



Evaluating the Applicability of Training-Free Style Transfer for Test-Time Augmentation under Domain Shift

Master Thesis

Master of Science in Applied Computer Science

Noah Ikechukwu Emmler

September 15, 2026

Supervisor:

1st: Prof. Dr. Christian Ledig
2nd: M.Sc Sebastian Dörrich

Chair of Explainable Machine Learning
Faculty of Information Systems and Applied Computer Sciences
Otto-Friedrich-University Bamberg

Abstract

Deep learning classifiers deployed in real-world scenarios frequently suffer performance degradation under domain shift due to mismatches in low-level appearance statistics. Existing test-time adaptation methods often require modifying model parameters or per-instance optimization, limiting applicability to black-box models. Test-time augmentation (TTA) avoids model updates but traditionally relies on geometric or photometric transformations that fail to bridge appearance gaps. This thesis investigates whether training-free style transfer, driven by semantically retrieved source-domain reference images, can serve as an effective input-based TTA strategy. We propose STyle-TTA, a framework operating entirely at inference time. STyle-TTA uses a self-supervised vision encoder (DINOv3) to retrieve semantically compatible source-domain style references in a label-agnostic manner. It then applies a training-free style transfer operator—primarily evaluated via diffusion-based style injection (StyleID) alongside a lightweight baseline (AdaIN)—to generate style-conditioned observations of the target input, aggregating predictions across transformed views without modifying the classifier. We evaluate STyle-TTA across natural image, remote sensing, and computational pathology benchmarks (ImageNet-R, EuroSAT $\rightarrow$ UCMerced, Camelyon17-WILDS, MIDOG2022, and Epithelium-Stroma). Empirical results demonstrate that StyleID-based STyle-TTA yields substantial accuracy gains for convolutional neural networks (CNNs) and standard vision transformers—achieving Top-1 accuracy improvements of up to 12.32 percentage points on ImageNet-R and up to 14.67 percentage points on remote sensing data. However, its effectiveness depends heavily on classifier architecture: while standard CNNs and ViTs benefit significantly, pre-trained Foundation Models (CLIP and DINOv2) possess existing appearance invariance, making geometric TTA the superior strategy.

Evaluation and method code can be found on GitHub:

https://github.com/Ikechukwue/STyle-TTA

Contents

List of Figures iv

List of Tables vi

List of Acronyms ix

1 Introduction 1

2 Theoretical Background 3

2.1 Distribution Shift . . . 3

2.2 Test-Time Augmentation . . . 5

2.3 Image Style Transfer . . . 8

2.4 Diffusion Models for Image Generation . . . 11

3 Related Work 14

3.1 Style-Based and Generative Test-Time Adaptation . . . 14

3.1.1 Style-Transfer-Based TTA . . . 14

3.1.2 Generative TTA . . . 14

3.1.3 Learned Feed-Forward Style-Transfer Networks . . . 15

3.2 Positioning of the Proposed Approach . . . 16

4 Methods 18

4.1 Overview of the Proposed Framework . . . 18

4.2 Style Reference Retrieval . . . 20

4.3 Style Transfer Operators . . . 22

4.4 Prediction Aggregation . . . 24

4.5 Implementation Detail . . . 25

5 Experiments & Results 27

5.1 Datasets . . . 27

5.2 Evaluation Metrics . . . 31

5.3 Classifiers . . . 32

5.3.1 Baseline (No-TTA) Classification Performance . . . 33

5.3.2 Remaining Benchmarks . . . 34

5.4 Comparison of TTA Strategies on ImageNet-R . . . 36

5.5 Additional Benchmarks . . . 38
5.6 Feature-Space Domain Gap . . . 41
5.6.1 Global Discrepancy . . . 41
5.6.2 Per-Class Gap and Performance Correlations . . . 42
5.7 Calibration and Confidence Quality . . . 43
5.8 Ablation Studies . . . 45
5.8.1 Style Reference Retrieval Strategy . . . 46
5.8.2 Number of Generated Views . . . 48
5.8.3 Prediction Aggregation Strategy . . . 50
5.9 Computational Cost . . . 55
5.10 Hybrid Approach . . . 56

6 Discussion **60**

7 Limitations **65**
7.1 Statistical Rigor and Computational Constraints . . . 65
7.2 Scope of Empirical and Mechanistic Claims . . . 65

8 Future Work **66**

9 Conclusion **68**

10 Appendix **70**

Bibliography **76**

List of Figures

1 Illustration of class-space distribution and domain gap. The feature space demonstrates the decision boundary between the target class (triangle) and other classes (circle). Blue and green markers highlight samples affected by domain shift relative to the decision boundary. . . 3

2 Illustration of a view ensemble for TTA classification tasks 6

3 Adapted from Chung et al. (2024). Examples of artistic style transfer, target content images stylized with specific artistic reference images using StyleID. 8

4 **AdaIN Framework.** Adapted from Huang and Belongie (2017). The first few layers of a fixed VGG-19 network are used to encode the content and style images. An AdaIN layer performs style transfer in feature space. A decoder is learned to invert the AdaIN output back into image space. The same VGG encoder is used to compute content and style losses. 10

5 **StyleID Framework** Adapted from Chung et al. (2024). (Left) Style transfer pipeline: content (z_T^c) and style (z_T^s) inverted noise are combined via Initial Latent AdaIN (z_T^{cs}) and refined during reverse diffusion using attention-based style injection and temperature scaling. (Right) Detailed mechanisms for style injection and initial noise alignment. 13

6 Overview of the proposed Style-Based Test-Time Augmentation (STyle-TTA) framework. 19

7 ImageNet-R test image (left) with its top-5 nearest-neighbour picks, from left to right. (All class="scuba diver") 22

8 Image comparison roster between StyleID and AdaIN. Content images are from ImageNet-R and style images from ImageNet-1K. 24

9 Examples of in-domain and out-of-domain images from the different benchmarks. For EuroSAT and ImageNet-1K, the test sets are different datasets (UCMerced/ImageNet-R). The pathology datasets use their respective OOD data. 28

10 Reliability diagram for DenseNet-121 and DINOv2 comparing Geometric TTA ($N = 64$) and Style-TTA ($N = 16$). Accuracy per confidence bin is computed across 15 bins using pooled predictions across seeds, alongside the aggregated Expected Calibration Error (ECE). . 45

11 Stylized outcomes for different retrieval methods. The left image uses a style reference retrieved by DINOv3, whereas the right image is a randomly selected reference. Style source is ImageNet-1K while content source is ImageNet-R 47

12 Top-1 accuracy of individual stylized views of ImageNet-R evaluated independently of the ensemble. 48

13 Effect of the number of created style-augmented STyle-TTA views (N) on TTA performance across representative classifier architectures against standard geometric TTA. Results are shown for DenseNet (CNN), Swin (ViT), and Swin (FM). The black cross denotes the corresponding baseline performance without style augmentation ($N = 1$). StyleID-based STyle-TTA is evaluated up to $N = 16$ the other up to $N = 64$. 49

14 Construction of the View Ensemble under Hybrid Strategies S1 and S2 ($N_{\text{Views}} = 8$). The $T(\cdot)$ either a geometric transformation or a style-based one. 57

15 Top-1 accuracy (%) on ImageNet-R across varying ensemble sizes $N \in \{8, 16, 32, 64\}$ using a fixed budget of three stylized views. The x marks the baseline accuracy relative to ImageNet. 59

16 Qualitative style transfer examples across datasets (Part 1). 72

17 Qualitative style transfer examples across datasets (Part 2). 73

List of Tables

1 Categorization of style-based test-time adaptation methods by source data dependence, model updating requirement, transformation training, and underlying mechanism, highlighting the zero-retraining paradigm of STyle-TTA. 16

2 Overview of the datasets used in our experiments across natural images, remote sensing, and histopathology. For each dataset, we report the number of classes, the number of samples in the training, validation, and test splits, and as a measure of class-distribution balance Shannon's Equitability (Eq). It ranges from 0 to 1, with values closer to 1 indicating a more uniform distribution of samples across classes. 28

3 Construction of the shared label space for the EuroSAT–UCMerced transfer benchmark. The original class taxonomies are mapped to four semantically aligned categories to enable cross-dataset evaluation, while classes without a correspondence are excluded. 29

4 Overview of the computational pathology benchmarks, their domain origins, and experimental splits. 29

5 Overview of evaluation classifiers. 32

6 Baseline (no-TTA) performance, ImageNet-1K (200 shared classes) → ImageNet-R. Reporting Top-1 accuracy, as well as balanced accuracy. The output logit vector $z \in \mathbb{R}^{1000}$ was sliced using a predefined class mask to yield $z_{\text{target}} \in \mathbb{R}^{200}$ prior to Softmax, directly matching the 200 shared target classes. 34

7 Baseline (no-TTA) performance across the four additional benchmark datasets, comparing CNN (DenseNet-121), ViT (Swin-B), and Foundation Model (DINOv2) architectures. We report top-1 accuracy, balanced accuracy, and the change in top-1 accuracy between the source and target domains. 35

8 Evaluation of baseline ImageNet-R and TTA methods across CNN, Vision Transformer, and Foundation Model backbones. Bold indicates the best performance per metric. All Top-1 accuracy improvements were statistically significant based on McNemar's test ($p < 0.001$). 37

9 Evaluation of baseline and TTA methods on Camelyon. Bold indicates highest accuracy. 38

10 Evaluation of baseline and TTA methods on EpiStr. Bold indicates highest accuracy. 39

11 Evaluation of baseline and TTA methods on UCMerced. Bold indicates highest accuracy. 40

12 Evaluation of baseline and TTA on MIDOG. Bold indicates highest accuracy. 40

13 Feature-space discrepancy on ImageNet-1K → ImageNet-R. Net shift calculated as OOD distance minus in-domain baseline. 41

14 Spearman rank correlations (ρ) between per-class domain gap and classification metrics. Baseline Gap: d_c^{val} vs. baseline metrics. Stylized Gap: $d_c^{k,\text{val}}$ (after StyleID) vs. baseline metrics. ΔGap vs. Gain: correlation between per-class Wasserstein reduction and STyle-TTA improvement. Negative values indicate inverse relationships. 42

15 Calibration and confidence-ranking performance of test-time augmentation strategies on ImageNet-1K → ImageNet-R. Each strategy is evaluated at its best-performing view count (STyle-TTA: $N = 16$, Geometric: $N = 64$). Lower ECE indicates better calibration, while higher AUC indicates better confidence-based ranking. Bold marks the best result per backbone and metric. 44

16 Evaluation of source style reference retrieval strategies for STyle-TTA. Performance on ImageNet-R across six ImageNet-1K classifiers comparing Random, Class-Balanced Random, and DINO-based nearest-neighbourretrieval ($N_{\text{views}} = 64$, AdaIN operator). Δ Acc reports the absolute change relative to the unadapted baseline with 95% bootstrap confidence intervals. Bold values indicate optimal performance per backbone. 47

17 Comparison aggregation modes (`vanilla`, `zero`, `tpt`) for StyleID ($N_{\text{refs}} = 16$) on ImageNet-R. 51

18 Ensemble accuracy and oracle upper bound for increasing numbers of stylized views. 53

19 Prediction agreement and pairwise diversity across 15 stylized views. 54

20 Correlation between prediction diversity measures and accuracy gain. 54

21 Per-view latency breakdown and peak memory consumption (ResNet-18, 224×224, NVIDIA RTX A5000). Reporting Augmentation Latency and Evaluation Latency per view, summarized in Total Per-View Latency, as well as peak VRAM Usage at inference. 55

22 Evaluation of baseline ImageNet-R and TTA methods across CNN, Vision Transformer, and Foundation Model backbones. Ensemble size per method was $N = 64$ for STyle-TTA it's $N = 16$. Bold indicates the best performance per metric. 58

A.1 Detailed architectural specifications of the evaluated classifiers. 70

A.2 Hyperparameter settings for fine-tuning and linear probing. 70

A.3 Classifier performance metrics across different view counts for DINO and Balanced Random retrieval methods. 71

A.4 Performance of the AdaIN test-time augmentation strategy ($N = 64$) against the no-TTA baseline on ImageNet-1K $\rightarrow$ ImageNet-R. Reported are balanced accuracy, absolute change from baseline with 95% CI, and significance (paired bootstrap test). 71

A.5 **Impact of augmented view count (N) on TTA performance across backbones.** Evaluation of balanced accuracy (%) on ImageNet-R across varying numbers of generated views ($N \in \{2, 4, 8, 16, 32, 64\}$) for AdaIN, StyleID, and Geometric TTA. The unadapted baseline ($N = 1$) corresponds to single-image inference. $\Delta(N_{\text{min}} \rightarrow N_{\text{max}})$ indicates the absolute percentage point change from the minimal evaluated view count ($N = 2$) to the maximum ($N = 64$ for AdaIN/Geometric; $N = 16$ for StyleID). Bold text denotes the peak balanced accuracy for each backbone. 74

A.6 Calibration (ECE & AUC) of Hybrid approach of STyle-TTA against (Baseline $\rightarrow$ TTA), ImageNet $\rightarrow$ ImageNet-R. Ensemble size for all methods $N = 64$ except regular STyle-TTA $N = 16$ 75

List of Acronyms

AdaIN	Adaptive Instance Normalization
AUC	Area Under the ROC Curve (AUC)
BA / BAcc	Balanced Accuracy
CI	Confidence Interval
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
DDIM	Denoising Diffusion Implicit Models
DINOv2/v3	Self-supervised Vision Transformer (Meta AI)
ECE	Expected Calibration Error
FM	Foundation Model
H&E	Hematoxylin and Eosin (Staining)
HPC	high-performance computing
IHC	Immunohistochemistry
LDM	Latent Diffusion Model
LP	Linear Probe / Linear Probing
OOD	Out-Of-Distribution
pp	Percentage Points
SGD	Stochastic Gradient Descent
TPT	Test-Time Prompt Tuning
TTA	Test-Time Augmentation
TTIA	Test-Time Instance Adaptation
UMAP	Uniform Manifold Approximation and Projection
ViT	Vision Transformer
WILDS	Benchmark for in-the-Wild Distribution Shifts

Notation

The following notation is used throughout this thesis.

Domains and Datasets

$\mathcal{D}_S$	Source domain dataset
$\mathcal{D}_T$	Target domain dataset
$\mathcal{D}_S^{\mathcal{C}}$	Source domain restricted to the shared classes
$\mathcal{D}_T^{\mathcal{C}}$	Target domain restricted to the shared classes
$\mathcal{C}_S$	Set of source-domain classes
$\mathcal{C}_T$	Set of target-domain classes
$\mathcal{C}$	Shared class set, $\mathcal{C}_S \cap \mathcal{C}_T$
x	Input image
y	Ground-truth class label
$\hat{y}$	Predicted class label
N_S	Number of source-domain samples
N_T	Number of target-domain samples

Models and Feature Representations

f_θ	Classifier with parameters θ
$\phi(x)$	Feature embedding of image x
z	Feature representation of an image ($z = \phi(x)$)
$\mathbb{R}^d$	d-dimensional feature space
c	Class index
$f_\theta(x)_c$	Predicted probability for class c

Style Transfer and Test-Time Augmentation

$\mathcal{S}$	Style reference pool
s	Style reference image
$T(x, s)$	Style transfer operator applied to image x using style image s
$\tilde{x}$	Stylized image
$\mathcal{A}(x)$	Set of augmented views generated from image x
N	Number of generated views
a_i	Augmentation operator for view i
$p(a)$	Distribution over augmentation operators

Domain Shift Analysis

$F_l(x)$	Feature activation tensor at layer l
$G_l(x)$	Gram matrix computed from layer l
d_{Gram}	Gram matrix distance between two images
$\|\cdot\|_F$	Frobenius norm

Mathematical Operators

$\mathbb{E}$	Expectation
$\arg\max$	Argument of the maximum value
$\sim$	"Distributed according to"
$\|\cdot\|_2$	Euclidean norm
$\|\cdot\|_F$	Frobenius norm
$\top$	Matrix transpose

1 Introduction

Deep learning has enabled substantial advances in visual recognition through the learning of hierarchical representations from data LeCun et al. (2015). In practice, however, deployed classifiers frequently encounter *domain shift*: the statistical properties of the target data differ from those observed during training, causing performance to degrade substantially (Zhou et al., 2022). This problem arises across a wide range of applications, from natural images captured under varying weather and lighting conditions to medical images acquired with different scanners or protocols, and to satellite imagery collected by sensors with divergent spectral characteristics.

A large body of research has addressed domain shift through model-level adaptation, where the classifier's internal weights or their statistics are updated after observing target-domain samples. While effective, these adaptation methods require access to model parameters and often involve per-instance optimization during inference. This limits their applicability when the classifier is deployed as a frozen or black-box model whose weights cannot be modified after training.

An alternative strategy is *test-time augmentation* (TTA), which preserves the classifier parameters and instead improves robustness by aggregating predictions from multiple transformed views of each test image. Existing TTA approaches predominantly rely on manually designed transformations such as geometric crops, flips, and photometric perturbations. Although these augmentations increase prediction diversity, they do not explicitly align the appearance of target inputs with the source-domain characteristics the classifier was trained on, leaving a substantial gap when source and target differ in low-level style statistics.

A complementary line of research has explored style manipulation as a mechanism for domain robustness. Methods in this space range from closed-form feature-statistic matching and learned feed-forward stylization networks to deep generative models that resynthesize inputs under a source-domain prior. What unites these approaches is that the style-transfer mechanism itself is fitted specifically to the source domain, whether through calibration, dedicated training, or generative modeling and its benefits cannot be transferred to a new classifier without repeating that source-specific step.

This thesis investigates whether training-free style transfer can serve as a general test-time augmentation strategy. We find that its effectiveness is fundamentally conditioned on the target classifier's sensitivity to low-level appearance statistics. The proposed *Style-Based Test-Time Augmentation* (STyle-TTA) framework operates entirely at inference time: it retrieves semantically compatible source-domain images based on visual similarity in a pre-trained embedding space, uses them as style references to generate source-like observations of the target input via a training-free style-transfer operator, and aggregates the classifier's predictions over these augmented views. The framework requires no modification of the classifier, no target-domain labels, and no training or fine-tuning of the style-transfer mechanism itself.

The central research question addressed in this work is therefore:

Can training-free style transfer, driven by semantically retrieved source-domain references, serve as an effective test-time augmentation strategy for improving image classification under domain shift?

To answer this question, the thesis makes the following contributions:

1. **A novel test-time augmentation framework.** We propose STyle-TTA, which generates style-conditioned views of target inputs using training-free style transfer and aggregates their predictions to improve robustness under domain shift without modifying the classifier.
2. **A label-agnostic retrieval strategy.** We introduce a retrieval mechanism based on a pre-trained self-supervised vision encoder (DINOv3) that selects source-domain style references by semantic visual similarity rather than class label, allowing style information to be drawn from semantically related but potentially different classes.
3. **An empirical evaluation across diverse domains.** We evaluate STyle-TTA on natural image benchmarks (ImageNet-1K and ImageNet-R) as well as domain-specific datasets from medical imaging and satellite imagery.

2 Theoretical Background

2.1 Distribution Shift

Distribution shift refers to the situation in which the distribution of data encountered at deployment differs from that observed during training. Machine learning models are typically optimized under the assumption that training and test samples are drawn from the same underlying distribution, an assumption that is frequently violated in real-world applications (Quiñonero-Candela et al., 2008). It can arise from changes in the objects, environments, acquisition conditions, or visual characteristics represented in the data. At the level of the object itself, shift can arise from attributes the classifier was not adequately exposed to during training as illustrated in Figure 1.

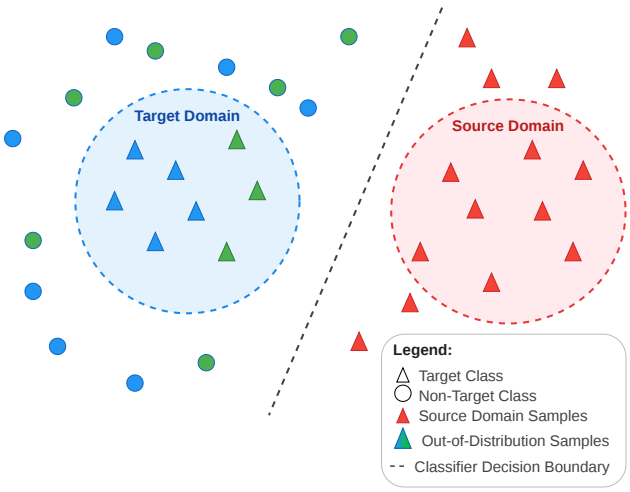


Figure 1: Illustration of class-space distribution and domain gap. The feature space demonstrates the decision boundary between the target class (triangle) and other classes (circle). Blue and green markers highlight samples affected by domain shift relative to the decision boundary.

A well-documented example is reduced detection accuracy for pedestrians with darker skin tones in object detection systems trained on unbalanced data (Wilson et al., 2019). At a more nuanced level, autonomous driving systems must recognize roads and pedestrians consistently across weather conditions, e.g. sunny, foggy, or rainy, despite substantial low-level appearance changes (Chen et al., 2017). In the medical domain, imaging hardware differs across hospitals and scanners, introducing systematic differences in noise, color tone, or texture even where the underlying anatomical structure is identical (Koh et al., 2021; Aubreville et al., 2024; Li et al., 2022a). Similarly, satellite imagery classifiers encounter severe domain shifts driven by temporal, geographic, and environmental variations—such as seasonal land cover changes, architectural differences across regions, or sensor and atmospheric interference (Koh et al., 2021). In many of these cases, the task-relevant semantics are intended to remain stable, while nuisance factors such as texture, color statistics, illumination, or sensor noise vary across domains. A robust classifier should therefore distinguish task-relevant information from source-domain-specific appearance characteristics. This distinction is particularly relevant for image classifiers that rely heavily on texture-based cues. (Geirhos et al., 2022) demonstrated that ImageNet-trained convolutional neural networks ex-

hibit a strong texture bias, often relying on local texture information rather than global object shape for recognition. Their findings suggest that changes in texture and appearance can substantially influence classification decisions even when the underlying object identity remains unchanged.

This gap becomes most directly observable in the feature space of a trained classifier, typically inspected at the penultimate (pre-logit) layer. Out-of-distribution samples may be represented less consistently with the learned source-domain structure, resulting in increased class overlap and degraded predictive performance (Hendrycks and Gimpel, 2017; Xie et al., 2023). The extent to which this occurs depends not only on the magnitude of the distribution shift, but also on the features learned by the model and on the relationship between the shifted attributes and the prediction task.

