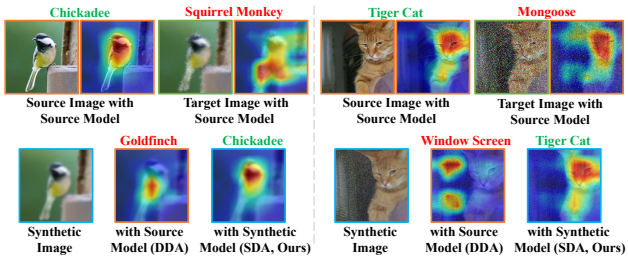


Figure 5. **Grad-CAM visualization comparison.** The first row shows activation maps for *source* and *target* images tested with the *source* model. The second row displays activation maps for diffusion *synthetic* images tested with the *source* model (DDA) and our *synthetic*-domain model (SDA). SDA aligns closely with the *source* model's response to *source* images.

underscoring the performance degradation due to domain shifts. Despite using adapted synthetic images, DDA still risks focusing on inappropriate regions and producing incorrect predictions. This highlights the domain misalignment of the synthetic data and source model. In contrast, SDA aligns both the data and model within the same synthetic domain, thereby producing activation maps and predictions that closely resemble those produced by the source model on source images.

Data stream sensitivity. In Tab. 9, we test data stream sensitivity using the UniTTA [12] benchmark which contains 12 class/domain balance/imbalance settings closely aligned with diverse real-world TTA scenarios. In addition to diffusion-driven TTA methods, we include 13 popular traditional TTA methods [4, 5, 11, 12, 18, 38, 40, 47, 60, 66, 68, 70] for comparison. We report the average accuracy of 12 settings using ResNet-50. The results indicate that diffusion-driven TTA methods are preferred in these challenging settings since they are insensitive to different variants of data streams. SDA maintains this insensitivity and showcases the best performance. Detailed comparisons for each setting are provided in the supplementary materials.

	Gaussian	Shot	Impluse	Defocus	Glass	Motion	Zoom	Frost	Snow	Fog	Brightness	Contrast	Elastic	Pixelate	JEPG	Avg.
Source-Zero	75.2	76.6	79.6	81.4	74.5	83.5	81.1	87.0	88.9	89.5	95.6	84.1	76.2	90.1	88.7	83.5
Source	86.8	86.0	88.6	90.1	85.5	93.5	89.7	93.6	95.0	94.9	98.1	89.5	85.6	96.2	94.8	91.2
DDA	92.4	91.5	92.5	88.6	90.1	92.2	88.5	93.0	93.1	89.6	97.8	75.8	89.6	96.4	97.0	91.2
SDA (Ours)	94.9	94.1	94.6	91.0	92.9	94.0	91.5	93.6	95.2	92.2	98.5	83.6	92.8	97.3	97.6	**93.6** (+2.4)

Table 8. **Comparisons of SDA and baselines with LLaVA on ImageNet-C.** SDA shows the best average accuracy.

Source	TENT [66]	ROID [38]	NOTE [18]	CoTTA [68]	TRIBE [60]	BN [40]	EATA [47]	SANTA [5]	RMT [11]	UnMIX [64]	RoTTA [70]	LAME [4]	UniTTA [12]	DDA [17]	SDA (Ours)
18.3	4.2	6.2	6.3	8.8	9.3	9.8	10.2	10.6	11.5	17.4	23.1	24.9	30.3	29.3	**31.6** (+2.3)

Table 9. **Data stream sensitivity comparison.** We additionally compare SDA with 13 traditional TTA methods on the UniTTA [12] benchmark which contains 12 class/domain balance/imbalance settings. Results are reported as average accuracy across settings.

Component	ResNet-50	Swin-T	ConvNeXt-T
DDA [17]	29.7	40.0	44.2
+ Conditional Data Generation	30.4	40.1	44.6
+ Unconditional Data Alignment	**32.5**	**42.5**	**47.0**

Table 10. **Impact of different components in SDA.** Results are evaluated on ImageNet-C.

Method	Number (N)	ResNet-50	Swin-T	ConvNeXt-T
DDA [17]	0	29.7	40.0	44.2
SDA	1K	31.9	42.3	45.5
	10K	31.9	**42.5**	46.6
	50K (default)	32.5	**42.5**	**47.0**
	100K	**32.6**	42.2	46.8

Table 11. **Impact of different numbers of fine-tuning images.** Results are evaluated on ImageNet-C.

4.4. Ablation Studies

Components. We examine the impact of two key components in our SDA framework's Mix of Diffusion technique in Tab. 10: (1) synthetic data generation using the conditional diffusion model and (2) synthetic data alignment using the unconditional diffusion model. Fine-tuning source models with synthetic data generated solely by the conditional diffusion model (+ Conditional Data Generation) yields only marginal gains. As discussed in Sec. 3.4, this limited gain arises from the domain misalignment between conditional and unconditional diffusion models. Specifically, using only conditional synthetic data results in models aligned with the conditional diffusion domain, whereas the test data belongs to the unconditional diffusion domain. Therefore, further aligning the synthetic data through the unconditional diffusion model (+ Unconditional Data Alignment) leads to significant performance gains over the baseline DDA. This demonstrates that bridging the misalignment of different diffusion domains is essential for the success of our SDA framework.

Number of fine-tuning images. We examine the impact of different numbers of images (N) used during synthetic-domain model fine-tuning in Tab. 11. Interestingly, even with only one image per class (N = 1K), SDA still significantly outperforms DDA. This finding suggests a key attribute of the fine-tuning process: source models are primarily learning to adapt to the synthetic domain itself, rather than acquiring class-specific knowledge. Increasing the number of images helps the fine-tuning process capture the synthetic domain more accurately, thereby enhancing performance. Based on a balance between performance improvement and image generation resources, we select N = 50K as our default experimental setting.

5. Conclusion

In this paper, we proposed **S**ynthetic-**D**omain **A**lignment (**SDA**), a novel test-time adaptation (TTA) framework that simultaneously aligns the domains of the source model and target data with the synthetic domain of a diffusion model. For the source model, SDA introduces a Mix of Diffusion (MoD) technique, which generates synthetic data to adapt the source model to a synthetic-domain model. MoD involves a conditional diffusion model for data generation and an unconditional diffusion model for data alignment. For the target data, SDA utilizes the aforementioned unconditional diffusion model to project the target data to synthetic data. As the domains of the model and data are aligned, SDA converts the cross-domain TTA task into an easier in-domain prediction task. Compared to existing diffusion-driven TTA methods, SDA significantly mitigates the source-synthetic domain misalignment issue. Compared to traditional TTA methods, SDA maintains insensitivity to different data streams. Extensive experiments across classifiers, segmenters, and MLLMs indicate that SDA achieves enhanced domain alignment and superior performance.

Acknowledgements

The work is supported in part by the National Key R&D Program of China under Grant 2024YFB4708200 and National Natural Science Foundation of China under Grant U24B20173.

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

[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

[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

[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

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

[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

[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

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

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

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

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

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

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

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

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

[32] 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

[33] 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

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

[35] 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

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

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

[38] 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

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

[40] 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

[41] 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

[42] 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

[43] 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.

[44] 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

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

[46] 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

[47] 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

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

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

[50] 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

[51] 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

[52] 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

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

[54] 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

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

[56] 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

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

[58] 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

[59] 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

[60] 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

[61] 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

[62] 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

[63] 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

[64] 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

[65] 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

[66] 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

[67] 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

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

[69] 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

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

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