Formally, following Zhou et al. (2022), a *domain* $\mathcal{D}$ is defined as a pair consisting of a feature space $\mathcal{X}$ and a marginal probability distribution over that space, $\mathcal{D} = \{\mathcal{X}, P(X)\}$, where $X \in \mathcal{X}$. A *task* $\mathcal{T}$ is defined by a label space $\mathcal{Y}$ and a conditional distribution $P(Y \mid X)$ that the classifier is trained to approximate, $\mathcal{T} = \{\mathcal{Y}, P(Y \mid X)\}$.

A classifier f_θ is trained on samples drawn from a source domain $\mathcal{D}_S = \{\mathcal{X}, P_S(X)\}$ under task $\mathcal{T}_S$, and is deployed on a target domain $\mathcal{D}_T = \{\mathcal{X}, P_T(X)\}$.

Domain shift (or *distribution shift*) occurs when

$$P_S(X) \neq P_T(X), \tag{1}$$

while the label space is shared, $\mathcal{Y}_S = \mathcal{Y}_T$. Depending on which components of the joint distribution change, different forms of distribution shift can be distinguished. In *label shift*, the label distribution changes while the class-conditional distribution is assumed to remain fixed, such that $P_S(Y) \neq P_T(Y)$ and $P_S(X \mid Y) = P_T(X \mid Y)$ (Saerens et al., 2002; Lipton et al., 2018). In *concept shift*, the conditional relationship between inputs and labels changes, meaning that $P_S(Y \mid X) \neq P_T(Y \mid X)$ (Lyu et al., 2026).

This thesis is concerned primarily with the setting in which the conditional distribution is assumed to remain invariant,

$$P_S(Y \mid X) = P_T(Y \mid X), \tag{2}$$

while the marginal input distribution changes. This setting is commonly referred to as *covariate shift* (Shimodaira, 2000; Zhou et al., 2022). In the context of this thesis, the relevant shifts are particularly associated with changes in image appearance, including texture, color, contrast, illumination, and other acquisition-related characteristics. The underlying object identity or task-relevant semantics are assumed to remain unchanged, while the visual realization of the input may differ between source and target domains.

Bridging this gap, i.e. improving generalization from $\mathcal{D}_S$ to $\mathcal{D}_T$, can be approached at several points in the pipeline. The most direct approach is to adapt the model itself, e.g. via retraining or fine-tuning on target-domain data (Yosinski et al., 2014). This is in many cases impractical. Going back to the previous medical applications, target-domain data are often scarce or highly homogeneous. Similarly, time constraints may preclude retraining or the model weights may not be accessible at all (e.g. behind an inference API). An alternative family of methods instead adapts the model at test time, without full retraining (Liang et al., 2024). If the model itself cannot be modified, a further alternative is to adapt what the model is shown instead. This is done by transforming the input at test time so that it more closely resembles the distribution the classifier was trained on.

2.2 Test-Time Augmentation

Test-Time Augmentation (TTA) denotes a family of methods that tries to improve a classifier's predictive performance by modifying the inference process itself, rather than the model's parameters. '*Model-based*' Adaptation methods update internal model statistics at inference time. These methods achieve this by recalibrating batch-normalization statistics on the test batch (Nado et al., 2020; Schneider et al., 2020) and require access to the model's internals. '*Input-based*' TTA methods meanwhile leave the model untouched and instead focus on the test-time input samples. By constructing multiple transformed views of the test instance itself, for example through geometric transformations or color-space perturbations, and aggregating the resulting predictions into a single output, model sensitivity to any single view is smoothed out (Shanmugam et al., 2021). An early instance of this idea is ten-crop testing, which averages predictions over ten spatially cropped and mirrored versions of a test image (He et al., 2016). Test-time augmentation does not by itself model or correct for distribution shift, but has been shown to be a useful component within broader pipelines, for instance by providing the unsupervised training signal that model-based methods optimize against (Liang et al., 2024). This thesis follows the input-based paradigm, since it does not require access to model weights or architecture and is therefore applicable to black-box classifiers. In the taxonomy of Liang et al. (2024), the setting considered in this thesis corresponds to *test-time instance adaptation* (TTIA): a special case of test-time batch adaptation in which a single unlabeled target instance x_t is processed at a time, as opposed to a mini-batch or the entire target domain.

Ensemble Formulation Given a test instance x and a classifier f_θ, an input-based TTA procedure generates a set of n augmented views $\{\tilde{x}_1, \ldots, \tilde{x}_n\}$ via a transformation family $\mathcal{T}$, where $\tilde{x}_k = t_k(x)$ and $t_k \in \mathcal{T}$. Conventionally, the original, unaugmented instance is included as $\tilde{x}_0 = x$ (Shanmugam et al., 2021)(Figure 2).

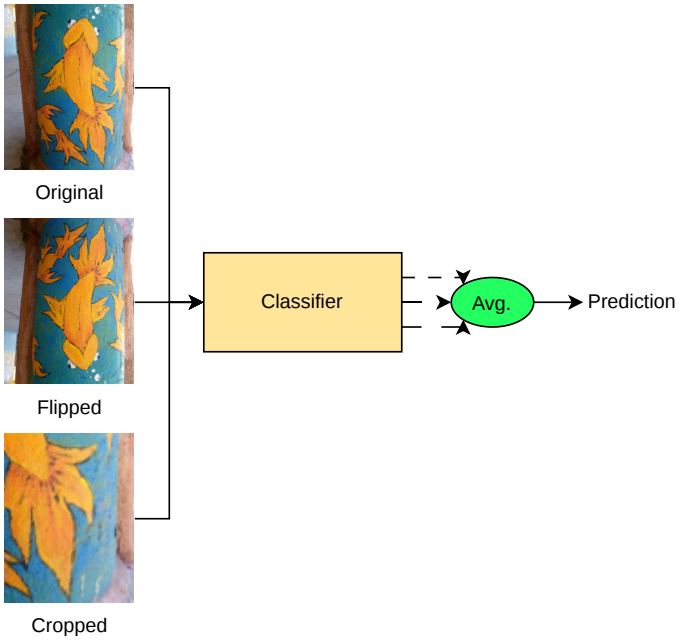


Figure 2: Illustration of a view ensemble for TTA classification tasks

Each view is passed through the classifier independently to obtain a prediction distribution $p_k = f_\theta(\tilde{x}_k) \in \mathbb{R}^{|\mathcal{Y}|}$. The final prediction $\hat{y}$ is obtained by aggregating the $n+1$ per-view predictions, most commonly via averaging (soft aggregation). MEMO, for instance, optimizes model confidence at test time by minimizing the entropy of exactly this averaged prediction over multiple randomly augmented views of a single input (Zhang et al., 2022), and related methods instead average the augmented predictions against a fixed prediction under an entropy-thresholding scheme to avoid diluting confident base predictions with noisy views (Sivaprasad and Fleuret, 2021).

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \frac{1}{n+1} \sum_{k=0}^{n} p_k(y), \tag{3}$$

or via majority vote over the per-view hard predictions $\hat{y}_k = \arg\max_y p_k(y)$ (hard aggregation) (Shanmugam et al., 2021),

$$\hat{y} = \text{mode}\big(\hat{y}_0, \ldots, \hat{y}_n\big). \tag{4}$$

Confidence-Based Aggregation Rather than assigning equal weight to all augmented views, aggregation can also be restricted to predictions from views on which the classifier is most confident. The motivation is that some transformations may produce ambiguous or less informative views whose predictions could adversely affect the final aggregation. Two confidence-based approaches considered in this thesis are Test-Time Prompt Tuning (TPT) (Shu et al., 2022) and ZERO (Farina et al., 2024). Both use prediction confidence to select a subset of augmented views before aggregation, but differ in how the selected predictions are combined. TPT

uses an entropy-based selection mechanism followed by soft probability aggregation, whereas ZERO uses the selected predictions for hard aggregation. Since the original TPT method performs test-time prompt optimization, the implementation used in this thesis adapts its confidence-based view-selection principle to conventional classifiers without updating model parameters. The precise implementation of these aggregation strategies is described in Sec. 5.8.3.

The use of an ensemble of views is intended to reduce the variance of the model's decision under transformations that are not expected to change the ground-truth label. If the classifier's learned decision boundary is well-aligned with the true class-conditional distribution $P(Y \mid X)$, predictions across label-preserving views of the same instance should agree, and aggregation should converge toward the correct class even where an individual view is misclassified. This relies on the assumption that classification errors are more often associated with low-confidence, "borderline" predictions than with high-confidence, systematically wrong ones (Guo et al., 2017). Under this assumption, creating varying perturbations of the input via augmentation shifts a low-confidence, borderline prediction either further toward the correct class or further away from it in a way that is on average corrected by ensembling, so that aggregating over several such perturbations increases the net probability mass on the correct label. Shanmugam et al. (2021) observed that the degree of divergence between views is itself predictive of the resulting gain: augmentations that are more distinct from one another and from the original instance tend to produce a larger improvement upon aggregation than near-duplicate views.

Standard (Geometric) TTA The conventional instantiation of input-based TTA applies geometric transformations to the test instance, such as random cropping, horizontal flipping, or resizing (He et al., 2016; Shorten and Khoshgoftaar, 2019). These transformations are label-preserving under the assumption that the object's identity does not depend on its position, scale, or orientation within the frame. Geometric TTA is most effective in exactly the setting where this assumption is only partially satisfied by the classifier itself: if a model were perfectly invariant to these transformations by construction, all views would yield identical predictions and aggregation would provide no benefit. In practice, convolutional and transformer-based architectures exhibit only approximate spatial invariance (Zhang, 2019; Bhojanapalli et al., 2021), which geometric TTA exploits by sampling several points in this remaining sensitivity and averaging over them.

Limits of Geometric TTA under Style-Based Shift Geometric transformations address a specific class of nuisance variation: position, scale, and orientation. These variations primarily affect the spatial layout of the target object's features and direct surroundings but leave the low-level appearance statistics of the instance (texture, color, local contrast) unchanged. As established previously, not all distribution shifts are geometric in nature: shifts arising from sensor characteristics, imaging conditions, or stylistic domain differences manifest primarily as changes in texture and appearance statistics rather than in object pose or framing. For such



Figure 3: Adapted from Chung et al. (2024). Examples of artistic style transfer, target content images stylized with specific artistic reference images using StyleID.

shifts, geometric augmentation offers no direct mechanism for closing the domain gap, since none of its transformations alter the appearance-level properties responsible for the shift (Geirhos et al., 2022). This motivates the style-based augmentation approaches.

2.3 Image Style Transfer

Style transfer refers to the process of generating an image that preserves the semantic content or spatial structure of a *content image* while adopting the visual appearance of a separate *style image* (Figure 3). The notion of style is deliberately broad: it can encompass characteristics such as color distribution, texture, brushwork, contrast, and other recurring visual patterns, while content generally refers to the objects, structures, and spatial relationships represented in the image. The central challenge of style transfer is therefore to modify appearance without substantially altering the underlying content. Although this distinction appears intuitive from a human perceptual perspective, separating content from style is not straightforward in pixel space, since both are entangled within the same image representation. Consequently, the development of style-transfer methods has largely been driven by different strategies for representing, disentangling, and manipulating these two factors. Prior to the use of deep neural networks, style transfer was predominantly approached through low-level image statistics, such as global color-histogram matching (Reinhard et al., 2001) or manipulation of frequency components under a Fourier decomposition, under the assumption that an image's coarse structure (content) is encoded in low-frequency components while fine appearance detail (style) is encoded in higher-frequency components (Oliva et al., 2006). A parallel line of work approached appearance transfer through texture synthesis, modeling texture

as a stationary stochastic process characterized by summary statistics of filter responses, which could then be resynthesized onto a target structure (Heeger and Bergen, 1995; Efros and Leung, 1999). These methods were comparatively brittle: they often failed to capture texture at a perceptually convincing level and could not reliably disentangle content from style, since frequency-based decomposition is only a coarse proxy for the semantic and appearance factors an observer actually perceives.

Neural Style Transfer Gatys et al. (2016) showed that deep convolutional networks trained for image classification, specifically VGG-19, implicitly learn representations in which content and style information can be separated. Neural networks construct hierarchical abstractions of input data across their layers. They found that content is captured by the feature activations of deeper convolutional layers, which preserve spatial/semantic structure, while style is captured by the correlations between feature channels within a layer, computed via the *Gram matrix*

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{jk}, \tag{5}$$

where $F^l \in \mathbb{R}^{C_l \times M_l}$ denotes the vectorized feature activations of layer l, with C_l channels and M_l spatial locations. The Gram matrix G^l discards spatial layout entirely and retains only the pairwise channel correlations, which Gatys et al. showed to correlate strongly with perceived texture and style, independent of object position. They synthesized a stylized image by optimizing a random noise image, or the content image itself, to jointly minimize a content loss (distance between deep-layer activations of the output and the content image) and a style loss (distance between Gram matrices of the output and a style reference image, summed across several layers) via gradient descent. This iterative, per-image optimization made the original formulation computationally expensive, motivating a wave of follow-up work. Many other frameworks made use of these findings, such as feed-forward networks trained to approximate the optimization result in a single pass (Johnson et al., 2016), instance normalization as a lightweight mechanism for style-consistent feed-forward stylization (Ulyanov et al., 2016), and adaptive instance normalization (AdaIN), which aligns the channel-wise mean and variance of content features to those of a style reference to enable arbitrary, real-time style transfer without per-style retraining (Huang and Belongie, 2017).

Feed-Forward and Statistic-Matching Methods Subsequent methods replaced the per-image optimization with a feed-forward network trained in advance to approximate the same objective in a single pass, substantially reducing inference cost but typically requiring one trained network per style or per small set of styles (Johnson et al., 2016; Ulyanov et al., 2016).

Huang and Belongie (2017) introduced Adaptive Instance Normalization (AdaIN), which enables arbitrary style transfer, i.e. transfer to a previously unseen style image

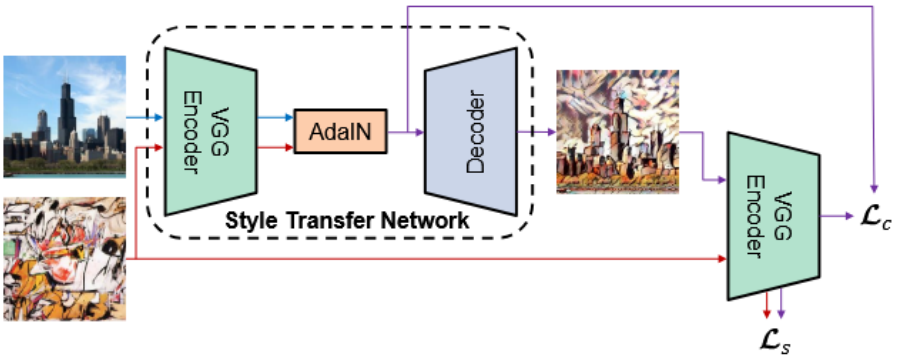


Figure 4: **AdaIN Framework.** Adapted from Huang and Belongie (2017). The first few layers of a fixed VGG-19 network are used to encode the content and style images. An AdaIN layer performs style transfer in feature space. A decoder is learned to invert the AdaIN output back into image space. The same VGG encoder is used to compute content and style losses.

at inference time, without retraining by directly aligning the channel-wise feature statistics of a content encoding to those of a style encoding. Given content and style feature maps F_c and F_s extracted from the same pre-trained encoder. AdaIN is defined as

$$\text{AdaIN}(F_c, F_s) = \sigma(F_s) \left(\frac{F_c - \mu(F_c)}{\sigma(F_c)} \right) + \mu(F_s), \tag{6}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the channel-wise mean and standard deviation, computed independently per channel and per instance. The content feature map is first normalized to remove its own style statistics, then re-scaled and re-shifted to match the style reference's statistics. Style is transferred as a shift in per-channel feature mean and variance, without altering the underlying spatial (content) structure encoded by the normalized activations. The stylized output is subsequently produced by a decoder trained to invert the encoder given the AdaIN-transformed features (Figure 4). To account for higher-order correlations beyond first- and second-order marginal statistics, Li et al. (2017) introduced Universal Style Transfer via Feature Transforms (WCT). Instead of matching scalar means and variances, WCT applies a Whitening and Coloring Transform to align full feature covariance matrices between content and style representations. Both AdaIN and WCT demonstrate that style can be efficiently transferred in a training-free manner by manipulating feature statistics.

GAN- and Diffusion-Based Approaches In parallel, generative adversarial networks (GANs) were applied to style and domain translation, typically learning a mapping between two image domains from unpaired data (Heusel et al., 2017). By training a generator to synthesize realistic images and a discriminator to distinguish them from real data, GANs learn complex domain distributions. While achieving

state-of-the-art visual quality, their application is constrained by the initial training setup. Similarly to early feed-forward methods, most GAN-based approaches require training on a fixed style or domain pair and do not generalize to arbitrary, previously unseen style references without retraining (Zhu et al., 2017). Like most parametric methods, GAN-based approaches are limited by the need for representative target-domain training data. More recently, diffusion-based generative models have been adapted for style transfer, offering training-free style transfer via manipulation of the image generation process itself rather than a dedicated feed-forward stylization network.

2.4 Diffusion Models for Image Generation

Diffusion models have recently emerged as one of the dominant paradigms for high-fidelity image generation (Sohl-Dickstein et al., 2015; Ho et al., 2020). Unlike GANs, which learn to generate images directly through adversarial training, diffusion models formulate image synthesis as a gradual denoising process. Beginning from random noise, the model iteratively reconstructs a coherent image by reversing a stochastic corruption process learned during training. This iterative refinement has proven highly effective for producing realistic images while maintaining stable optimization and broad generative capability.

Training consists of learning the inverse of a predefined forward diffusion process. During the forward process, Gaussian noise is progressively added to an image over a sequence of timesteps until the original signal becomes indistinguishable from pure noise (Ho et al., 2020),

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{7}$$

where β_t denotes the variance schedule controlling the amount of injected noise at timestep t. After sufficiently many diffusion steps, the resulting distribution approaches an isotropic Gaussian. The reverse process is parameterized by a neural network, typically a U-Net, which predicts the noise component present at each timestep. Starting from Gaussian noise, the model repeatedly removes the predicted noise, gradually reconstructing an image sampled from the learned data distribution. Rather than predicting the image directly, modern implementations predict the added noise, $epsilon_\theta(x_t, t)$, which has been shown to simplify optimization and improve sample quality (Ho et al., 2020). Although early diffusion models operated directly in pixel space, contemporary systems commonly employ latent diffusion models (LDMs) (Rombach et al., 2022). Instead of diffusing high-dimensional images, an encoder first compresses an image into a lower-dimensional latent representation. The diffusion process is then performed within this latent space before a decoder reconstructs the final image. This significantly reduces computational cost while preserving image quality, enabling high-resolution image synthesis such as that used in Stable Diffusion (Rombach et al., 2022). The ability of large pre-trained diffusion models to generate realistic images while conditioning on auxiliary information makes them particularly well suited for image editing tasks. Rather

than synthesizing images from scratch, diffusion models can modify existing images while preserving their semantic content (Meng et al., 2022). By conditioning the denoising process on textual prompts, reference images, or learned feature representations, these models can selectively alter appearance attributes—including illumination, texture, color palette, artistic style, or weather conditions—while maintaining object identity and scene geometry. This capability has made diffusion models an attractive foundation for style-transfer methods aimed at reducing domain shift. However, many existing approaches either rely on computationally expensive optimization procedures or struggle to simultaneously preserve the semantic structure of the content image while accurately transferring stylistic characteristics. These limitations are addressed by StyleID through a training-free adaptation of a pre-trained latent diffusion model.

StyleID Attention-based control mechanisms have emerged as a dominant paradigm for training-free image editing in latent diffusion models. Early approaches demonstrated that structural content can be preserved during generation by fixing or injecting cross-attention and self-attention maps across diffusion trajectories (Hertz et al., 2023; Meng et al., 2022). Building directly on DDIM inversion (Song et al., 2021), StyleID (Chung et al., 2024) adapts these principles specifically for arbitrary style transfer without optimization or model fine-tuning. DDIM provides a deterministic sampling formulation that enables the inversion of an image into a corresponding latent trajectory under the diffusion process. This property is important for image editing methods such as StyleID, where the latent representations of a content image and a style reference are used to guide the generation of a stylized output.

Rather than updating model parameters or learning additional networks, StyleID manipulates the internal self-attention representations of the diffusion model during the reverse denoising process. The method first performs DDIM inversion on both the content and style images to extract their intermediate latent trajectories. During image generation, the Key (K) and Value (V) representations within the self-attention layers are replaced by those extracted from the style image, while the Query (Q) representation is largely retained from the content image (Figure 5).

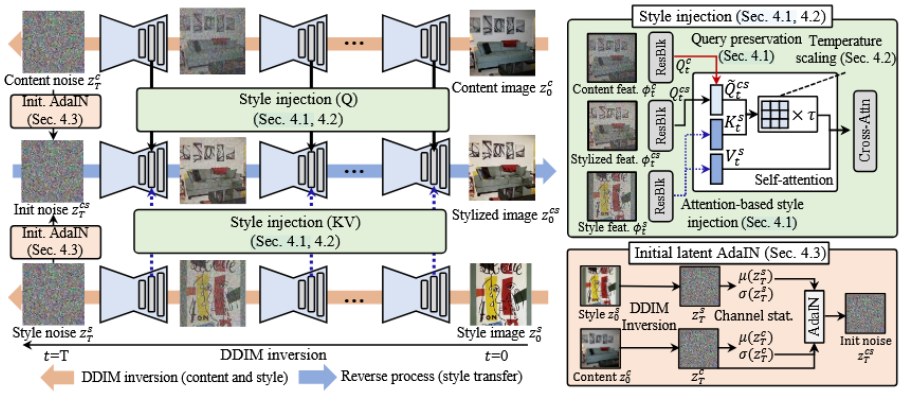


Figure 5: **StyleID Framework** Adapted from Chung et al. (2024). (Left) Style transfer pipeline: content (z_T^c) and style (z_T^s) inverted noise are combined via Initial Latent AdaIN (z_T^{cs}) and refined during reverse diffusion using attention-based style injection and temperature scaling. (Right) Detailed mechanisms for style injection and initial noise alignment.

Since keys and values determine which visual characteristics are retrieved during attention, whereas queries determine where these characteristics are applied (Hertz et al., 2023), this spatial-decoupling mechanism enables stylistic attributes—such as texture and local appearance statistics—to be transferred while strictly preserving the spatial layout of the original content. From the perspective of domain generalization, StyleID can be interpreted as a feature-preserving input transformation. The semantic content of the image is largely maintained through query preservation, while appearance-related characteristics are modified through style injection. Consequently, StyleID provides a mechanism for moving samples between visual domains without requiring changes to the downstream classifier. This property makes it particularly suitable as the style-transfer operator employed in the test-time augmentation framework investigated in this thesis.

3 Related Work

3.1 Style-Based and Generative Test-Time Adaptation

A growing line of work makes target inputs (or intermediate representations) resemble the source domain at test time, rather than relying solely on model-parameter updates. These methods span a spectrum of model capacity and training cost: closed-form or lightweight operators that match low-level style statistics directly; learned, feed-forward stylization networks trained once on source data; and deep generative models that resynthesize the input under a learned source-domain prior. All three are motivated by the same style-augmentation intuition established during training by Style Augmentation (Jackson et al., 2019), Stylized-ImageNet (Geirhos et al., 2022), and MixStyle (Zhou et al., 2021), which perturb or mix feature statistics across instances to prevent classifiers from over-fitting to domain-specific textures.

3.1.1 Style-Transfer-Based TTA

A first family of methods aligns feature- or pixel-level style statistics between target and source *without* learning a full generative model of the source distribution. SFIT (Hou and Zheng, 2021) matches Gram-matrix-based style statistics between a frozen source model's activations and those of a trainable image translator, combined with entropy minimization to preserve semantic content. LODS (Li et al., 2022b) suppresses domain-specific style cues within the network's own representations, while GDA (Zhou et al., 2022) relies on batch-normalization-based style matching together with feature- and phase-level consistency constraints.

A related sub-family performs style adaptation in the frequency domain, motivated by the observation that low-level style statistics are largely encoded in the amplitude spectrum of an image's Fourier decomposition, while structural content is preserved in its phase (Yang and Soatto, 2020). OST (Termöhlen et al., 2021) maps target inputs onto the source domain's style manifold via Fourier-based style transfer as a pre-processing step ahead of the frozen source network. TAF-Cal (Zhao et al., 2022) extends this idea by calibrating target images using the average amplitude spectrum computed over the training distribution, aligning target and source style statistics at both training and test time.

What unites this family is that the transformation is a closed-form or lightweight statistic-matching operation: no image is fully resynthesized, and the mechanism does not model the source data distribution beyond its low-order statistics.

3.1.2 Generative TTA

A second family instead uses a generative model, typically trained on source data, to project the target input back onto the source manifold, treating adaptation as conditional resynthesis rather than statistic matching. DDA (Gao et al., 2023)

projects test inputs toward the source domain using a diffusion model trained only on source data, combining image guidance with self-ensembling to control how strongly each input is adapted; both the diffusion model and the downstream classifier remain frozen across all target domains. Diffusion-TTA (Prabhudesai et al., 2023) instead adapts the discriminative model itself, using gradient feedback from a generative diffusion model to update the classifier at test time on a per-instance basis. Related purification-style approaches such as DiffPure (Nie et al., 2022), originally proposed for adversarial robustness, use the same denoising-diffusion mechanism to remove input-level perturbations before classification, and have since been adopted as a baseline in diffusion-driven TTA comparisons (Gao et al., 2023). More recent work such as Synthetic-Domain Alignment (Guo et al., 2024) highlights a limitation of this family: the synthetic images produced by the diffusion model are not perfectly aligned with the true source distribution, motivating additional fine-tuning of the source classifier on synthetic data.

3.1.3 Learned Feed-Forward Style-Transfer Networks

Between closed-form statistic matching (Section 3.1.1) and deep generative modeling of the source distribution (Section 3.1.2) lies a family of methods that train a dedicated, feed-forward stylization network on source data and apply it, unchanged, as a test-time transform. Adaptive Instance Normalization (Huang and Belongie, 2017) established this paradigm: an encoder-decoder network re-normalizes content features with the channel-wise mean and variance of a style image, enabling arbitrary, real-time style transfer without a style-specific decoder or per-pair training. Because the transform is cheap and architecture-agnostic, it has since been adopted directly as a test-time mechanism for aligning target inputs with the source style distribution, rather than only as a training-time augmentation.

Several works apply this idea explicitly at inference time. Test-Time Style Shifting (Park et al., 2023) uses AdaIN to shift the style of each test sample toward the source style space at test time, requiring no backpropagation or model update, and reports competitive accuracy. S^3-TTA (Xie et al., 2024) combines scale augmentation with AdaIN-based style transfer inside a test-time augmentation pipeline for biomedical segmentation, using a rotational-consistency criterion to select the best-stylized version of each input at inference. T3s (Meng et al., 2025) extends the same underlying mechanism with a bidirectional mapping into a learned style representation space and a diversity-enhancing module, dynamically adjusting test-sample style to align with source-domain features for cross-organ generalization.

Stylizing ViT (Doerrich et al., 2026) applies the same paradigm using a Vision Transformerencoder with weight-shared attention blocks to fuse anatomical structure with target style features. A frozen VGG19 encoder supervises the content and style losses during training. Unlike the three methods above, which are designed and evaluated specifically as test-time procedures, Stylizing ViT is trained primarily as a domain-generalization augmentation method and only subsequently repurposed for test-time augmentation, where the same network is used to align an unseen input's

style with the training distribution at inference, yielding a further accuracy gain over training-time augmentation alone.

3.2 Positioning of the Proposed Approach

Recent approaches for improving robustness under domain shift can broadly be divided into three categories: model adaptation (Nado et al., 2020; Schneider et al., 2020), input-level augmentation (Shanmugam et al., 2021; Shorten and Khoshgoftaar, 2019), and style-based domain robustness (Hou and Zheng, 2021; Gao et al., 2023). The proposed Style-Based Test-Time Augmentation (STyle-TTA) framework occupies a different point on this spectrum: rather than fitting a style-transfer mechanism to the source domain, it applies a training-free, off-the-shelf style-transfer operator (i.e., a pre-trained diffusion model operating as a transformation function). This operator is trained on neither the source nor the target domain, which raises the question of whether it can still stand in for the source-specific mechanisms above. Instead of updating model parameters or generating training data, STyle-TTA modifies only the input samples presented to a frozen classifier. Source-domain images are retrieved based on semantic similarity and used as style references for this off-the-shelf style-transfer operator, producing multiple source-like observations of the same target-domain sample. Predictions from these observations are then aggregated to obtain the final classification result. Note that this retrieval step still draws on source-domain data at inference time. What STyle-TTA removes is not access to source images, but any training, fine-tuning, or calibration of the style-transfer mechanism itself against that data. Table 1 provides a structural summary of this.

Table 1: Categorization of style-based test-time adaptation methods by source data dependence, model updating requirement, transformation training, and underlying mechanism, highlighting the zero-retraining paradigm of STyle-TTA.

Method	**Source images at test**	**Model update**	**Source-trained transform**	**Approach**
SFIT (Hou and Zheng, 2021)	No	No/Yes	Yes	Style matching
OST (Termöhlen et al., 2021)	No	No	No	Fourier transfer
TAF-Cal (Zhao et al., 2022)	No	No	Calibration	Fourier statistics
DDA (Gao et al., 2023)	No	No	Yes	Diffusion
Diffusion-TTA (Prabhudesai et al., 2023)	No	Yes	Yes	Diffusion
Test-Time Style Shifting (Park et al., 2023)	No	No	Yes	AdaIN
S^3-TTA (Xie et al., 2024)	No	No	Yes	AdaIN + TTA
T3s (Meng et al., 2025)	No	No	Yes	Learned style mapping
Stylizing ViT (Doerrich et al., 2026)	No	No	Yes	Learned style transfer
STyle-TTA (ours)	**Yes**	**No**	**No**	**Off-the-shelf transfer**

This formulation positions STyle-TTA as a source-assisted test-time augmentation approach. Unlike source-free test-time adaptation methods, STyle-TTA assumes access to a collection of labeled source-domain samples during inference but does not require target-domain labels or classifier modification, and unlike every style-based method reviewed above, it requires no source-specific training or calibration

of the stylization mechanism itself. This setting reflects practical scenarios where the original training data are available while deployed models cannot be retrained or updated. The framework is evaluated across diverse domain-shift scenarios, including natural image recognition benchmarks such as ImageNet-R using ImageNet-1K as source domain, as well as additional domain-specific datasets involving medical imaging and satellite imagery. By operating entirely at the input level and requiring no source-specific training of the style-transfer mechanism, STyle-TTA provides a general framework that can be applied to different classifiers and domain shifts without requiring task-specific optimization.

4 Methods

4.1 Overview of the Proposed Framework

This chapter presents *Style-Based Test-Time Augmentation* (STyle-TTA), a framework for improving image classification under domain shift. Conventional test-time augmentation generates multiple geometric or photometric transformations of a test image. In contrast, STyle-TTA constructs an ensemble of style-transferred variants whose appearance is adapted toward source-domain characteristics, while preserving the semantic content of the target image required for classification.

We consider a source domain

$$\mathcal{D}_S = \{(x_i^S, y_i^S)\}_{i=1}^{N_S}, \qquad x_i^S \in \mathcal{X},\ y_i^S \in \mathcal{C}_S, \tag{8}$$

consisting of labeled images available during classifier training, and a target domain

$$\mathcal{D}_T = \{(x_j^T, y_j^T)\}_{j=1}^{N_T}, \qquad x_j^T \in \mathcal{X},\ y_j^T \in \mathcal{C}_T, \tag{9}$$

drawn from a distribution shifted relative to $\mathcal{D}_S$. A dataset is treated as a target domain precisely when no samples from it are observed during classifier training; its labels y_j^T are used only for evaluation, never to adapt f_θ or construct augmentations.

Since $\mathcal{D}_S$ and $\mathcal{D}_T$ may originate from different data collections, their label sets $\mathcal{C}_S$ and $\mathcal{C}_T$ need not coincide. We therefore restrict evaluation to the shared label space

$$\mathcal{C} = \mathcal{C}_S \cap \mathcal{C}_T, \tag{10}$$

and define the class-restricted subsets

$$\begin{aligned} \mathcal{D}_S^{\mathcal{C}} &= \{(x, y) \in \mathcal{D}_S : y \in \mathcal{C}\}, \\ \mathcal{D}_T^{\mathcal{C}} &= \{(x, y) \in \mathcal{D}_T : y \in \mathcal{C}\}, \end{aligned} \tag{11}$$

so that every classifier considered predicts over the same $|\mathcal{C}|$ classes.

Let $f_\theta : \mathcal{X} \to \mathbb{R}^{|\mathcal{C}|}$ denote a classifier trained using only $\mathcal{D}_S^{\mathcal{C}}$, with parameters θ that remain frozen at inference time. For an input $x \in \mathcal{X}$, the predicted label is $\hat{y}(x) = \arg\max_{c \in \mathcal{C}} f_\theta(x)_c$.

Goal. Under distribution shift, the accuracy of $\hat{y}(x_j^T)$ for $x_j^T \in \mathcal{D}_T^{\mathcal{C}}$ degrades relative to the source domain. The objective of this thesis is to improve $\hat{y}(x_j^T)$ without updating θ, by transforming the input rather than the model. We formalize this as test-time augmentation: instead of presenting f_θ with the raw input x_j^T, we construct K transformed views $\{t_1(x_j^T), \ldots, t_K(x_j^T)\}$ and aggregate their predictions,

$$\hat{y}(x_j^T) = \arg\max_{c \in \mathcal{C}} \frac{1}{K} \sum_{k=1}^{K} f_\theta\big(t_k(x_j^T)\big)_c. \tag{12}$$

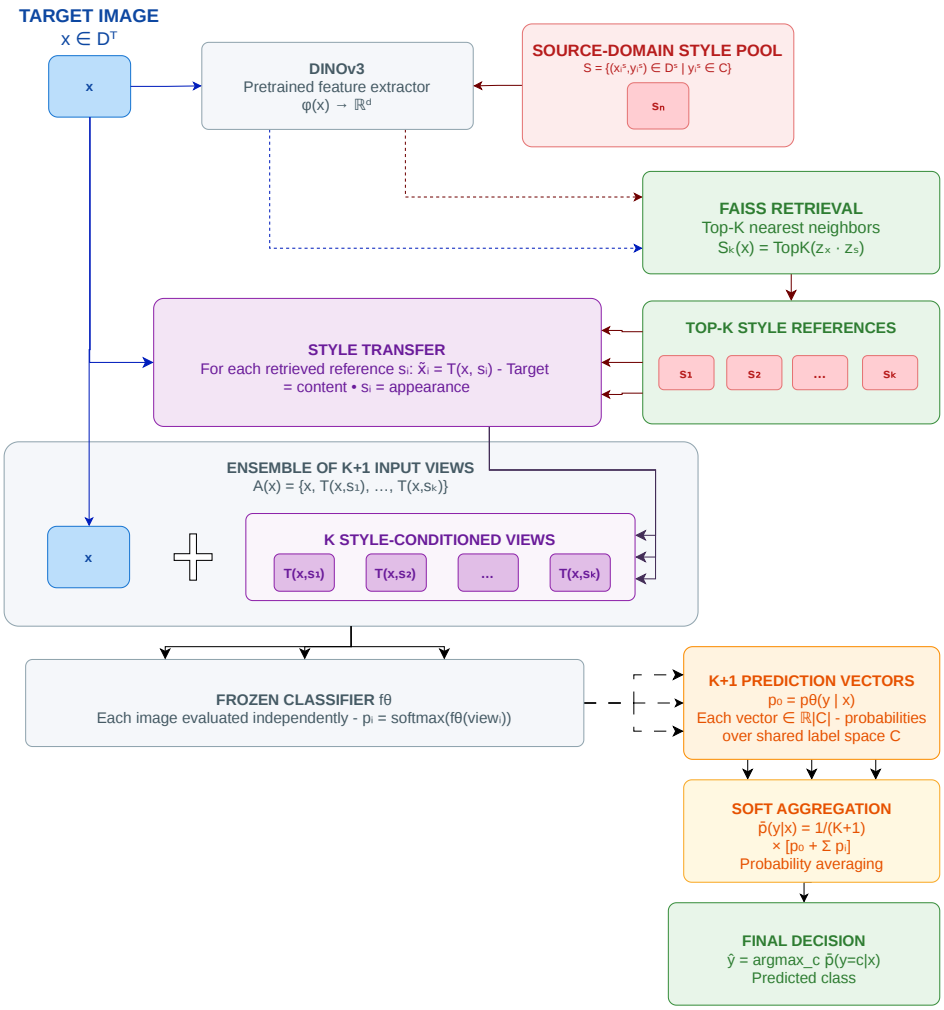


Figure 6: Overview of the proposed Style-Based Test-Time Augmentation (STyle-TTA) framework.

The framework operates entirely at inference time. It requires neither modifications to the classifier architecture nor access to target-domain labels, making it applicable to frozen or black-box classifiers whose parameters cannot be updated after deployment.

The framework proceeds in four stages (Figure 6). First, a target image is encoded by a pre-trained feature extractor and compared against a reference pool constructed from $\mathcal{D}_S^{\mathcal{C}}$. This style pool contains only source-domain samples from the shared label space, ensuring compatibility with the downstream classification task. Based on feature similarity, the K nearest source images are retrieved as candidate style references.

Second, each retrieved reference is combined with the target image via a training-free style transfer operator to generate a stylized view. This step modifies appearance-related characteristics while preserving the semantic structure of the target image.

The framework treats style transfer as a modular component. In this work, we investigate two training-free operators: StyleID, an high-end diffusion-based style-injection method, and alternatively, Adaptive Instance Normalization (AdaIN), which performs feed-forward feature-statistic matching. This modular design allows us to determine whether the observed benefits arise from style augmentation itself or from the specific properties of diffusion-based transfer. Third, the original target image together with all K stylized views are independently evaluated by the frozen classifier, each producing a probability distribution over the shared label space. Finally, the individual predictions are aggregated via soft averaging to obtain the final classification. By combining multiple appearance-adapted observations of the same semantic instance, the framework reduces the influence of domain-specific appearance variations while retaining the semantic information necessary for accurate classification.

The following sections describe each stage in detail, beginning with the problem formulation and proceeding through style reference retrieval, style transfer operators, and prediction aggregation.

4.2 Style Reference Retrieval

The proposed STyle-TTA framework requires source-domain reference images that provide source-domain information for the style transfer operation. Since arbitrary style references may introduce appearance characteristics that are inconsistent with the semantic structure of the target image, the framework retrieves source-domain images that are visually similar according to a semantic embedding space of the target sample before performing style transfer.

The style reference pool is constructed from the source-domain training set. To ensure compatibility with the classification task, only source-domain samples belonging to classes shared between the source and target domains are included:

$$\mathcal{S} = \{(x_i^S, y_i^S) \in \mathcal{D}_S \mid y_i^S \in \mathcal{C}\}. \tag{13}$$

Although class information is used when constructing the reference pool by limiting the source domain to the shared label space, the retrieval process itself is completely label-agnostic. Retrieved references are selected exclusively based on visual similarity in an embedding space and are not constrained to share the same class label as the target image. This design choice allows the framework to evaluate whether style information from semantically related but potentially different classes can improve robustness under domain shift.

Feature Representation To measure similarity between target images and source-domain references, images are embedded using a pre-trained DINOv3 vision transformer:

$$\phi(x) \in \mathbb{R}^d. \tag{14}$$

Specifically, the model `vit_base_patch16_dinov3.lvd1689m` is used as a feature extractor. The resulting final image representation is used as the retrieval embedding. Meta DINOv3 (Siméoni et al., 2025) is selected due to its strong semantic representation capability arising from self-supervised pretraining over 1.689 billion images (LVD-1689M), which allows the extracted features to capture object-level similarity without requiring task-specific labels. Furthermore, self-supervised representations are expected to provide improved robustness across visual domains compared to representations optimized solely for a specific classification task.

The extracted embeddings are L2-normalized:

$$\hat{\phi}(x) = \frac{\phi(x)}{\|\phi(x)\|_2}. \tag{15}$$

Normalization allows similarity computation through cosine similarity, which measures the angular distance between feature representations while removing the influence of embedding magnitude:

$$\mathrm{sim}(x, s) = \frac{\hat{\phi}(x)^T \hat{\phi}(s)}{\|\hat{\phi}(x)\|_2 \, \|\hat{\phi}(s)\|_2}. \tag{16}$$

Since all embeddings are normalized, cosine similarity is equivalent to the inner product between feature vectors.

Nearest-neighbour Retrieval For computational efficiency, source-domain embeddings are computed once for each dataset and stored before inference. During evaluation, only the embedding of the target image needs to be computed online. The complete source reference pool is indexed using the FAISS library with an inner-product search index, enabling efficient nearest-neighbourretrieval over the normalized embeddings.

Given a target image x, the retrieval operation selects the K most similar source-domain references:

$$\mathcal{S}_K(x) = \{s_1, \ldots, s_K\}, \tag{17}$$

where $s_i \in \mathrm{TopK}_{s \in \mathcal{S}} \, \mathrm{sim}(x, s)$. The value of K is fixed for each experimental configuration and determined before inference. No additional selection based on class labels, duplicate removal, or manual filtering is performed during retrieval. This ensures that the retrieved reference set is determined solely by the learned visual representation of DINOv3.

Each retrieved image is subsequently used as a style donor:

$$\tilde{x}_i = T(x, s_i), \qquad s_i \in \mathcal{S}_K(x), \tag{18}$$

where T denotes the selected transformation operator.

By retrieving semantically compatible source-domain examples before style transfer, STyle-TTA aims to expose the classifier to source-like appearance variations while

avoiding arbitrary stylistic changes that could interfere with the semantic content of the target image. As shown in Figure 7, and further examined in the ablation section 5.8.1. The high semantic similarity of the retrieved references increases the likelihood that style images associated with the correct label are selected.

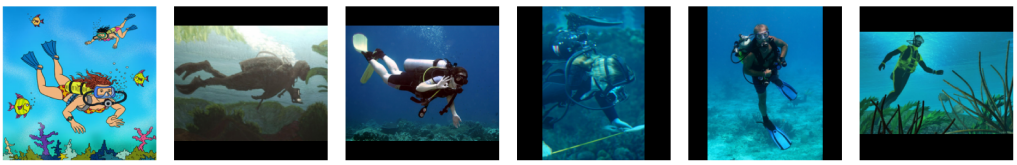


Figure 7: ImageNet-R test image (left) with its top-5 nearest-neighbour picks, from left to right. (All class="scuba diver")

4.3 Style Transfer Operators

The style transfer operator represents the transformation component of STyle-TTA that generates alternative views of a target-domain sample. Unlike conventional test-time augmentations, which modify images through predefined transformations such as cropping, flipping, or color perturbations, the proposed framework generates appearance-shifted observations by transferring source-domain style characteristics to the target sample.

Per equation 18, given a target-domain image x and a retrieved source-domain reference image s_i, a style transfer operator $T(\cdot)$ generates: $\tilde{x}_i$.

The generated image $\tilde{x}_i$ is intended to preserve the semantic identity of the original sample while modifying appearance-related factors that may contribute to the domain discrepancy. The STyle-TTA framework is designed to be independent of the specific style transfer mechanism. In theory, any operator capable of generating content-preserving style transformations from arbitrary reference images can be incorporated. In practise, feature preservation and high-quality stylisation are required, for which we use diffusion-based style transfer. In this work, two training-free operators are investigated: StyleID as the primary transformation method and AdaIN as a computationally efficient comparison operator.

StyleID The main style transfer operator used in STyle-TTA is StyleID (Chung et al., 2024). The method is selected for its training-free design, which aligns with the requirements of test-time augmentation under domain shift. In contrast to approaches that require domain-specific training or optimization for each target image, StyleID enables direct transfer from an arbitrary reference image at inference time.

For STyle-TTA, StyleID is interpreted as an appearance adaptation mechanism rather than a generative image synthesis model. The target image provides the semantic structure that should remain unchanged, while the retrieved source-domain image provides appearance information. Internally, StyleID performs style transfer

by manipulating intermediate diffusion representations, preserving content information through the content image attention structure while injecting appearance-related information from the style reference. The resulting image represents an alternative observation of the same semantic sample under source-like appearance conditions. Multiple retrieved references therefore produce multiple style-conditioned views, using an adapted version of equation 18:

$$\mathcal{A}_{StyleID}(x) = \{T_{\text{StyleID}}(x, s_1), \ldots, T_{\text{StyleID}}(x, s_K)\}. \tag{19}$$

These views are subsequently evaluated by the frozen classifier and combined through probability aggregation as described in Section 4.4.

In the proposed implementation, StyleID is used purely as an inference-time transformation. The pre-trained diffusion model remains unchanged, no fine-tuning is performed, and no textual conditioning is used.

To maintain image fidelity while meeting model specifications, image resolutions are explicitly managed throughout the pipeline. Content images and retrieved style references are first resized to the diffusion model's native processing resolution of 512×512 (R_D). To prevent distortion, this transformation preserves the original aspect ratio prior to processing. Following style generation, the stylized outputs are resized to 224×224 (R_C), matching the input resolution expected by the downstream classifier:

$$\tilde{x}_i = R_C \left(T_{\text{StyleID}}(R_D(x), R_D(s_i)) \right), \tag{20}$$

where R_D and R_C denote the aspect-ratio-preserving resizing operations for the diffusion model (512×512) and the classifier (224×224), respectively.

AdaIN Comparison Operator Although StyleID is used as the primary operator, the STyle-TTA formulation does not solely depend on diffusion-based style transfer. To investigate whether the observed benefits originate from style-based augmentation itself rather than the specific properties of StyleID, AdaIN Huang and Belongie (2017) is additionally evaluated.

AdaIN provides a lightweight alternative by transferring style information through feature statistic alignment. Unlike StyleID, whose DDIM process is an iterative, AdaIN aligns the channel-wise feature statistics of the content image through its feed-forward network much faster.

Within STyle-TTA, AdaIN follows the identical transformation interface: $\tilde{x}_i = T_{\text{AdaIN}}(x, s_i)$.

The same retrieved style references, classifier evaluation procedure, and probability aggregation strategy are used for both operators. As StyleID is a computationally more expensive operation, both in VRAM and inference time, it limits the variations of experiments that can be reliably done within the timeframe of a master thesis. AdaIN is included primarily due to its computational efficiency while still not requiring additional training to perform arbitrary reference-based style transfer. Figure 8 gives a small comparison of their output.

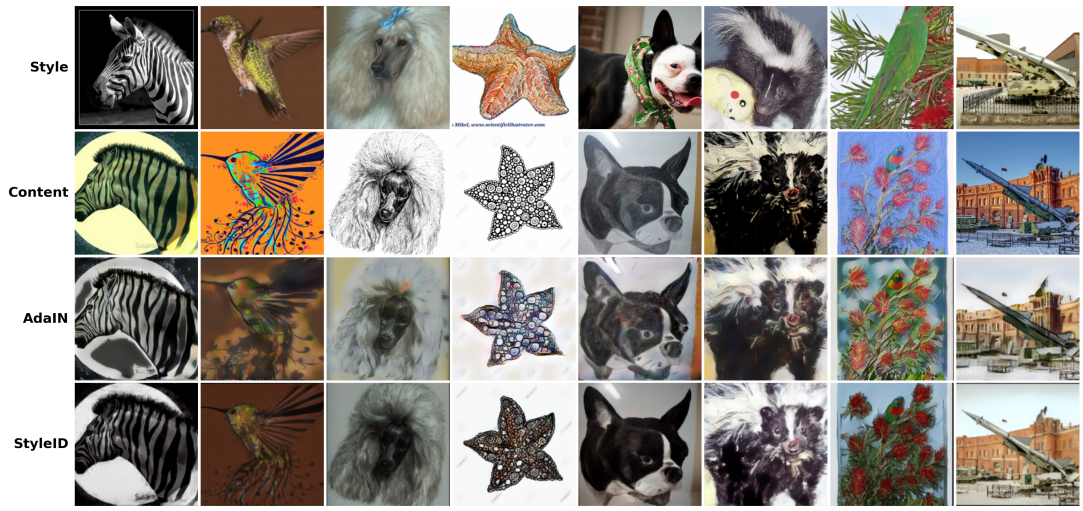


Figure 8: Image comparison roster between StyleID and AdaIN. Content images are from ImageNet-R and style images from ImageNet-1K.

This ensures that differences between StyleID and AdaIN reflect the properties of the style transfer mechanism itself rather than differences in the augmentation pipeline.

4.4 Prediction Aggregation

After generating style-transferred views, STyle-TTA aggregates predictions across the original target image x and its style-conditioned observations $\tilde{x}_i = T(x, s_i)$ for $s_i \in \mathcal{S}_K(x)$. While conventional TTA combines predictions from predefined geometric or photometric transformations, STyle-TTA constructs an ensemble over source-style adapted views generated through the style transfer operator.

The augmented observation set $\mathcal{A}(x) = \{x\} \cup \{\tilde{x}_i\}_{i=1}^{K}$ explicitly retains the original image to preserve the classifier's performance under the unmodified target-domain appearance, preventing decisions from relying exclusively on generated samples.

Let $p_\theta(y \mid x_i) = \text{softmax}(f_\theta(x_i))$ denote the probability distribution produced by the frozen classifier for a view $x_i \in \mathcal{A}(x)$ over label space $\mathcal{C}$. The final prediction distribution is obtained by averaging predicted probabilities across all $K+1$ views:

$$\bar{p}(y \mid x) = \frac{1}{K+1}\left(p_\theta(y \mid x) + \sum_{i=1}^{K} p_\theta(y \mid \tilde{x}_i)\right). \tag{21}$$

The final decision is then assigned via $\hat{y} = \arg\max_{c \in \mathcal{C}} \bar{p}(y = c \mid x)$.

Unlike hard majority voting, soft probability averaging preserves output confidence and allows complementary predictions across style views to contribute to the final decision. Furthermore, ensembling multiple independently retrieved style references

reduces sensitivity to individual style-transfer artifacts. Because the classifier parameters θ remain frozen throughout, this step forms a pure, inference-time input-space ensemble.

The complete STyle-TTA procedure is summarized in Algorithm 1.

Algorithm 1 Style-Based Test-Time Augmentation (STyle-TTA)

Require: Frozen classifier f_θ, target image x, style pool $\mathcal{S}$, embedding model ϕ, number of references K, style operator T

Ensure: Predicted class $\hat{y}$

(Offline) Compute style pool embeddings: $z_{\mathcal{S}} \leftarrow \phi(\mathcal{S})$

Compute target embedding: $z_x \leftarrow \phi(x)$

Retrieve top-K nearest style references:

$$\mathcal{S}_K(x) \leftarrow \mathrm{TopK}_{s \in \mathcal{S}} \left(\mathrm{sim}(z_x, z_{\mathcal{S}})\right)$$

Initialize prediction set with original image: $P \leftarrow \{f_\theta(x)\}$

for $s_i \in \mathcal{S}_K(x)$ **do**

  Generate style-transferred image: $\tilde{x}_i \leftarrow T(x, s_i)$

  Obtain prediction: $p_i \leftarrow f_\theta(\tilde{x}_i)$

  Add prediction to ensemble: $P \leftarrow P \cup \{p_i\}$

end for

Aggregate probabilities:

$$p_{\text{final}} \leftarrow \frac{1}{K+1} \sum_{p \in P} p$$

return $\hat{y} \leftarrow \mathrm{argmax}_c \, p_{\text{final},c}$

4.5 Implementation Detail

All experiments were conducted using PyTorch (v2.11.0)-based implementations built on `torchvision` (v0.26.0), `timm` (v1.0.22), `open_clip` (v3.3.0), and HuggingFace `diffusers` (v0.35.2) for the diffusion-based style transfer component. Style-reference retrieval utilized the FAISS library (`faiss-cpu` v1.13.2) with an inner-product index over L_2-normalized DINOv3 embeddings, precomputed once per source-domain pool and reused across all target-domain queries.

To handle the computational demands of style-based test-time augmentation (STyle-TTA), particularly for diffusion-based style transfer, experiments were executed across different high-performance computing (HPC) environments. ImageNet-based experiments were conducted on the Erlangen HPC infrastructure using NVIDIA A100 GPUs, distributed across one to four GPUs depending on the number of generated style references. To eliminate inference overhead, all generated stylized views were precomputed and cached prior to evaluation; classifiers subsequently operated

directly on these cached views. Experiments on the remaining datasets were performed on faculty computing infrastructure using a single NVIDIA RTX A5000 GPU (24 GB VRAM). All input images were converted to floating-point tensors and scaled to the range $[0, 1]$. Images were resized to the classifier input resolution while preserving their original aspect ratio via padding. ResNet-18, DenseNet-121, ViT-B/16, Swin-B, and DINOv2 were normalized with standard ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$); CLIP uses its own pretraining normalization statistics instead ($\mu = [0.48145466, 0.4578275, 0.40821073]$, $\sigma = [0.26862954, 0.26130258, 0.27577711]$), consistent with the OpenAI CLIP preprocessing pipeline.

StyleID was executed using its official reference configuration (Chung et al., 2024): Stable Diffusion v1.4 (`CompVis/stable-diffusion-v1-4`, `float16` precision), 50 DDIM sampling/inversion steps, a query-preservation factor of $\gamma = 0.75$, an attention temperature of $T = 1.5$, self-attention injection at UNet decoder layers 6–11, AdaIN applied to the initial latent, and attention injection activated from step 49 (out of 50) onward. AdaIN utilized a VGG-19 encoder–decoder architecture, following the formulation of Huang and Belongie (2017) (Section 2.3).

All experiments were conducted strictly in an inference-time adaptation setting: classifier parameters remained completely frozen during evaluation, and no target-domain samples or labels were used for optimization. The default inference batch size was set to 1 due to memory constraints associated with processing multiple augmented views per instance, using four parallel data-loading workers. Given the substantial computational cost of generating diffusion-stylized views, STyle-TTA experiments were evaluated using a single random seed, whereas geometric TTA experiments were averaged over three random seeds to capture variance from augmentation sampling.

5 Experiments & Results

This chapter evaluates STyle-TTA across a diverse set of domain-shift benchmarks and classifier architectures. We first conduct an in-depth analysis on the primary ImageNet-1K → ImageNet-R benchmark and comparing STyle-TTA against a geometric TTA baseline in detail (Section 5.4). We then assess generalization across our four additional remote sensing and computational pathology benchmarks (Section 5.5), followed by an analysis of calibration effects (Section 5.7), ablations over the core design choices (Section 5.8), and a hybrid approach combining style-based and geometric augmentation (Section 5.10). Sections 5.1 and 5.3 introduce the full set of datasets and classifiers used throughout.

5.1 Datasets

Experiments were conducted on three source–target dataset groups covering natural image classification, remote sensing imagery, and computational pathology, comprising six datasets in total. Table 2 details each dataset, its modality, including the data splits used, and the corresponding number of classes and samples. Since in some cases the D_S and D_T did not have a fully overlapping label space, we restricted it to their overlapping label subset $\mathcal{D}_S^{\mathcal{C}}$ and $\mathcal{D}_T^{\mathcal{C}}$. For the evaluation, the target was adjusted to C as was the source dataset to restrict the style images to only task relevant labels. As a measure to quantify the class distribution of each dataset split, we report Shannon's Equitability (Eq) (Shannon, 1948), which is defined as the normalized Shannon entropy,

$$\frac{-\sum_{c=1}^{C} p_c \log p_c}{\log C}, \tag{22}$$

where ($\mathrm{Eq} \in [0, 1]$). Values closer to 1 indicate a more uniform distribution of samples across classes, with ($\mathrm{Eq} = 1$) corresponding to a perfectly balanced distribution. Figure 9 shows a small overview of the datasets.

Natural Images Classification Benchmark

The first benchmark consists of the ImageNet-1K and ImageNet-R datasets. ImageNet-1K is used as the source domain (D_S), comprising 1,281,167 training images and 50,000 validation images across 1,000 classes (Russakovsky et al., 2015). ImageNet-R is used as the target domain (D_T), comprising 30,000 test images across 200 classes (Hendrycks et al., 2021). Since ImageNet-R classes are a subset of the full ImageNet-1K, with only 200 classes. Evaluation is therefore restricted to these $C = 200$ shared classes. Target-domain evaluation was performed on the full ImageNet-R test split.

Table 2: Overview of the datasets used in our experiments across natural images, remote sensing, and histopathology. For each dataset, we report the number of classes, the number of samples in the training, validation, and test splits, and as a measure of class-distribution balance Shannon's Equitability (Eq). It ranges from 0 to 1, with values closer to 1 indicating a more uniform distribution of samples across classes.

Modality	**Dataset**	**# Classes**	**Train / Val / Test**	**Eq**
Natural Images	ImageNet	1000	1,281,167 / 50,000 / -	0.9998 / 1.00
	$ImageNet_C$	200	258,951 / 10,000 / -	0.9999 / 1.0000
	ImageNet-R	200	- / - / 30,000	0.9797
Remote Sensing	EuroSAT	10	21,600 / 2,700 / -	0.9966 / 0.9966
	$EuroSAT_C$	4	12,800 / 1,600 / -	0.9182 / 0.9182
	UCMerced	21	- / - / 2,100	1.00
	$UCMerced_C$	4	- / - / 600	0.8962
Histopathology	MIDOG2022	2	12,270 / 3,006 / 3,299	0.9866 / 0.9882 / 0.9470
	Camelyon17-WILDS	2	302,436 / 34,904 / 2,000	1.00 / 1.00 / 1.00
	Epithelium-Stroma	2	8,337 / 5,920 / 1,376	0.9783 / 0.9992 / 0.9712

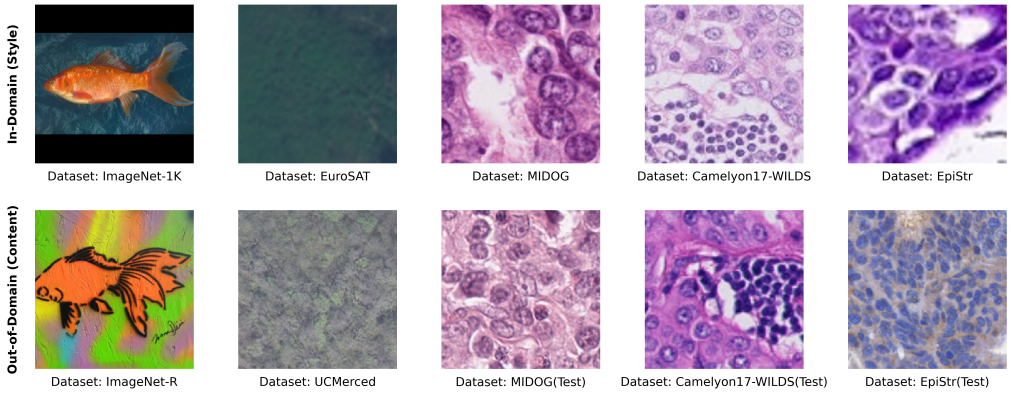


Figure 9: Examples of in-domain and out-of-domain images from the different benchmarks. For EuroSAT and ImageNet-1K, the test sets are different datasets (UCMerced/ImageNet-R). The pathology datasets use their respective OOD data.

Remote Sensing Benchmark

The second benchmark evaluates transfer from EuroSAT (D_S) to UCMerced Land Use (D_T). EuroSAT contains 27,000 images across 10 classes from Sentinel-2 satellite imagery (Helber et al., 2019), while UCMerced Land Use contains 21 classes with 100 images per class, for a total of 21,000 images (Yang and Newsam, 2010). Because the original label taxonomies only partially overlapping, evaluation is restricted to a manually constructed four-class shared taxonomy,

$$\mathcal{C} = \{\text{agricultural, forest, residential, river}\}, \tag{23}$$

with the class correspondences shown in Table 3. The table describes how the original classes are aggregated into the shared taxonomy; classes without a correspondence are excluded. Applying this mapping to EuroSAT produces the four-class $EuroSAT_C$, comprising 12,800 training and 1,600 validation images. The same mapping applied to UCMerced produces the four-class $UCMerced_C$, comprising 600 images in total: 100 agricultural, 100 forest, 300 residential, and 100 river images. Since UCMerced is used exclusively for target-domain evaluation in this benchmark, the complete mapped subset is retained as a single target evaluation set and is not further divided into training, validation, or test splits.

Table 3: Construction of the shared label space for the EuroSAT–UCMerced transfer benchmark. The original class taxonomies are mapped to four semantically aligned categories to enable cross-dataset evaluation, while classes without a correspondence are excluded.

Shared class	**EuroSAT classes**	**UCMerced classes**
Agricultural	AnnualCrop, PermanentCrop, Pasture	agricultural
Forest	Forest	forest
Residential	Residential	denseresidential, mediumresidential, sparseresidential
River	River	river

Computational Pathology Benchmarks

The datasets used in this study are summarized in Table 4. Unlike the previous benchmark pairs, these datasets contain multiple acquisition domains, enabling source and target partitions to be defined from their native structure.

Table 4: Overview of the computational pathology benchmarks, their domain origins, and experimental splits.

Dataset	**Domain**	**Origin / Scanner**	**Tissue**	**Split**	**Samples**
Camelyon17	Centers 0–2	CWZ, RST, RUMC; 3DHistech	Breast, H&E	Source train	302,436
	Center 3	UMCU; Hamamatsu	Breast, H&E	Source val.	34,904
	Center 4	LPON; Philips	Breast, H&E	Target test	85,054 (2,000 eval.)
MIDOG2022	A, B, C, E	Multiple scanners; see text	Multiple tumors, H&E	Source train	12,270
	A, B, C, E	Multiple scanners; see text	Multiple tumors, H&E	Source val.	3,006
	D	Hamamatsu NanoZoomer XR	Neuroendocrine, H&E	Target test	3,299
	F	Unlabeled challenge set	–	Excluded	–
Epithelium–Stroma	NKI	Netherlands	Breast, H&E	Source train	8,337
	VGH	Vancouver	Breast, H&E	Validation	5,920
	IHC	Finland	Colorectal, IHC	Target test	1,376

Camelyon17-WILDS The Camelyon17-WILDS classification-challenge dataset is made of breast cancer metastases in whole-slide images of histological lymph node sections (Litjens et al., 2018; Bandi et al., 2018). Following the WILDS benchmark's hospital-based split, 96×96-pixel patches with binary tumor/normal labels are partitioned across five medical centers in the Netherlands. The source domain comprises 302,436 training patches (Centers 0, 1, and 2) and 34,904 validation patches (Center 3), while Center 4 serves as the official out-of-distribution target domain test set (85,054 samples). The five centers are CWZ, RST, UMCU, RUMC, and LPON; slides from CWZ, RST, and RUMC were scanned with a 3DHistech Pannoramic Flash II 250, UMCU with a Hamamatsu NanoZoomer-XR C12000-01, and LPON with a Philips Ultrafast Scanner (Litjens et al., 2018). To keep per-sample style retrieval and style-transfer inference tractable across all STyle-TTA configurations, a random balanced subset of 2,000 samples was drawn from the target domain for evaluation.

MIDOG2022 The MIDOG dataset is an open-source challenge dataset (Aubreville et al., 2024). Patches of 224×224 pixels were extracted centered on annotated mitotic and non-mitotic objects across various tumor types and digitization scanners. Source-domain data are drawn from scanner groups A, B, C, and E (comprising human breast carcinoma, canine lung carcinoma, canine cutaneous mast cell tumor, and canine lymphoma scanned via Hamamatsu NanoZoomer XR, Hamamatsu S360, Aperio ScanScope CS2, and 3DHistech Pannoramic Scan II at 0.23–0.25 μm/px). These are randomly split 80/20 into training (12,270 images) and validation (3,006 images). Scanner group D (human neuroendocrine tumor digitized on a Hamamatsu NanoZoomer XR) is held out entirely as the target evaluation domain (3,299 test images). Scanner group F is excluded from this experimental setup as it was the unlabeled challenge dataset.

Epithelium-Stroma (EpiStr) The aggregated Epithelium-Stroma dataset, consists of the publicly available NKI, VGH, (Beck et al., 2011) and IHC (Linder et al., 2012) datasets. The NKI and VGH datasets comprise H&E stained breast cancer tissue images, while the IHC dataset consists of IHC-stained colorectal cancer tissue images. This dataset evaluates tissue compartment classification (binary stroma vs. epithelium) across three cohort-level domains: NKI as source (8,337 training images, Hematoxylin and Eosin [H&E]-stained breast cancer tissue), VGH as validation (5,920 images, H&E-stained breast cancer tissue), and IHC as target (1,376 test images, Immunohistochemistry [IHC]-stained colorectal cancer tissue). Note that, unlike Camelyon17 and MIDOG, where source and target differ primarily in acquisition hardware, the EpiStr target domain differs from its source in both stain type and underlying tissue/cancer type, representing a substantially larger domain gap than the other two pathology datasets.

5.2 Evaluation Metrics

To evaluate the applicability of STyle-TTA under domain shift, we assess both predictive performance and the quality of the resulting confidence estimates. Since TTA methods modify the inference process by generating and aggregating multiple transformed views of each input sample, its effect is not limited to classification accuracy. In addition to measuring whether predictions become more accurate, we evaluate whether the confidence assigned to these predictions better reflects the actual likelihood of correctness. Therefore, the evaluation considers both classification performance and probabilistic calibration.

Classification Performance For each test sample, we record whether the predicted class with the highest probability corresponds to the ground-truth label. Top-1 accuracy is defined as:

$$\mathrm{Acc}_1 = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i), \tag{24}$$

where $\hat{y}_i$ denotes the predicted class for sample i, y_i denotes the corresponding ground-truth label, and N represents the total number of test samples.

Since domain shifts may affect individual classes unequally, standard accuracy can obscure class-specific degradation. Therefore, balanced accuracy is additionally reported. Unlike standard accuracy, which weights all samples equally regardless of class membership, balanced accuracy computes the average recall across all classes (Brodersen et al., 2010):

$$\mathrm{BAcc} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \frac{TP_c}{TP_c + FN_c}, \tag{25}$$

where $\mathcal{C}$ denotes the set of classes, and TP_c and FN_c represent the true-positive and false-negative counts for class c, respectively. Balanced accuracy is particularly relevant under domain shift, where changes in the input distribution may disproportionately affect specific classes.

Prediction Calibration In addition to classification performance, we evaluate whether the confidence estimates produced by the classifier are reliable. A calibrated classifier should assign confidence values that correspond to its empirical probability of being correct. For example, among predictions made with a confidence of 80%, approximately 80% should be correct.

We measure calibration using Expected Calibration Error (ECE) (Naeini et al., 2015; Guo et al., 2017). Predictions are grouped into M confidence bins according to their predicted probability. The ECE is computed as:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{26}$$

where B_m denotes the set of predictions assigned to confidence bin m, $\mathrm{acc}(B_m)$ is the empirical accuracy of the bin, and $\mathrm{conf}(B_m)$ is the average predicted confidence within that bin. Lower ECE values indicate better calibration, meaning that the predicted confidence more accurately reflects the actual likelihood of a correct prediction.

5.3 Classifiers

To evaluate the effect of STyle-TTA across diverse visual architectures, we consider six image-classification backbones spanning three architectural families: convolutional neural networks (CNNs; ResNet-18 and DenseNet-121), vision transformers (ViTs; ViT-B/16 and Swin-B), and foundation-model backbones (DINOv2 ViT-B/14 and CLIP ViT-B/16). The models differ not only in architectural design and pre-training objective, but also in how their representations are adapted to the downstream classification tasks.

Table 5 summarizes the evaluated models, their implementation frameworks, pre-trained weights, backbone dimensionality, classification head, number of trainable parameters, and adaptation strategy. For the ImageNet-1K → ImageNet-R benchmark, ResNet-18, DenseNet-121, ViT-B/16, and Swin-B are evaluated using their ImageNet-1K-pretrained classification heads, without additional source-domain training. DINOv2 and CLIP, whose pre-training objectives do not directly provide ImageNet-1K classification heads, are instead equipped with linear probes trained on ImageNet-1K. For the remaining domain-transfer benchmarks, ResNet-18, DenseNet-121, and ViT-B/16 are fully fine-tuned on the source domain, whereas Swin-B, DINOv2, and CLIP use frozen backbones with linear probes trained on source-domain data.

Table 5: Overview of evaluation classifiers.

Model	Source	Framework	Pretrained weights	Backbone params	Feature dim.	Classifier head	Adaptation
ResNet-18	He et al. (2016)	timm	ImageNet-1K	11.7M	512	Linear	Frozen / FT
DenseNet-121	Huang et al. (2017)	timm	ImageNet-1K	8.0M	1024	Linear	Frozen / FT
ViT-B/16	Dosovitskiy et al. (2021)	timm	ImageNet-1K	∼86M	768	Linear	Frozen / FT
Swin-B	Liu et al. (2021)	timm	ImageNet-1K	∼88M	1024	Linear	Frozen / LP
DINOv2 ViT-B/14	Oquab et al. (2023)	torch.hub	DINOv2	∼86M	768	Linear probe	LP
CLIP ViT-B/16	Radford et al. (2021)	open_clip	OpenAI	∼86M	512	Linear probe	LP

CNN and transformer classifiers were loaded via `timm`. CLIP was loaded via `open_clip` using the `ViT-B-16` architecture with OpenAI pre-trained weights, while DINOv2 was loaded via `torch.hub` from the `facebookresearch/dinov2` repository using the `dinov2_vitb14` model. Note that DINOv3, used exclusively for style-reference retrieval (Section 3.3), and DINOv2, evaluated separately as a classifier (Section 4.3),

were selected independently for their respective roles and are therefore not interchangeable within the pipeline.

Linear Probe Implementation For DINOv2 and CLIP, the backbone is kept entirely frozen (`requires_grad = False`), and only a linear classification layer is trained on top of the extracted image representations. The feature dimensionality is determined dynamically from the backbone at initialization rather than being hardcoded. For CLIP, image features are additionally L_2-normalized before being passed to the linear layer. The resulting logits are rescaled using CLIP's pre-trained logit-scale temperature, which is kept frozen. This preserves the cosine-similarity-based feature geometry of the pre-trained CLIP representation while allowing the linear probe to be trained for the downstream classification task.

5.3.1 Baseline (No-TTA) Classification Performance

Before evaluating STyle-TTA, each classifier was evaluated without test-time adaptation on both the source-domain validation split ($\mathcal{D}_S^{\mathcal{C}}$) and the target-domain evaluation split ($\mathcal{D}_T^{\mathcal{C}}$). These results provide the reference point for the STyle-TTA and geometric-TTA results reported in Section 5.3.

ImageNet-1K → ImageNet-R For the ImageNet-1K → ImageNet-R benchmark, ResNet-18, DenseNet-121, ViT-B/16, and Swin-B were evaluated using their ImageNet-1K-pretrained classification heads without further source-domain adaptation. DINOv2 and CLIP were instead equipped with linear probes trained on the full ImageNet-1K training set, corresponding to the full 1000-class label space ($C_S = 1000$).

For evaluation on ImageNet-R, the output logit space was restricted to the 200 target classes shared between ImageNet-1K and ImageNet-R. Specifically, the original 1000-dimensional logit vector was indexed using a predefined boolean class mask, yielding $z_{\text{target}} \in \mathbb{R}^{200}$ prior to the final prediction step, following the evaluation setup of Hendrycks et al. (2021).

Table 6 reports top-1 accuracy on the source and target domains, as well as balanced accuracy and the change in target-domain top-1 accuracy relative to the source-domain performance. In this setting, $\mathcal{D}_S^{\mathcal{C}}$ corresponds to the ImageNet-1K validation split, while $\mathcal{D}_T^{\mathcal{C}}$ corresponds to the ImageNet-R evaluation set restricted to the 200 shared classes.

The results show substantial degradation under the ImageNet-1K → ImageNet-R shift for all evaluated classifiers. The CNNs exhibit the largest drops in top-1 accuracy, with decreases of 51.74 percentage points for ResNet-18 and 50.47 percentage points for DenseNet-121. The transformer-based models show smaller, but still substantial, degradation, with decreases of 45.87 and 37.03 percentage points for ViT-B/16 and Swin-B, respectively. The two foundation-model backbones are the most robust in this setting, with CLIP and DINOv2 showing drops of 28.37 and

Table 6: Baseline (no-TTA) performance, ImageNet-1K (200 shared classes) → ImageNet-R. Reporting Top-1 accuracy, as well as balanced accuracy. The output logit vector $z \in \mathbb{R}^{1000}$ was sliced using a predefined class mask to yield $z_{\text{target}} \in \mathbb{R}^{200}$ prior to Softmax, directly matching the 200 shared target classes.

Backbone	$\mathcal{D}_S^{\mathcal{C}}$ **Top-1 Acc. (%)**	$\mathcal{D}_T^{\mathcal{C}}$ **Top-1 Acc. (%)**	$\mathcal{D}_T^{\mathcal{C}}$ **Bal. Acc. (%)**	**Δ Top-1 Acc. (%)**
ResNet-18	85.36	33.62	35.44	−51.74
DenseNet-121	89.07	38.60	40.60	−50.47
Swin-B	95.82	58.79	60.41	−37.03
ViT-B/16	92.58	46.71	47.89	−45.87
CLIP ViT-B/16	92.79	64.42	64.23	−28.37
DINOv2 ViT-B/14	95.64	67.47	69.04	−28.17

28.17 percentage points, respectively. DINOv2 achieves the highest target-domain top-1 accuracy (67.47%), followed by CLIP (64.42%).

5.3.2 Remaining Benchmarks

To complement the ImageNet-1K evaluation, Table 7 summarizes baseline (no-TTA) domain-transfer performance across four additional benchmark datasets: Satellite (EuroSAT → UCMerced), Camelyon17-WILDS, MIDOG2022, and EpiStr.

For each benchmark, models are grouped into three architectural families: CNN (DenseNet-121), ViT (Swin-B), and Foundation Model (DINOv2). Across these experiments, the CNN and ViT models that are adapted through fine-tuning use the source-domain training split, whereas Swin-B, DINOv2, and CLIP use frozen backbones with linear probes where specified by the experimental setup. Fine-tuning uses stochastic gradient descent (SGD) with momentum of 0.9, zero weight decay, an initial learning rate of 3×10^{-4}, a batch size of 32, and a maximum of 20 epochs. Linear probing uses the same SGD optimizer configuration with an initial learning rate of 1×10^{-2}, a batch size of 32, and a maximum of 50 epochs. In both settings, the learning rate is decayed using cosine annealing over the total number of epochs. Training uses seed 42, with random horizontal flipping and random resized cropping as the only training-time augmentations. Early stopping is based on validation loss. Full fine-tuning and linear-probing hyperparameters are again listed in Appendix A.2 and A.1

EuroSAT case For the EuroSAT → UCMerced benchmark, the source and target datasets do not share the same label space. We therefore manually defined a mapping from the EuroSAT classes to the four UCMerced target categories: agricultural, forest, residential, and river. Source-model logits were first converted to probabilities using Softmax. The probabilities of source classes corresponding to the same target category were then summed, and the resulting four-class probabilities

were converted back to log-space using $\log(p + \epsilon)$. This projection produces predictions in the target label space while preserving a logit representation for subsequent evaluation.

Table 7: Baseline (no-TTA) performance across the four additional benchmark datasets, comparing CNN (DenseNet-121), ViT (Swin-B), and Foundation Model (DINOv2) architectures. We report top-1 accuracy, balanced accuracy, and the change in top-1 accuracy between the source and target domains.

Dataset	Backbone	$\mathcal{D}_S^{\mathcal{C}}$ Top-1 Acc. (%)	$\mathcal{D}_T^{\mathcal{C}}$ Top-1 Acc. (%)	$\mathcal{D}_T^{\mathcal{C}}$ Bal. Acc. (%)	Δ Top-1 Acc. (%)
Satellite	CNN	78.40	32.67	33.83	−45.73
	ViT	80.25	53.83	68.58	−26.42
	FM	80.78	57.83	72.92	−22.95
Camelyon	CNN	82.75	67.95	67.66	−14.80
	ViT	90.07	92.75	92.76	+2.68
	FM	92.07	93.00	92.97	+0.93
MIDOG	CNN	57.32	40.47	48.98	−16.85
	ViT	58.05	54.62	51.80	−3.43
	FM	55.82	52.11	50.28	−3.71
EpiStr	CNN	92.35	59.45	50.12	−32.90
	ViT	93.70	75.73	69.87	−17.97
	FM	93.21	72.46	65.70	−20.75

The baseline results in Table 7 indicate that the magnitude of domain degradation varies substantially across datasets and architectural families. CNNs exhibit the largest target-domain degradation on three of the four benchmarks, while the transformer and foundation-model configurations generally retain higher target-domain performance.

The nature of the domain shifts differs across the individual benchmarks:

- **Satellite (EuroSAT → UCMerced):** All three model families experience a substantial reduction in top-1 accuracy when transferred from EuroSAT to UCMerced. The degradation is largest for the CNN (−45.73 percentage points), followed by the ViT (−26.42) and foundation model (−22.95). Balanced accuracy is substantially higher than top-1 accuracy for the ViT and FM configurations, reaching 68.58% and 72.92%, respectively.
- **Camelyon17-WILDS:** The CNN exhibits a 14.80 percentage-point decrease in top-1 accuracy on the target domain. In contrast, both the ViT and foundation-model configurations slightly improve their target-domain top-1 accuracy relative to the source domain, with changes of +2.68 and +0.93 percentage points, respectively. Their balanced accuracies closely track their top-1 accuracies.

- **MIDOG2022:** All three architectures experience a decrease in target-domain top-1 accuracy. The CNN shows the largest drop (−16.85 percentage points), while the ViT and FM configurations show smaller decreases of −3.43 and −3.71 percentage points, respectively. Target-domain balanced accuracy is close to 50% for all three configurations, ranging from 48.98% for the CNN to 51.80% for the ViT.

- **EpiStr:** Although all models achieve high source-domain top-1 accuracy (above 92%), their target-domain performance is substantially lower. The CNN suffers the largest degradation (−32.90 percentage points), followed by the FM (−20.75) and ViT (−17.97) configurations. The corresponding target-domain balanced accuracies are 50.12%, 69.87%, and 65.70%, respectively.

5.4 Comparison of TTA Strategies on ImageNet-R

In this section, we evaluate the two style-transfer operators in STyle-TTA against a conventional geometric TTA baseline on the ImageNet-1K → ImageNet-R domain shift. All classifiers remain frozen; only the inference-time augmentation strategy varies. Evaluation is performed over the shared 200-class label space.

Experimental protocol. Geometric TTA views were generated online during evaluation, following standard practice. Style-transferred views were pre-cached to amortize inference cost. Geometric TTA conditions were averaged over three random seeds (71397589, 133560673, 265017005); seed-to-seed variation was negligible (≤ 0.2 pp Top-1 accuracy). A single seed was used for style-based conditions due to their substantially higher computational cost.

We evaluate augmentation budgets using a logarithmic progression of the number of total views, with $N \in \{1, 2, 4, 8, 16, 32, 64\}$. This exponential progression provides increasingly larger augmentation budgets at a fixed number of evaluation points, allowing us to characterize performance scaling without exhaustively evaluating every possible view count. For Geometric TTA, we generate up to 63 augmented images, corresponding to 64 total views including the original image. Because STyle-TTA is computationally expensive, it is capped at 15 augmented views, corresponding to 16 total views.

Table 8 reports the best-performing configuration for each strategy, selected according to balanced accuracy: STyle-TTA at $N = 16$ and Geometric TTA at $N = 64$. Results for the complementary AdaIN-based style operator are reported separately in Appendix A.4.

The statistical analysis combines bootstrap confidence intervals with McNemar's test for the Top-1 predictions. The 95% confidence intervals reported in Table 8 are obtained by bootstrap resampling, while McNemar's test is used to assess the significance of paired Top-1 differences. All reported Top-1 improvements over the

corresponding baseline or competing TTA condition that are identified as significant satisfy $p < 0.001$. To visually inspect the fidelity of these transformations, qualitative style-transferred examples for all five benchmark datasets are provided in Appendix 10 (Figures 16 and 17).

Table 8: Evaluation of baseline ImageNet-R and TTA methods across CNN, Vision Transformer, and Foundation Model backbones. Bold indicates the best performance per metric. All Top-1 accuracy improvements were statistically significant based on McNemar's test ($p < 0.001$).

Backbone	Baseline (%)	Strategy	Top-1 Acc. (%)	Bal. Acc. (%)	Δ (95% CI)
ResNet-18	33.62	STyle-TTA	**45.94**	**46.80**	+12.32 [11.85, 12.79]
		Geometric	37.53 ± 0.02	39.46 ± 0.02	+3.91 [3.55, 4.27]
DenseNet-121	38.60	STyle-TTA	**49.24**	**50.32**	+10.64 [10.19, 11.10]
		Geometric	41.90 ± 0.07	44.07 ± 0.07	+3.30 [2.95, 3.65]
ViT-B/16 (224)	46.71	STyle-TTA	**56.91**	**57.82**	+10.20 [9.75, 10.66]
		Geometric	51.44 ± 0.05	52.80 ± 0.04	+4.73 [4.37, 5.11]
Swin-B (224)	58.79	STyle-TTA	**63.25**	**63.88**	+4.46 [4.03, 4.89]
		Geometric	60.93 ± 0.01	62.45 ± 0.01	+2.14 [1.80, 2.48]
CLIP ViT-B/16	64.42	STyle-TTA	62.99	62.78	−1.44 [-1.91, -0.97]
		Geometric	**66.15** ± 0.08	**65.64** ± 0.06	+1.73 [1.35, 2.12]
DINOv2 ViT-B/14	67.47	STyle-TTA	66.18	67.23	−1.29 [-1.66, -0.90]
		Geometric	**69.03** ± 0.05	**70.58** ± 0.07	+1.57 [1.28, 1.86]

Table 8 shows a clear dependence of the optimal TTA strategy on the backbone. For the conventional CNN and supervised transformer models—ResNet-18, DenseNet-121, ViT-B/16, and Swin-B—STyle-TTA achieves the highest Top-1 and balanced accuracy. The corresponding Top-1 accuracies are 45.94%, 49.24%, 56.91%, and 63.25%, respectively, compared with baseline accuracies of 33.62%, 38.60%, 46.71%, and 58.79%. This corresponds to absolute Top-1 improvements of 12.32, 10.64, 10.20, and 4.46 percentage points (pp), respectively.

STyle-TTA also consistently outperforms geometric TTA on these four backbones. For example, on ResNet-18, STyle-TTA reaches 45.94% Top-1 accuracy compared with 37.53% for geometric TTA. The advantage is similarly pronounced for DenseNet-121 (49.24% vs. 41.90%) and ViT-B/16 (56.91% vs. 51.44%). For Swin-B, the difference is smaller but remains positive, with STyle-TTA achieving 63.25% Top-1 accuracy compared with 60.93% for geometric TTA.

The pattern reverses for the two foundation-model backbones. For CLIP ViT-B/16 and DINOv2 ViT-B/14, geometric TTA produces the strongest results, achieving 66.15% and 69.03% Top-1 accuracy and 65.64% and 70.58% balanced accuracy, respectively. In comparison, STyle-TTA reduces performance relative to the no-TTA

baseline for both models, reaching 62.99% Top-1 accuracy for CLIP and 66.18% for DINOv2. Geometric TTA therefore provides the more reliable adaptation strategy for these pretrained foundation models.

5.5 Additional Benchmarks

To evaluate the generalization of our method beyond the datasets considered previously, we extended our benchmarking to four additional datasets: Camelyon, EpiStr, UCMerced, and MIDOG. We compare STyle-TTA against geometric TTA and the corresponding baseline across three representative backbone families.

Based on our observation that StyleID-based STyle-TTA achieves most of its effect within the first few stylized views, we reduced the number of generated stylized images from 15 to 3 per test image. This setting was chosen in preparation for the geometric hybrid approach described in Section 5.4 and is further supported by the accuracy trends observed in the ablation study in Section 5.3.

For each dataset, we evaluate the baseline, geometric TTA ($N_{\text{Views}} = 64$), and STyle-TTA ($N_{\text{Views}} = 4$). We report both Top-1 accuracy and balanced accuracy, together with the absolute change in Top-1 accuracy relative to the baseline (Δ) and its 95% confidence interval. Classifiers were trained on the full training set of $\mathcal{D}_S^{\mathcal{C}}$, while the baseline comparison is performed on the corresponding target-domain evaluation set.

As we observed architecture-specific behavior in our earlier experiments, we simplify the presentation in this section by selecting one representative model from each architectural group. For convolutional neural networks (CNNs), we report the results of DenseNet; for vision transformers (ViTs), we use Swin-B (224); and for foundation models (FMs), we use DINOv2. All other experimental settings remain unchanged.

Table 9: Evaluation of baseline and TTA methods on Camelyon. Bold indicates highest accuracy.

Backbone	**Baseline (%)**	**Strategy**	**Top-1 Acc. (%)**	**Bal. Acc. (%)**	**Δ (95% CI)**
CNN	67.95	STyle-TTA	80.40	80.40	+12.45 [10.65, 14.25]
		Geometric	**85.40** ± 0.20	**85.40** ± 0.20	+17.45 [15.65, 19.20]
ViT	92.75	STyle-TTA	**95.25**	**95.25**	+2.50 [1.55, 3.55]
		Geometric	92.65 ± 0.06	92.65 ± 0.06	−0.10 [-1.05, 0.85][1]
FM	93.00	STyle-TTA	**95.00**	**95.00**	+2.00 [1.15, 2.85]
		Geometric	93.85 ± 0.10	93.85 ± 0.10	+0.85 [0.00, 1.70][2]

Note: Unless indicated by a footnote, all unannotated changes are statistically significant with $p < 0.001$.
[1] $p = 0.916$; [2] $p = 0.068$.

Camelyon The results for Camelyon17 are shown in Table 9. STyle-TTA improves performance across all three backbone families, with the largest gain observed for the CNN, where Top-1 accuracy increases from the baseline of 67.95% to 80.40%, corresponding to a gain of 12.45 percentage points. The improvement is also consistent for the ViT and foundation model, with gains of 2.50 and 2.00 percentage points, respectively. Geometric TTA produces an even larger improvement for the CNN, reaching 85.40% and exceeding STyle-TTA by 5.00 percentage points. Meanwhile, its effect on the ViT is negligible, with performance remaining essentially unchanged relative to the baseline. For the foundation model, geometric TTA provides a smaller gain of 0.85 percentage points.

For both TTA strategies transfer effectively to Camelyon, particularly for the CNN backbone. STyle-TTA provides the strongest result among the two methods for the ViT and foundation model, whereas geometric TTA achieves the highest accuracy for the CNN.

Table 10: Evaluation of baseline and TTA methods on EpiStr. Bold indicates highest accuracy.

Backbone	**Baseline (%)**	**Strategy**	**Top-1 Acc. (%)**	**Bal. Acc. (%)**	**Δ (95% CI)**
CNN	59.45	STyle-TTA	73.69	67.36	+14.24 [12.43, 16.06]
		Geometric	**78.92** ± 0.26	**76.94** ± 0.27	+19.48 [17.37, 21.58]
ViT	75.73	STyle-TTA	**75.15**	**70.41**	−0.58 [-2.62, 1.45][1]
		Geometric	73.84 ± 0.44	67.48 ± 0.53	−1.89 [-3.27, -0.51][2]
FM	72.46	STyle-TTA	72.60	66.45	+0.15 [-1.60, 1.82][3]
		Geometric	**79.29** ± 0.32	**74.35** ± 0.37	+6.83 [5.45, 8.21]

Note: Unless indicated by a footnote, all unannotated changes are statistically significant with $p < 0.001$.
[1] $p = 0.624$; [2] $p = 0.010$; [3] $p = 0.932$.

EpiStr The results for EpiStr are shown in Table 10. STyle-TTA substantially improves CNN performance, increasing Top-1 accuracy from 59.45% to 73.69%. Geometric TTA provides a further improvement, reaching 78.92% Top-1 accuracy. For the ViT, neither method improves substantially over the baseline. STyle-TTA results in a small decrease in Top-1 accuracy of 0.58 percentage points, while geometric TTA decreases it by 1.89 percentage points. For the foundation model, STyle-TTA again has little effect (+0.15 pp), whereas geometric TTA produces a clear improvement of 6.83 percentage points in Top-1 accuracy.

UCMerced Table 11 reports the results on UCMerced. In contrast to Camelyon and EpiStr, the two TTA strategies exhibit substantially different behavior across architectures. STyle-TTA strongly improves CNN performance, increasing Top-1 accuracy by 14.67 percentage points relative to the baseline. Geometric TTA, however,

Table 11: Evaluation of baseline and TTA methods on UCMerced. Bold indicates highest accuracy.

Backbone	Baseline (%)	Strategy	Top-1 Acc. (%)	Bal. Acc. (%)	Δ (95% CI)
CNN	32.67	STyle-TTA	**47.33**	**52.67**	+14.67 [10.50, 18.83]
		Geometric	22.50 ± 0.35	27.25 ± 0.29	−10.17 [-13.83, -6.50]
ViT	53.83	STyle-TTA	48.50	68.92	−5.33 [-8.50, -2.17][1]
		Geometric	**55.50** ± 0.00	**72.08** ± 0.10	+1.67 [-1.17, 4.67][2]
FM	57.83	STyle-TTA	51.83	69.75	−6.00 [-9.17, -2.83]
		Geometric	**54.67** ± 0.51	**72.17** ± 0.42	−3.17 [-5.83, -0.33][3]

Note: Unless indicated by a footnote, all unannotated changes are statistically significant with $p < 0.001$.
[1] $p = 0.002$; [2] $p = 0.320$; [3] $p = 0.034$.

substantially degrades CNN performance, with Top-1 accuracy falling from 32.67% to 22.50%. For the ViT and foundation model, STyle-TTA decreases Top-1 accuracy relative to the baseline by 5.33 and 6.00 percentage points, respectively. Notably, despite the drop in Top-1 accuracy for these two architectures, balanced accuracy remains substantially higher than the raw Top-1 performance, reaching 68.92% for the ViT and 69.75% for the foundation model, reflecting noticeable differences in class distribution handling. Geometric TTA performs better than STyle-TTA in both cases, achieving higher Top-1 and balanced accuracies for both ViT and foundation model backbones.

Table 12: Evaluation of baseline and TTA on MIDOG. Bold indicates highest accuracy.

Backbone	Baseline (%)	Strategy	Top-1 Acc. (%)	Bal. Acc. (%)	Δ (95% CI)
CNN	40.47	STyle-TTA	**47.07**	**50.03**	+6.61 [4.64, 8.64]
		Geometric	36.71 ± 0.05	49.97 ± 0.07	−3.76 [-5.06, -2.46]
ViT	54.62	STyle-TTA	40.28	49.33	−14.34 [-16.34, -12.34]
		Geometric	**46.56** ± 0.14	**50.31** ± 0.18	−8.06 [-9.70, -6.46]
FM	52.11	STyle-TTA	**56.62**	**51.05**	+4.52 [2.58, 6.49]
		Geometric	50.05 ± 0.23	49.50 ± 0.26	−2.06 [-3.67, -0.48][1]

Note: Unless indicated by a footnote, all unannotated changes are statistically significant with $p < 0.001$.
[1] $p = 0.014$.

MIDOG The MIDOG results are presented in Table 12. Overall, the changes induced by TTA are more limited than those observed on Camelyon and EpiStr.

STyle-TTA improves CNN performance from a baseline Top-1 accuracy of 40.47% to 47.07%, corresponding to a gain of 6.61 percentage points. It also improves the foundation model, with a gain of 4.52 percentage points. Style-based STyle-TTA decreases ViT Top-1 accuracy by 14.34 percentage points. Geometric TTA produces smaller changes for the CNN and foundation model, decreasing Top-1 accuracy by 3.76 and 2.06 percentage points, respectively. For the ViT, geometric TTA performs better than STyle-TTA, although it still remains below the baseline.

5.6 Feature-Space Domain Gap

Prior to evaluating augmentation strategies, we quantify the feature-space discrepancy between source and target domains. This characterization provides context for the architectural comparison of the previous section.

5.6.1 Global Discrepancy

To quantify the feature-space discrepancy between source and target distributions, we employ the 2-Wasserstein distance (W_2). Unlike centroid-based measures, W_2 accounts for both mean displacement and differences in distribution spread by computing the minimum transportation cost between two distributions. We estimate W_2 between penultimate-layer feature distributions using a Gaussian approximation (Fréchet distance), which provides a closed-form solution while capturing second-order statistics. To isolate genuine domain shift from inherent sampling noise, we report the net domain shift ΔW_2, calculated by subtracting the in-distribution baseline (source train→val) from the out-of-distribution distance (source train→target).

Table 13 shows that supervised CNNs and Vision Transformers exhibit net shifts of 0.28–0.40, while Foundation Models show lower values (0.23–0.31) despite higher baseline accuracy.

Table 13: Feature-space discrepancy on ImageNet-1K → ImageNet-R. Net shift calculated as OOD distance minus in-domain baseline.

Family	Backbone	In-domain W_2	OOD W_2	Net Shift (ΔW_2)	Top-1 Acc. (%)
CNN	ResNet-18	0.212 ± 0.038	0.493 ± 0.088	+0.281	35.44
	DenseNet-121	0.303 ± 0.047	0.629 ± 0.099	+0.325	40.60
ViT	ViT-B/16	0.213 ± 0.050	0.604 ± 0.124	+0.391	47.89
	Swin-B	0.204 ± 0.051	0.600 ± 0.121	+0.396	60.41
FM	CLIP ViT-B/16	0.150 ± 0.043	0.378 ± 0.051	+0.228	64.23
	DINOv2 ViT-B/14	0.217 ± 0.056	0.528 ± 0.109	+0.310	69.04

5.6.2 Per-Class Gap and Performance Correlations

To analyze where domain degradation concentrates, we compute per-class Wasserstein distances d_c and correlate them with class-level baseline metrics across the 200 shared classes.

Table 14: Spearman rank correlations (ρ) between per-class domain gap and classification metrics. Baseline Gap: d_c^{val} vs. baseline metrics. Stylized Gap: $d_c^{k,\text{val}}$ (after StyleID) vs. baseline metrics. ΔGap vs. Gain: correlation between per-class Wasserstein reduction and STyle-TTA improvement. Negative values indicate inverse relationships.

Family	Backbone	Gap vs Acc	Gap vs Conf	Gap vs Ent	Stylized Gap vs Acc	ΔGap vs Gain
CNN	ResNet-18	−0.547	−0.453	+0.468	−0.547	−0.415
	DenseNet-121	−0.699	−0.656	+0.670	−0.784	−0.605
ViT	ViT-B/16	−0.721	−0.643	+0.656	−0.811	−0.413
	Swin-B	−0.678	−0.622	+0.634	−0.746	−0.412
FM	CLIP ViT-B/16	−0.450	−0.448	+0.446	−0.597	−0.314
	DINOv2 ViT-B/14	−0.600	−0.674	+0.702	−0.666	−0.302

Note: ΔGap vs Gain reports the correlation against the *ensembled* STyle-TTA accuracy gain (original image + stylized views averaged), not the style-only gain discussed separately in the prose below. The two are computed from the same per-class Wasserstein reduction but correlated against different performance deltas, and can differ noticeably for a given backbone (e.g. CLIP ViT-B/16: −0.314 ensembled vs. −0.498 style-only).

Table 14 shows that per-class domain gap correlates negatively with baseline accuracy across all backbones ($\rho = -0.450$ to -0.721). After stylization, this correlation strengthens for most backbones (e.g., DenseNet-121: $-0.699 \rightarrow -0.784$; ViT-B/16: $-0.721 \rightarrow -0.811$). The reduction in per-class gap correlates negatively with STyle-TTA gains ($\rho = -0.302$ to -0.605), indicating that larger gap reduction is associated with larger accuracy improvements.

Feature-Space Stylization To verify that StyleID's accuracy gains correspond to actual feature-space alignment rather than ensemble averaging alone, we measured the per-class Wasserstein distance (W_2) between target features and the source distribution before and after stylization.

For texture-biased backbones, the raw domain gap is strongly predictive of where STyle-TTA helps most: under W_2, the per-class out-of-domain gap correlates negatively with baseline accuracy for ResNet-18 ($\rho = -0.547$), DenseNet-121 ($\rho = -0.699$), and ViT-B/16 ($\rho = -0.721$), indicating that classes suffering the largest appearance-driven displacement are precisely those with the worst pre-augmentation performance. Stylization effectively reduces this domain gap ($\Delta W_2 < 0$), where a larger reduction in feature distance strongly drives higher *style-only* STyle-TTA accuracy gains (i.e. gains from the stylized view alone, prior to ensembling with the

original image) across ResNet-18 ($\rho = -0.441$), DenseNet-121 ($\rho = -0.644$), and ViT-B/16 ($\rho = -0.435$). This style-only correlation is distinct from, and in most cases larger in magnitude than, the ensembled ΔGap vs Gain figures reported in Table 14.

For DINOv2, while the baseline out-of-domain gap also correlates negatively with baseline accuracy ($\rho = -0.600$), stylization operates differently: feature distance reduction correlates with style-only gains ($\rho = -0.520$, again the style-only rather than ensembled metric), yet yields little net performance benefit. This confirms that style transfer effectively benefits classifiers whose domain gap is dominated by correctable appearance statistics, but provides minimal benefit when the backbone feature space is already invariant to low-level style shifts.

5.7 Calibration and Confidence Quality

Beyond predictive accuracy, test-time augmentation may affect the quality of the confidence estimates produced by the model. We therefore evaluate the calibration and confidence-ranking behaviour of the different TTA strategies using Expected Calibration Error (ECE) and the area under the receiver operating characteristic curve (AUC). Lower ECE indicates better agreement between predicted confidence and empirical accuracy, whereas higher AUC indicates better separation between correct and incorrect predictions based on the confidence scores. The two metrics therefore capture complementary aspects of confidence quality.

Table 15 reports the results on the ImageNet-1K $\rightarrow$ ImageNet-R benchmark at the ensemble sizes selected in the preceding experiments. For each backbone, the TTA strategy is evaluated at its respective best-performing view count: $N = 16$ for STyle-TTA and $N = 64$ for Geometric TTA.

Table 15: Calibration and confidence-ranking performance of test-time augmentation strategies on ImageNet-1K → ImageNet-R. Each strategy is evaluated at its best-performing view count (STyle-TTA: $N = 16$, Geometric: $N = 64$). Lower ECE indicates better calibration, while higher AUC indicates better confidence-based ranking. Bold marks the best result per backbone and metric.

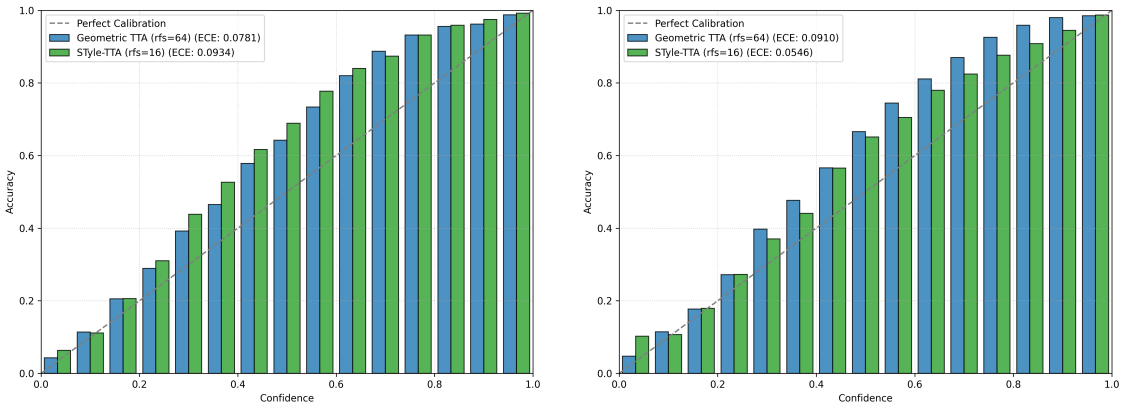
Backbone	**Strategy**	**ECE ↓**	**AUC ↑**
ResNet-18	STyle-TTA	0.05	**0.94**
	Geometric	**0.04**	0.92
DenseNet-121	STyle-TTA	0.09	**0.95**
	Geometric	**0.08**	0.93
ViT-B/16 (224)	STyle-TTA	0.13	**0.97**
	Geometric	**0.11**	0.95
Swin-B (224)	STyle-TTA	**0.13**	**0.98**
	Geometric	0.19	0.97
CLIP ViT-B/16	STyle-TTA	**0.07**	0.98
	Geometric	0.13	**0.99**
DINOv2 ViT-B/14	STyle-TTA	**0.05**	0.98
	Geometric	0.09	**0.98**

The results reveal a clear distinction between predictive performance and confidence quality. Geometric TTA achieves the lowest ECE for ResNet-18, DenseNet-121, and ViT-B/16, with ECE values of 0.04, 0.08, and 0.11, respectively. For the three remaining backbones, STyle-TTA provides the lowest ECE, achieving values of 0.13 for Swin-B, 0.07 for CLIP ViT-B/16, and 0.05 for DINOv2 ViT-B/14. Thus, neither style-based nor geometric augmentation consistently produces the best-calibrated predictions across backbone architectures.

The AUC results show a somewhat different pattern. STyle-TTA achieves the highest AUC for ResNet-18, DenseNet-121, ViT-B/16, and Swin-B, with values ranging from 0.94 to 0.98. For CLIP ViT-B/16, Geometric TTA obtains the highest AUC of 0.99, while STyle-TTA and Geometric TTA are tied at 0.98 for DINOv2. This indicates that STyle-TTA can provide better separation between correct and incorrect predictions even in cases where Geometric TTA produces slightly better-calibrated confidence values.

The difference between the two metrics is particularly apparent for ResNet-18 and ViT-B/16. For ResNet-18, Geometric TTA has the lowest ECE (0.04), whereas STyle-TTA achieves the highest AUC (0.94). Similarly, for ViT-B/16, Geometric TTA improves ECE from 0.13 with STyle-TTA to 0.11, while STyle-TTA achieves a higher AUC of 0.97 compared with 0.95 for Geometric TTA. Consequently, the strategy that produces confidence values that are best calibrated is not necessarily the strategy that ranks correct predictions most effectively.

To provide a visual assessment of calibration, Figure 10 shows reliability diagrams for the evaluated TTA strategies. Each diagram compares predicted confidence with empirical accuracy across confidence bins. The diagonal represents perfect calibration, where predicted confidence matches observed accuracy. The reliability diagrams provide a visual comparison of the confidence estimates for DenseNet-121 and DINOv2. For DenseNet-121, the STyle-TTA and Geometric TTA curves exhibit different calibration behaviour across confidence bins. For DINOv2, the relative position of the two curves differs from that observed for DenseNet-121. These visual differences complement the ECE values reported in Table 15.



(a) DenseNet-121 (b) DINOv2 ViT-B/14

Figure 10: Reliability diagram for DenseNet-121 and DINOv2 comparing Geometric TTA ($N = 64$) and Style-TTA ($N = 16$). Accuracy per confidence bin is computed across 15 bins using pooled predictions across seeds, alongside the aggregated Expected Calibration Error (ECE).

5.8 Ablation Studies

The following experiments isolate the contribution of three pipeline components: (i) style-reference retrieval, (ii) number of generated views, and (iii) prediction aggregation. Each ablation changes one factor of the implementation pipeline while holding all others fixed. This is done in an effort to observe the effect degree the pipeline variables have on the success of the method. To ensure evaluations remain invariant to potential class imbalances in out-of-domain test data, we prioritize balanced accuracy as our primary classification performance metric.

5.8.1 Style Reference Retrieval Strategy

To analyse the effect of the style-reference images has on performance we compare three retrieval-strategies for selecting source-domain style references:

1. **Random retrieval:** Uniform random sampling from the source style pool.
2. **Balanced random retrieval:** Random sampling constrained to maintain an approximately balanced class distribution in the retrieved reference images.
3. **DINO-based retrieval:** Nearest-neighbour selection using L2-normalized DINOv3 embeddings (Sec. 4.2).

We created two additional stylized test view sets based on random and balanced random retrieved reference images. To limit computational cost, AdaIN was evaluated with 31 generated stylized views for random and balanced random(32 total with the original image). For this experiment we use AdaIN as the style operator, as we create 31 images per test-image of ImageNet-R (2,880,000 additional images). All other parameters were held constant: probability averaging for aggregation, and the same source-domain pool.

Results. DINO-based retrieval achieves the highest balanced accuracy for every backbone (Table 16). Gains over random retrieval range from $\sim$5 pp (DINOv2) to $\sim$9 pp (ResNet-18); balanced-random retrieval performs almost identically to unbalanced random retrieval, indicating that class balancing alone does not explain the improvement. How the retrieval method influences true-label dominance can be found in the Appendix A.3. These results demonstrate that retrieving semantically compatible references is critical: arbitrary style transfer hurts performance, whereas semantic-neighbour style transfer helps.

Calibration follows the opposite trend: DINO-based retrieval produces the highest ECE for all six backbones. For example, ResNet-18 improves from 30.10% to 39.37% balanced accuracy when switching from random to DINO retrieval, but ECE increases from 0.249 to 0.312. The accuracy gains motivate DINO-based retrieval as the default strategy, but the calibration cost should be kept in mind when deploying in safety-critical settings.

Looking at Figure 11 as an example where conceptually less relevant style images cause a breakdown in feature structure of the content image while also moving the test image further away from its label space. Given that DINO-based retrieval is beneficial, we next ask how many stylized views are needed to realise these gains.

Table 16: Evaluation of source style reference retrieval strategies for STyle-TTA. Performance on ImageNet-R across six ImageNet-1K classifiers comparing Random, Class-Balanced Random, and DINO-based nearest-neighbourretrieval ($N_{\text{views}} = 64$, AdaIN operator). Δ Acc reports the absolute change relative to the unadapted baseline with 95% bootstrap confidence intervals. Bold values indicate optimal performance per backbone.

Backbone	Retrieval	Balanced Acc. (%) ↑	Δ Acc. (95% CI)	ECE ↓
ResNet-18	Random	30.10	-5.33 [-5.84, -4.82]	**0.249**
	Balanced	29.30	-6.14 [-6.66, -5.64]	0.251
	DINO	**39.37**	**+3.93 [+3.44, +4.41]**	0.312
DenseNet-121	Random	34.11	-6.48 [-7.01, -5.96]	**0.235**
	Balanced	32.98	-7.62 [-8.15, -7.08]	0.242
	DINO	**42.48**	**+1.88 [+1.38, +2.38]**	0.290
ViT-B/16	Random	47.03	-0.85 [-1.37, -0.31]	0.164
	Balanced	47.06	-0.83 [-1.36, -0.28]	**0.161**
	DINO	**52.45**	**+4.57 [+4.06, +5.07]**	0.213
Swin-B	Random	53.60	-6.81 [-7.33, -6.27]	**0.165**
	Balanced	52.93	-7.48 [-7.99, -6.95]	0.171
	DINO	**59.25**	**-1.16 [-1.63, -0.67]**	0.193
CLIP ViT-B/16	Random	54.03	-10.20 [-10.77, -9.64]	**0.157**
	Balanced	53.77	-10.47 [-11.05, -9.90]	0.158
	DINO	**60.47**	**-3.77 [-4.31, -3.23]**	0.174
DINOv2 ViT-B/14	Random	60.59	-8.44 [-8.95, -7.93]	**0.186**
	Balanced	60.59	-8.45 [-8.95, -7.96]	**0.186**
	DINO	**65.32**	**-3.72 [-4.18, -3.24]**	0.195



Figure 11: Stylized outcomes for different retrieval methods. The left image uses a style reference retrieved by DINOv3, whereas the right image is a randomly selected reference. Style source is ImageNet-1K while content source is ImageNet-R

5.8.2 Number of Generated Views

To understand how the number of stylized views affects STyle-TTA performance, we first examine the informativeness of the individual generated views before evaluating them cumulatively as an ensemble. This distinction is important because adding more views can only be beneficial if the additional views provide useful and sufficiently complementary predictions. We therefore first evaluate each stylized view independently, followed by an analysis of the ensemble performance as progressively more views are aggregated.

Standalone performance of stylized views. Each stylized view is first treated as an independent prediction, without information from the other views. This provides a baseline for assessing whether individual stylized transformations remain informative despite potentially introducing substantial visual changes.

Figure 12 shows the Top-1 accuracy obtained when each stylized view is evaluated independently on ImageNet. Across all three backbones, the first stylized view achieves the highest accuracy among the generated views, after which performance generally remains within a relatively narrow range. For DenseNet-121, accuracy decreases from 41.41% for the first view to 39.61% for the fifteenth view. A similar pattern is observed for Swin-B, with accuracy decreasing from 54.42% to 52.12%, while DINOv2 decreases from 59.90% to 57.62%.

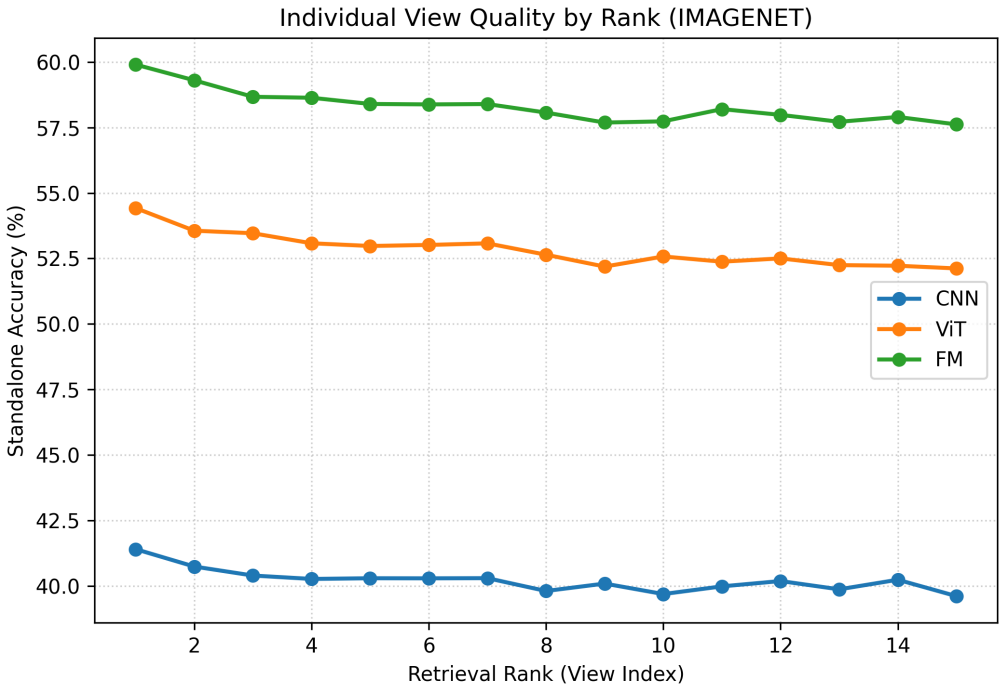


Figure 12: Top-1 accuracy of individual stylized views of ImageNet-R evaluated independently of the ensemble.

The relatively small degradation across view rank suggests that the generated views remain individually informative throughout the sequence. However, individual-view accuracy alone does not determine whether additional views improve test-time prediction. To assess this, we next vary the number of stylized views aggregated at inference time.

Ensemble performance as the number of views increases. We vary the number of stylized views aggregated at inference time, $N \in 2, 4, 8, 16, 32, 64$. Each condition contains the original target image together with $N - 1$ generated stylized views, and predictions are mean-averaged across all views. Due to the computational budget, AdaIN is evaluated for all view counts, while StyleID is evaluated up to $N = 16$. For visualization, we use our architectural representatives: DenseNet for CNNs, Swin for ViTs, and DINOv2 for foundation models. The complete results for each backbone are reported in Appendix A.5.

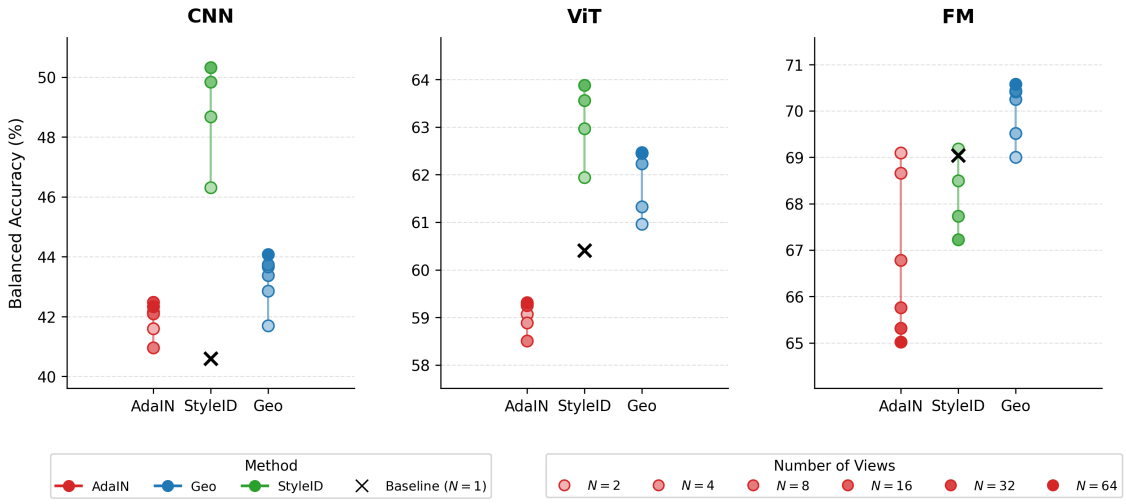


Figure 13: Effect of the number of created style-augmented STyle-TTA views (N) on TTA performance across representative classifier architectures against standard geometric TTA. Results are shown for DenseNet (CNN), Swin (ViT), and Swin (FM). The black cross denotes the corresponding baseline performance without style augmentation ($N = 1$). StyleID-based STyle-TTA is evaluated up to $N = 16$ the other up to $N = 64$.

Results. Figure 13 reveals substantially different effects of increasing the view count across architectures and style operators. StyleID-based STyle-TTA outperforms AdaIN and geometric TTA for both CNN and ViT, achieving higher accuracy with a comparatively small view budget.

For the CNN, StyleID already outperforms AdaIN with a single additional stylized view ($N = 2$), achieving 46.31% compared with 41.60%, a gain of 4.71 percentage points. It also exceeds geometric TTA by 4.62 percentage points (46.31% vs. 41.69%). For the ViT, AdaIN does not benefit from increasing the number of

views and ultimately falls below the baseline by 1.10 percentage points at $N = 64$ (59.31% vs. 60.41%). In contrast, StyleID exceeds the performance of geometric TTA using only eight views, achieving 63.56% compared with 62.45% for geometric TTA at $N = 64$.

The behavior differs for the foundation model. Here, geometric TTA is the only method that consistently benefits from increasing the number of views. With one additional stylized view ($N = 2$), both style operators perform close to the 69.04% baseline, with all methods remaining within 0.14 percentage points of the baseline. As more stylized views are aggregated, however, STyle-TTA performance deteriorates. AdaIN drops by 4.02 percentage points at $N = 64$, while StyleID drops by 1.81 percentage points at $N = 16$.

Standard geometric TTA exhibits relatively consistent behavior across the three architectures, with additional views generally providing steady improvements. In contrast, the benefit of stylized views is strongly architecture-dependent. The optimal number of views therefore varies considerably, ranging from very few views for the foundation model to substantially larger ensembles for the CNN and ViT.

5.8.3 Prediction Aggregation Strategy

Finally, we evaluate the prediction aggregation strategy, i.e., how the predictions from the generated views are combined into a final output. This is particularly relevant for style-based TTA because the generated views are not necessarily equally informative for the classifier. Although the style transformations are intended to preserve the semantic content of the original instance, individual stylized views may vary in how well they preserve features that are useful for classification. Consequently, averaging all views implicitly assumes that each generated view provides equally reliable evidence for the final prediction. We compare three aggregation modes: vanilla mean probability averaging, and two confidence-based aggregation strategies based on ZERO and TPT.

TPT. The TPT condition adapts the confidence-selection component of Test-Time Prompt Tuning (Shu et al., 2022) to conventional classifiers. No prompt or model parameters are optimized at test time. Instead, the predictive entropy of each view is calculated as

$$H(p_i) = -\sum_c p_{i,c} \log(p_{i,c}), \tag{27}$$

where p_i denotes the softmax probability distribution for view i. The views are ranked by entropy, with lower entropy corresponding to higher predictive confidence. For a total of N views, the lowest-entropy

$$k = \max(1, \lfloor N\gamma \rfloor) \tag{28}$$

views are retained, where $\gamma = 0.1(Top - 10\%)$ denotes the filtering fraction. The softmax probability distributions of the retained views are then averaged at the

classifier's native temperature. Thus, the TPT condition performs entropy-based view selection followed by soft aggregation.

ZERO. ZERO applies the same entropy-based view selection but differs in the subsequent aggregation step (Farina et al., 2024). After retaining the k lowest-entropy views, their logits are converted into an effectively zero-temperature softmax distribution. In the limit of zero temperature, this assigns the probability mass to the highest-logit class of each view, thereby converting each retained prediction into a hard class vote. The votes are summed across the selected views and normalized to obtain the final prediction distribution. Consequently, ZERO can be expressed as a majority-vote aggregation over the classes predicted by the selected, highest-confidence views. Both confidence-based methods therefore use the same entropy criterion to select views but differ in how the selected predictions are aggregated: TPT retains their probability information through soft averaging, whereas ZERO reduces them to class-level votes.

This comparison addresses a complementary question to the preceding experiments on view generation and selection. The earlier experiments determine which stylized views should be generated, whereas this experiment asks whether all generated views should contribute equally to the final prediction. All three aggregation modes use the same generated views and are evaluated under identical conditions, with StyleID and $N_{\text{refs}} = 16$. Consequently, differences between the conditions can be attributed to the aggregation strategy rather than to differences in the generated view ensemble.

Table 17: Comparison aggregation modes (`vanilla`, `zero`, `tpt`) for StyleID ($N_{\text{refs}} = 16$) on ImageNet-R.

Classifier	**Method**	**Bal Acc (%) ↑**	**AUC ↑**	**ECE ↓**
CNN	Vanilla	50.32	**0.9478**	**0.0934**
	Zero	50.29	0.7980	0.2658
	TPT	**50.55**	0.9215	0.2817
ViT	Vanilla	**63.88**	**0.9768**	**0.1338**
	Zero	63.47	0.8601	0.1770
	TPT	62.68	0.9465	0.2360
FM	Vanilla	67.23	**0.9836**	**0.0546**
	Zero	68.59	0.8793	0.1895
	TPT	**69.57**	0.9734	0.1991

Results. The balanced accuracy results differ across the three classifier categories (Table 17). For the CNN, TPT achieves the highest balanced accuracy at 50.55%, compared with 50.32% for vanilla averaging and 50.29% for ZERO. For the ViT, vanilla averaging achieves the highest value at 63.88%, followed by ZERO at 63.47% and TPT at 62.68%. For the Foundation Model, TPT achieves the highest balanced accuracy at 69.57%, followed by ZERO at 68.59% and vanilla averaging at 67.23%. The AUC results show a different ordering. Vanilla averaging achieves the highest AUC for all three classifier categories, with values of 0.9478 for the CNN, 0.9768 for the ViT, and 0.9836 for the Foundation Model. TPT obtains AUC values of 0.9215, 0.9465, and 0.9734, respectively, while ZERO obtains 0.7980, 0.8601, and 0.8793. The same ordering is observed for ECE. Vanilla averaging produces the lowest ECE for all three classifier categories, with values of 0.0934 for the CNN, 0.1338 for the ViT, and 0.0546 for the Foundation Model. ZERO produces ECE values of 0.2658, 0.1770, and 0.1895, while TPT produces 0.2817, 0.2360, and 0.1991, respectively. In summary, TPT obtains the highest balanced accuracy for two of the three classifier categories, while vanilla averaging obtains the highest AUC and lowest ECE for all three categories. ZERO does not achieve the highest value for any of the three reported metrics across the classifier categories.

Oracle Analysis. The ensemble accuracy and corresponding oracle upper bound were evaluated for increasing numbers of stylized views in Table 18. Increasing the view count creates a pronounced divergence between actual ensemble performance and the theoretical oracle ceiling. For DenseNet-121, while standard aggregation yields steady gains from 44.63% ($N = 2$) to 49.00% ($N = 16$), the oracle accuracy surges from 49.29% to 66.81%. This expands the performance gap nearly fourfold (from 4.66 pp to 17.82 pp), indicating that while individual generated views frequently contain the correct visual signals, naïve probability averaging increasingly fails to isolate them as view counts scale.A similar trend emerges for Swin-B and DINOv2. Swin-B achieves an ensemble accuracy of 62.69% at 16 views against an oracle ceiling of 78.39% (a 15.70 pp gap), whereas DINOv2 reaches 65.38% ensemble accuracy compared to a 79.27% oracle bound (a 13.90 pp gap). Notably, DINOv2 exhibits the smallest relative gap at lower view counts (3.81 pp at $N = 2$), underscoring the higher baseline reliability of its individual view representations. These growing gaps demonstrate that the primary bottleneck in scaling stylized TTA is not a lack of informative views, but rather the limitation of standard aggregation mechanisms in retrieving correct predictions from noisy ensemble sets.

Prediction Agreement and Diversity.

Prediction Agreement and Diversity. To evaluate how stylized augmentations influence ensemble behavior, we analyze prediction consistency across the 15 generated views alongside pairwise distributional diversity (Table 19). Model representations exhibit a clear gradient in stability moving from convolutional to vision

Table 18: Ensemble accuracy and oracle upper bound for increasing numbers of stylized views.

Backbone	**Views**	**Ensemble Top-1 (%)**	**Oracle Top-1 (%)**	**Gap (pp)**
DenseNet-121	Top-2	44.63	49.29	4.66
	Top-4	46.76	55.95	9.19
	Top-8	48.30	62.12	13.82
	Top-16	49.00	66.81	17.82
Swin-B	Top-2	57.78	62.29	4.51
	Top-4	60.35	69.19	8.84
	Top-8	61.90	74.43	12.52
	Top-16	62.69	78.39	15.70
DINOv2	Top-2	62.80	66.62	3.81
	Top-4	64.30	71.81	7.50
	Top-8	65.10	76.00	10.90
	Top-16	65.38	79.27	13.90

transformer and foundation architectures. DenseNet-121 displays the lowest consensus, reaching full 15-view agreement on only 17.11% of test instances, while exhibiting high distributional dispersion (symmetric KL divergence of 1.0123). Swin-B exhibits moderate stability with a 26.66% agreement rate and a KL divergence of 0.9267. Conversely, DINOv2 yields the highest prediction stability, maintaining complete consensus across all views for 35.73% of samples and recording the lowest KL divergence (0.7874) and cosine similarity (0.2342).

Table 19: Prediction agreement and pairwise diversity across 15 stylized views.

Backbone	**Full Agreement (%)**	**Avg. Sym. KL**	**Avg. Cosine**
DenseNet-121	17.11	1.0123	0.3631
Swin-B	26.66	0.9267	0.3182
DINOv2	35.73	0.7874	0.2342

To determine whether instance-level view diversity drives accuracy improvements, we evaluate the correlation between these diversity metrics and performance gains (Table 20). Across all backbones, both Pearson and Spearman correlation coefficients remain negligible, ranging between -0.094 and $+0.070$. This near-zero correlation demonstrates that prediction diversity alone is not a reliable predictor of test-time augmentation success; generating highly diverse views provides no guarantee of improved accuracy unless those views preserve semantically meaningful features.

Table 20: Correlation between prediction diversity measures and accuracy gain.

Measure	**Backbone**	**Pearson**	**Spearman**
Symmetric KL	DenseNet-121	+0.064	+0.070
	Swin-B	+0.031	+0.040
	DINOv2	-0.055	-0.062
Cosine Similarity	DenseNet-121	-0.006	-0.003
	Swin-B	-0.003	+0.005
	DINOv2	-0.094	-0.089

Ablation Summary In summary, the ablation studies reveal three practical principles for style-based test-time augmentation. First, semantically guided style retrieval (via DINO embeddings) is essential for accuracy gains, whereas non-semantic sampling introduces uninformative noise. Second, the optimal number of generated

views depends strongly on the backbone: CNN and ViT continuously benefit from larger view counts ($N = 16$–64), whereas Foundation Models reach peak performance with minimal view counts ($N = 2$) and are hurt by additional style-transfered views. Third, while confidence-based aggregation strategies like TPT and ZERO can yield marginal accuracy gains on Foundation Models, vanilla probability averaging remains the most reliable default due to its consistently superior calibration and AUC metrics across all architectures. However, the widening gap to the oracle bound indicates that developing selective aggregation methods capable of isolating correct predictions remains a critical open challenge.

5.9 Computational Cost

While overall accuracy gains are positive, the matter of computational cost must be explicitly addressed. For real-world test scenarios, execution latency and memory overhead are as critical as predictive performance. We evaluate the three transformation operators—standard geometric, AdaIN, and StyleID—on ImageNet-R using a ResNet-18 backbone at 224×224 resolution. Latency and peak VRAM consumption were measured over 97 samples evaluated on a single NVIDIA RTX A5000 GPU. Style transfer was performed using the random-retrieval configuration; feature-based retrieval adds negligible overhead (< 1 ms) given a pre-cached style pool.

Table 21 lists the resulting per-view latencies and memory footprints. Style transfer remains the main computational bottleneck: a single StyleID view generation requires $\sim$964.20 ms, whereas AdaIN requires $\sim$42.49 ms and standard geometric augmentation takes only $\sim$0.11 ms. Evaluation latency is low across all methods ($\sim$0.31–0.49 ms per view) owing to the lightweight classifier architecture.

Table 21: Per-view latency breakdown and peak memory consumption (ResNet-18, 224×224, NVIDIA RTX A5000). Reporting Augmentation Latency and Evaluation Latency per view, summarized in Total Per-View Latency, as well as peak VRAM Usage at inference.

Method	**Aug. [ms/view]**	**Eval. [ms/view]**	**Total [ms/view]**	**Peak VRAM [MB]**
Geometric TTA	0.11	0.31	0.42	82.5
STyle-TTA (AdaIN)	42.49	0.49	42.97	295.3
STyle-TTA (StyleID)	964.20	0.38	964.58	2594.5

Extrapolating these figures to a full test-time ensemble of $N = 32$ views per sample, pure StyleID TTA requires approximately 30.9 s per sample, whereas pure AdaIN TTA takes 1.38 s, and pure geometric TTA completes in just 13.4 ms. Peak VRAM is heavily dictated by the generative backbone: $\sim$2.6 GB for StyleID-based pipelines versus < 0.3 GB for AdaIN or geometric setups. This sharp trade-off between expressive style transfer and runtime efficiency motivates hybrid configurations that selectively combine geometric and style-based transformations within a single test-time ensemble.

5.10 Hybrid Approach

The preceding experiments characterize the main design choices of the proposed STyle-TTA pipeline. The view-generation experiments show that style-based transformations can provide useful variation beyond the original input, while the view-count analysis indicates that the benefit of additional stylized views depends on the classifier architecture. The aggregation experiment further shows that the manner in which these views are combined constitutes a separate design choice, with vanilla probability averaging providing a strong general-purpose baseline. These findings also highlight that stylized views and conventional geometric views address different sources of variation. The two types of transformations are therefore not mutually exclusive: they can generate complementary views of the same test instance. A second consideration is computational cost. As shown in the preceding analysis, style transfer is substantially more expensive than conventional geometric augmentation. Generating a large ensemble exclusively through style transfer therefore incurs a considerable inference-time cost, whereas geometric views can be generated at negligible additional cost.

This motivates the final question of the experimental study: rather than choosing between geometric and style-based augmentation, can their respective sources of variation be combined within a single test-time ensemble? The following experiment therefore evaluates hybrid configurations that combine a limited number of computationally expensive stylized views with computationally inexpensive geometric transformations. Two strategies are compared in this section as illustrated in Figure 14. **Hybrid S1** applies all geometric views to the original image only, yielding an ensemble of $\{x, T_{\text{Geo}}(x), T_{\text{ST}}(x, s_i)\}$. This configuration tests whether the contribution of style-based views can be combined with a conventional geometric TTA ensemble without increasing the number of style-transfer operations. **Hybrid S2** distributes the geometric budget across both the original and each stylized view, yielding $\{x, T_{\text{Geo}}(x), T_{\text{ST}}(x, s_i), T_{\text{geo}}(T_{\text{style}}(x, s_i))\}$. This configuration tests whether spatial variation can be introduced independently across the different stylistic realizations of the input. In this sense, S2 explores a larger transformation space than S1 by combining appearance variation with spatial variation within each stylized branch.

All evaluations use the same pipeline as before: DINOv2-based retrieval, StyleID as the style operator, and mean probability aggregation. The geometric baseline is a standard TTA pipeline with random resized crops and horizontal flips. As the goal is to limit the number of needed stylized views, we only use the first three stylised images per test sample for all benchmarks including ImageNet.

Multi-dataset overview. Table 22 reports balanced accuracy for the baseline classifier, pure geometric TTA, pure STyle-TTA, Hybrid S1, and Hybrid S2 across five domain-shift datasets and three backbone architectures. The geometric and hybrid configurations use $N = 64$ views, while STyle-TTA is evaluated at $N = 16$.

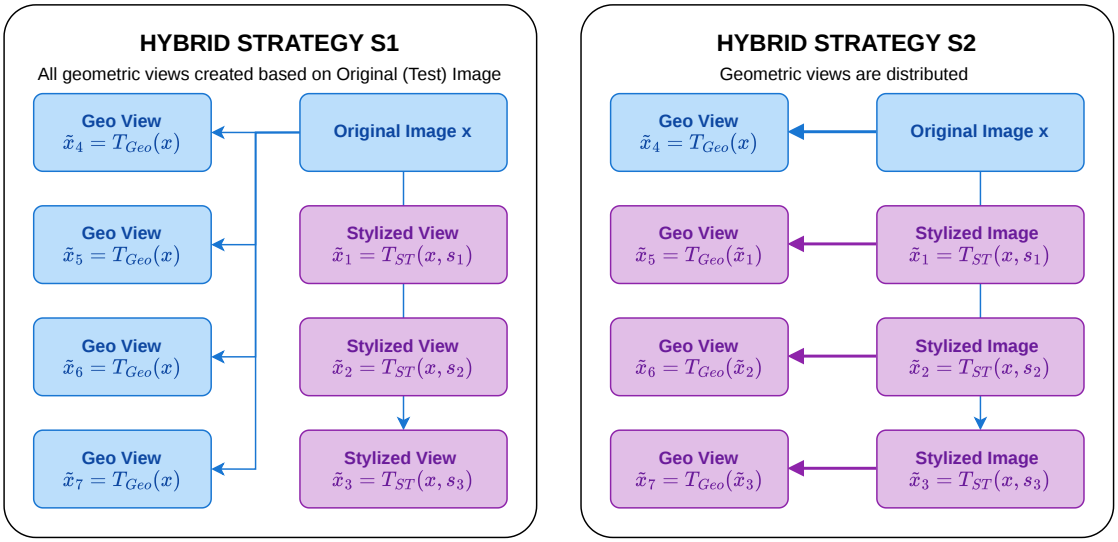


Figure 14: Construction of the View Ensemble under Hybrid Strategies S1 and S2 ($N_{\text{Views}} = 8$). The $T(\cdot)$ either a geometric transformation or a style-based one.

On MIDOG, none of the hybrid configurations improves upon the geometric baseline. Hybrid S2 remains close to the geometric baseline for DenseNet-121 (−0.15 pp), but underperforms it for Swin-B (−5.30 pp) and DINOv2 (−7.79 pp). STyle-TTA achieves the highest balanced accuracy for DenseNet-121 and DINOv2, while Hybrid S1 performs best among the hybrid configurations for Swin-B.

On UCMerced, Hybrid S2 achieves a substantial improvement over geometric TTA for DenseNet-121 (+24.17 pp), approaching the performance of STyle-TTA. For Swin-B and DINOv2, however, Hybrid S2 remains below geometric TTA by −8.00 pp and −6.33 pp, respectively. Hybrid S1 performs similarly to geometric TTA for Swin-B and DINOv2, but does not exceed the geometric baseline.

On Camelyon17-WILDS, Hybrid S2 achieves the highest balanced accuracy for DenseNet-121 and DINOv2, with gains over geometric TTA of +7.15 pp and +1.45 pp, respectively. For Swin-B, STyle-TTA achieves the highest performance, while Hybrid S2 still improves upon the geometric baseline by +2.00 pp.

On EpiStr, Hybrid S2 provides the strongest performance for DenseNet-121, improving upon geometric TTA by +6.54 pp. For Swin-B, Hybrid S2 provides only a marginal improvement of +0.51 pp, while for DINOv2 it remains below geometric TTA by −0.87 pp. STyle-TTA performs competitively for Swin-B but does not exceed the geometric baseline for DINOv2.

On ImageNet, Hybrid S2 improves upon geometric TTA for DenseNet-121 (+7.31 pp) and Swin-B (+2.13 pp), while remaining below geometric TTA for DINOv2 (−1.41 pp). Hybrid S1 achieves the highest performance for DINOv2, with a marginal gain of +0.15 pp over geometric TTA. STyle-TTA achieves the highest balanced accuracy for DenseNet-121 and Swin-B, but remains below the geometric baseline for DINOv2.

Overall, the results indicate that hybrid augmentation is most beneficial for DenseNet-121, particularly on UCMerced, Camelyon17-WILDS, EpiStr, and ImageNet. Hybrid S2 generally provides the strongest hybrid configuration, although its effectiveness depends on the backbone and dataset. For DINOv2, geometric TTA or Hybrid S1 is generally preferable to Hybrid S2, suggesting that distributing geometric transformations across stylized views is not universally beneficial.

Table 22: Evaluation of baseline ImageNet-R and TTA methods across CNN, Vision Transformer, and Foundation Model backbones. Ensemble size per method was $N = 64$ for STyle-TTA it's $N = 16$. Bold indicates the best performance per metric.

Dataset	Backbone	Baseline	Geo-TTA	STyle-TTA	Hybrid S1	Hybrid S2	Δ S2 vs Geo
MIDOG	CNN	40.47	36.71	**47.07**	36.71	36.56	-0.15
	ViT	54.62	46.56	40.28	**47.07**	41.25	-5.30
	FM	52.11	50.05	**56.62**	49.20	42.26	-7.79
UCMerced	CNN	32.67	22.50	**47.33**	23.67	46.67	+24.17
	ViT	53.83	**55.50**	48.50	55.00	47.50	-8.00
	FM	57.83	**54.67**	51.83	54.17	48.33	-6.33
Camelyon	CNN	67.95	85.40	80.40	85.75	**92.55**	+7.15
	ViT	92.75	92.65	**95.25**	93.15	94.65	+2.00
	FM	93.00	93.85	95.00	93.85	**95.30**	+1.45
EpiStr	CNN	59.45	78.92	73.69	79.51	**85.47**	+6.54
	ViT	75.73	73.84	**75.15**	74.56	74.35	+0.51
	FM	72.46	79.29	72.60	**79.58**	78.42	-0.87
ImageNet	CNN	38.60	41.90	**49.24**	42.75	49.21	+7.31
	ViT	58.79	60.93	**63.25**	61.89	63.06	+2.13
	FM	67.47	69.03	66.18	**69.18**	67.63	-1.41

Effects of Ensemble Size. The preceding view-count analysis showed that the effect of increasing the number of views is backbone-dependent, with some architectures benefiting from additional observations while others degrade as more stylized views are introduced. We therefore examine whether the same behaviour occurs for the hybrid strategy, and whether the allocation of stylized views within the ensemble affects this dependence.

We fix the number of stylized images to three and vary the total ensemble size $N \in \{8, 16, 32, 64\}$ and compare the Hybrid S1 and Hybrid S2. We use representatives again with DenseNet being CNN, Swin for ViT and DINO as FM. Figure 15 illustrates the top-1 accuracy on ImageNet-R, together with the change relative to the corresponding ImageNet baseline. This evaluation isolates the effect of ensemble size and geometric-view allocation while keeping the style-transfer budget fixed.

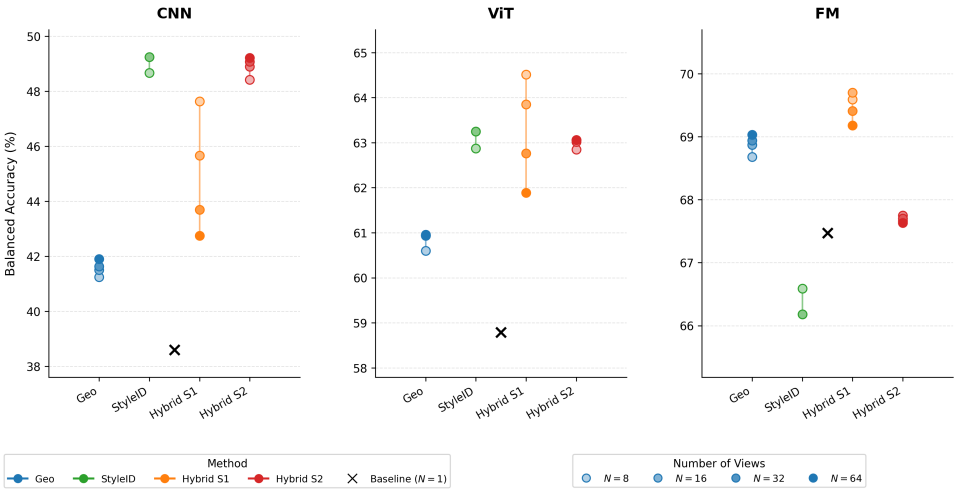


Figure 15: Top-1 accuracy (%) on ImageNet-R across varying ensemble sizes $N \in \{8, 16, 32, 64\}$ using a fixed budget of three stylized views. The x marks the baseline accuracy relative to ImageNet.

Results. The two hybrid configurations exhibit markedly different scaling behaviour across the evaluated backbones.

DenseNet-121 exhibits a similar pattern. Hybrid S1 declines from 47.63% at $N = 8$ to 42.75% at $N = 64$, whereas Hybrid S2 remains stable, increasing from 48.42% to 49.21% over the same range. Hybrid S2 outperforms Hybrid S1 at every ensemble size and approaches the performance of STyle-TTA.

Swin-B shows a different pattern. Hybrid S1 achieves its highest balanced accuracy at $N = 8$ (64.51%), after which performance declines to 61.89% at $N = 64$. Hybrid S2 remains relatively stable, ranging from 62.85% at $N = 8$ to 63.06% at $N = 64$. Although Hybrid S1 performs best at the smallest ensemble size, Hybrid S2 provides more stable scaling as the ensemble grows.

DINOv2 exhibits a similar preference for Hybrid S1. Hybrid S1 achieves its highest balanced accuracy at $N = 16$ (69.70%), while remaining above geometric TTA across all ensemble sizes. Its performance then gradually declines to 69.18% at $N = 64$. Hybrid S2 remains close to the baseline and does not improve over geometric TTA, ranging from 67.66% at $N = 8$ to 67.63% at $N = 64$. Calibration information can be found in the Appendix A.6 This is consistent with the earlier view-count results, where DINOv2 showed limited tolerance for additional style-conditioned variation.

6 Discussion

Central Empirical Finding The central empirical finding of this thesis is that style-based test-time augmentation (STyle-TTA) substantially improves classification performance under domain shift in several key settings, frequently outperforming conventional geometric test-time augmentation. On the ImageNet-1K → ImageNet-R benchmark, this effect is strongest for conventional convolutional architectures and supervised Vision Transformers, where STyle-TTA yields gains of 4–12 percentage points in Top-1 accuracy over the unaugmented baseline and clear advantages over geometric TTA.

However, cross-benchmark analysis reveals that the relative superiority of style-based versus geometric TTA is conditional rather than absolute across all model families. On ImageNet-R, Foundation Models (CLIP ViT-B/16 and DINOv2 ViT-B/14) consistently respond better to geometric TTA than to style-based views. Yet, on specialized domain-shift benchmarks (EuroSAT, Camelyon17, MIDOG2022, EpiStr), model architectures interact directly with dataset characteristics: Geometric TTA outperforms STyle-TTA on CNNs for histology benchmarks like Camelyon17 and EpiStr, whereas STyle-TTA yields top performance for Foundation Models on Camelyon17 and MIDOG2022. Rather than a universal rule governed strictly by backbone family, the effectiveness of TTA strategies reflects a combined dependency on model representation and target domain properties.

Texture Bias and Appearance Invariance A potential explanation for the observed differences in the effectiveness of style-based test-time augmentation can be found in the differing interpretations of feature reliance in ImageNet-trained neural networks. Geirhos et al. Geirhos et al. (2022) demonstrated that conventional ImageNet-trained convolutional neural networks tend to rely more heavily on texture than shape when making classification decisions. In contrast, Burgert et al. Burgert et al. (2026) challenge the interpretation of this behaviour as an inherent texture bias and argue that the visual cues exploited by CNNs may be more accurately understood in terms of local shape features. These differing perspectives provide a useful framework for interpreting the architecture-dependent response to STyle-TTA observed in this thesis.

The results on ImageNet-1K → ImageNet-R are consistent with the hypothesis that appearance-related information plays an important role in the predictions of conventional supervised models. STyle-TTA improved Top-1 accuracy by 12.32 percentage points for ResNet-18 and 10.64 percentage points for DenseNet-121, substantially exceeding the gains achieved by geometric TTA (+3.91 pp and +3.30 pp, respectively). A similar pattern was observed for the supervised ViT-B/16, where STyle-TTA improved Top-1 accuracy by 10.20 percentage points compared with 4.73 percentage points for geometric augmentation. These findings suggest that modifying image appearance at inference time can mitigate some of the errors introduced by the domain shift. In the context of Geirhos et al. (2022), this behaviour is compatible with the interpretation that supervised ImageNet-trained models rely on appearance-related

features that may become less reliable when the visual characteristics of the target domain differ from those encountered during training.

The effectiveness of STyle-TTA can therefore be interpreted as evidence that appearance-based augmentation addresses a relevant component of the domain shift for these backbones. By introducing stylized views that alter the visual appearance of the input while preserving its semantic content, STyle-TTA provides the classifier with multiple observations of the same object under different appearance conditions. The resulting improvements suggest that these additional observations can support more robust predictions when the original image contains appearance characteristics that are less compatible with the learned representation. This interpretation is particularly relevant for the CNN backbones, which showed the largest improvements from style-based augmentation across the evaluated benchmarks.

However, the findings of Burgert et al. (2026) suggest that the effectiveness of style-based augmentation should not be equated directly with the presence of a fixed texture bias. Their work questions whether the conventional texture-versus-shape interpretation fully captures the visual features used by ImageNet-trained CNNs. From this perspective, the improvements observed with STyle-TTA may reflect a broader sensitivity to local appearance statistics rather than a specific reliance on texture as defined by the cue-conflict paradigm. The distinction is relevant because style transfer can modify several properties of an image simultaneously, including texture, colour, contrast, and other appearance-related characteristics. Consequently, the present results indicate that appearance-based transformations can be beneficial for certain models, but do not establish which individual visual cues are responsible for the observed improvements.

Architecture and Pretraining Objective The contrasting behaviour of the foundation-model backbones provides further insight into the relationship between learned representations and the effectiveness of style-based augmentation. While STyle-TTA substantially improved performance for the supervised CNNs and both Swin-B and ViT-B/16 on ImageNet, it reduced accuracy relative to the baseline for both CLIP ViT-B/16 (-1.44 pp) and DINOv2 ViT-B/14 (-1.29 pp) on ImageNet-R. CLIP and DINOv2 achieved higher unaugmented baseline accuracies than the conventional supervised backbones, yet did not benefit from additional stylized views on ImageNet-R. This suggests that the usefulness of appearance-based TTA is not determined by model capacity or baseline performance alone. Instead, the learned representations and the type of visual information they exploit may influence whether stylized views provide useful complementary information or introduce additional variability.

The comparison between the supervised vision transformers and the foundation-model backbones is particularly informative in this regard. Although all four models use transformer-based architectures, their response to STyle-TTA differs substantially. Both supervised ViT-B/16 and hierarchical Swin-B benefited from style-based augmentation, whereas CLIP and DINOv2 experienced performance degradation. This pattern suggests that architectural category alone is insufficient to explain

the effectiveness of STyle-TTA. Differences in pretraining objectives and the resulting representations may contribute to the observed behaviour, with the foundation models potentially being less dependent on the specific appearance characteristics modified by STyle-TTA. The results therefore support a distinction between the architecture of a model and the feature reliance established through its training procedure.

The view-scaling experiments provide additional support for this interpretation. Conventional supervised backbones generally benefited from increasing the number of stylized views, whereas CLIP and DINOv2 achieved their highest performance at a small number of views and subsequently declined as more stylized images were introduced. This suggests that the additional views do not provide the same benefit across all backbones. For the supervised models, further style-conditioned observations may offer complementary information under the domain shift. For the foundation models, the same transformations may instead introduce variability that is not useful for the learned representation. The observed scaling behaviour thus indicates that the effectiveness of style-based TTA depends not only on the presence of a domain shift, but also on how the model responds to the specific appearance changes introduced by the augmentation. Taken together, the findings support a more nuanced interpretation of the relationship between texture bias and appearance invariance. The strong improvements observed for conventional supervised models are consistent with the motivation of Geirhos et al., in that modifying appearance-related information can improve robustness under domain shift. At the same time, the perspective of Burgert et al. highlights that such improvements do not necessarily demonstrate a specific texture bias, as the underlying features may involve local shape or other appearance-related information. The results of this thesis therefore suggest that the practical usefulness of style-based TTA is closely related to the visual representations learned by the backbone, rather than being a universal consequence of model architecture alone.

Semantic Matching Matters The retrieval-strategy ablation (Section 5.8.1) provides independent evidence that STyle-TTA's mechanism is genuinely about semantic style matching rather than style transfer in general. Feature-based (DINO) retrieval outperformed both random and balanced-random retrieval by 5–10 percentage points across every backbone tested, with balanced-random retrieval performing no better than uninformed random retrieval. This rules out the possibility that STyle-TTA's gains are attributable simply to exposing the classifier to some stylized variation of each image. The specific choice of style reference matters, and choosing one that is semantically close to the target image in embedding space does real work.

Aggregation Strategies: The Accuracy–Calibration Trade-off The aggregation strategy ablation (Section 5.8.3) revealed a systematic trade-off between raw accuracy and probabilistic calibration. Simple probability averaging (vanilla) achieved the lowest Expected Calibration Error for every backbone category; ECE

was up to 3.7× lower than ZERO or TPT (ranging from 1.3× to 3.7× across architectures), while confidence-based schemes traded substantially worse calibration for accuracy gains that were themselves backbone-dependent: meaningful for Foundation Models (up to +2.3 percentage points under TPT), negligible for CNNs (+0.23 points), and absent for the supervised ViT (where TPT reduced accuracy by 1.2 points). This trade-off is further compounded by the observation that all forms of TTA evaluated in this thesis tend to increase ECE relative to the unaugmented baseline, including geometric TTA, which is often treated as a "safe" default. This suggests that STyle-TTA's practical value depends not only on architecture but on how much a given deployment scenario weighs raw accuracy against calibrated confidence. This distinction matters in the medical-imaging benchmarks (Camelyon17-WILDS, MIDOG2022, EpiStr), where a confidently wrong prediction typically carries a different cost than an uncertain one, and where the well-calibrated vanilla-aggregation configuration may be the more defensible default even when a confidence-filtered alternative achieves marginally higher raw accuracy.

Generalization Across Benchmarks The results across the additional benchmarks demonstrate that the optimal TTA strategy depends heavily on the specific nature of the domain shift.

On satellite imagery (EuroSAT → UCMerced), where visual appearance and spectral properties vary significantly while spatial orientations contain domain-specific spatial semantics, STyle-TTA delivers large gains for CNNs (+14.67 pp Top-1) while Geometric TTA induces severe performance drops (−10.17 pp Top-1). Furthermore, on this dataset, STyle-TTA preserves balanced accuracy far better than raw Top-1 accuracy for ViT and FM models (reaching 68.92% and 69.75% balanced accuracy, respectively).

On microscopic histology imagery (Camelyon17-WILDS and EpiStr), both TTA modalities yield consistent improvements over baseline across CNN backbones. However, spatial perturbations prove more effective than style shifts for CNNs, with Geometric TTA outperforming STyle-TTA on EpiStr (78.92% vs. 73.69% Top-1) and Camelyon17 (85.40% vs. 80.40% Top-1). Interestingly, on these histology tasks, Vision Transformers and Foundation Models benefit more from STyle-TTA (e.g., 95.00% STyle-TTA vs. 93.85% Geometric TTA for FMs on Camelyon17), suggesting that higher-capacity backbones leverage stylized views more effectively when primary spatial features are already robustly encoded.

Finally, on MIDOG2022, overall performance remains constrained across all settings (with balanced accuracy remaining close to 50%), reflecting either task saturation or domain shifts that lie beyond the scope of simple spatial or style augmentations.

Hybrid Geometric–Style Configuration The hybrid experiments further suggest that geometric and style-based augmentations address different aspects of robustness and should not be viewed as interchangeable alternatives. Geometric transformations primarily modify the spatial presentation of an object or scene, whereas

style transformations modify its appearance while aiming to preserve its semantic content. Combining the two therefore creates views that vary along both spatial and appearance dimensions. This can be beneficial when the target domain differs from the source domain in multiple ways, since the model is exposed to a broader set of plausible input configurations. The hybrid configuration is consequently best understood as a form of complementary augmentation rather than simply a larger collection of views.

However, the effect of combining the two augmentation families depends on the model and benchmark. For models that benefit from both geometric and style-based views, the hybrid strategy can provide a more comprehensive robustness mechanism by addressing both spatial and appearance-related discrepancies. For models that already exhibit strong invariance to one of these factors, the additional views may provide limited benefit or may even introduce unnecessary prediction variability. This interpretation is consistent with the observed differences between conventional convolutional models, supervised vision transformers, and foundation models. In particular, the stronger response of convolutional models to style augmentation suggests that adding style views can be useful when appearance sensitivity remains a major source of error, while the more consistent geometric gains of foundation models indicate that spatial transformations may be the safer default for representations that are already relatively insensitive to style.

The hybrid results therefore motivate a model- and domain-dependent selection strategy. Rather than assuming that combining all available augmentations will always improve performance, test-time augmentation should be designed around the expected structure of the domain shift. Geometric views are most directly motivated when changes in viewpoint, scale, or spatial arrangement are expected, while style views are most useful when the source and target domains differ in illumination, color, texture, or acquisition conditions. A hybrid strategy is most compelling when both types of discrepancy are present and when the model can benefit from the resulting diversity without losing prediction consistency. In this sense, the experiments support the broader conclusion that effective test-time augmentation is not determined solely by the number of views, but by the semantic relevance of the transformations used to construct them.

7 Limitations

7.1 Statistical Rigor and Computational Constraints

All style-based augmentation conditions in this thesis were evaluated using a single random seed, whereas geometric TTA was averaged over three seeds. While geometric TTA's seed-to-seed variation was negligible ($\leq$0.2 percentage points in Top-1 accuracy), this provides no guarantee that style-based conditions exhibit similarly low variance. Specifically, because StyleID relies on stochastic DDIM sampling, the initial noise maps sampled during view generation introduce an additional source of variance alongside style-reference selection. The rankings and effect sizes reported for StyleID and AdaIN should therefore be interpreted as indicative rather than statistically definitive. A priority for any follow-up study is to quantify this combined generation variance directly, even if only on a representative subset of one benchmark.

A related constraint is the substantial computational cost of diffusion-based style transfer. StyleID was capped at $N = 16$ in the main ablations and $N \leq 4$ in the additional benchmarks, while AdaIN and geometric TTA were evaluated up to $N = 64$. For CNN and supervised-ViT backbones, accuracy had not clearly saturated at the maximum view counts tested for the cheaper methods, suggesting that StyleID's performance ceiling may be systematically underestimated for exactly the architectures it benefits most. The comparisons presented here are therefore conservative with respect to StyleID's potential advantage.

7.2 Scope of Empirical and Mechanistic Claims

The mechanistic account offered in Chapter 6 — that STyle-TTA's utility is governed by backbone texture bias—draws on established literature about supervised CNNs and foundation-model pretraining objectives (Geirhos et al., 2022), rather than on dedicated probing experiments conducted within this thesis. No independent texture-sensitivity diagnostic (e.g., cue-conflict datasets, controlled texture–shape manipulation) was used to validate this hypothesis for the specific classifiers and benchmarks evaluated here. The explanation is therefore plausible and consistent with the observed behavioural patterns, but it remains an indirect inference.

Similarly, the MIDOG2022 benchmark represents a genuine negative result that this thesis cannot fully diagnose. Balanced accuracy remained near chance for every backbone and condition, suggesting that the source-trained classifiers may never have acquired a scanner-invariant discriminative signal for this task.

Finally, STyle-TTA is inherently limited by its input-based formulation. It can only reweight and recombine predictions that the frozen classifier is already capable of making: it offers no mechanism to recover discriminative signal that the source classifier failed to learn during training. This was evident on MIDOG2022, but

it applies to any deployment scenario where the source model's representation is fundamentally inadequate for the target task.

8 Future Work

Scaling STyle-TTA to higher view counts. Given that CNN and supervised-ViT accuracies had not plateaued at $N = 64$ for cheaper augmentation strategies, an important empirical question is whether STyle-TTA continues to improve for these backbones at $N = 32$ or $N = 64$. If so, the performance gap between style-based and geometric TTA may be substantially larger than reported here. This, however, depends on either access to greater computational resources or the development of faster style-transfer operators.

Lightweight style-transfer operators. StyleID's computational cost is a practical barrier to deployment. Future work could investigate whether more efficient diffusion-based stylization methods—such as distilled few-step diffusion models or latent-consistency models—preserve the same corrective benefits at a fraction of the inference cost. Alternatively, flow-based generative models might provide a fruitful alternative by replacing iterative denoising with a direct path from noise to samples.

Direct texture-bias probing. Rather than inferring texture sensitivity from accuracy patterns, future work could directly measure it for each backbone using controlled stimuli (e.g., the cue-conflict dataset of Geirhos et al. (2022) or systematic texture–shape swaps). This would transform the architectural split observed here from an empirical correlation into a predictive diagnostic: one could test a new backbone *before* deploying STyle-TTA, and predict whether it will respond positively to style-based augmentation.

Adaptive and learned aggregation. This thesis employed a fixed number of retrieved references and a fixed aggregation strategy for all samples. Future work could explore per-instance adaptive view counts: generating additional stylized views only when the original prediction confidence is low, or when the retrieved references span a diverse range of feature-space neighborhoods. Similarly, learned aggregation weights, trained on source-domain data, could improve both accuracy and calibration beyond uniform averaging.

Source-free and reduced-source variants. STyle-TTA assumes access to a source-domain pool instances at test time. A valuable extension would be to relax this requirement by replacing the retrieval step with a compact, precomputed style bank (e.g., per-class mean style statistics or a small set of learned style anchors) that captures source-domain appearance without storing the full training set. This

would broaden applicability to deployment scenarios with strict storage or privacy constraints.

9 Conclusion

This thesis evaluated the effectiveness of style-based test-time augmentation (STyle-TTA) under domain shift, comparing style-transfer operators against conventional geometric TTA across diverse model architectures and benchmarks. The central finding is that modifying image appearance at inference time can yield substantial performance gains over unaugmented baselines and standard geometric transformations. However, the degree to which style-based TTA succeeds is not a simple function of model capacity or view budget; rather, it is governed by a conditional interaction between learned backbone representations and the specific structure of the target domain shift.

On the primary ImageNet-1K $\rightarrow$ ImageNet-R benchmark and appearance-dominated tasks such as EuroSAT, STyle-TTA demonstrated its clearest advantages for conventional supervised models. For convolutional networks and supervised Vision Transformers, style-based transformations mitigated severe performance drops caused by shifts in texture, color, and illumination. These results align with the perspective that standard supervised training leaves models sensitive to local appearance statistics. By presenting the classifier with semantically preserved yet stylized variants of target images, STyle-TTA provides complementary visual observations that help regularize predictions under domain shift.

Crucially, the empirical results across the extended benchmark suite reveal that style-based TTA is not a universal replacement for geometric transformations. The optimal augmentation strategy is strongly domain-dependent:

- **Appearance-Dominant Shifts:** On datasets like EuroSAT, where spectral properties and background appearances change drastically while spatial orientations carry semantic meaning, STyle-TTA delivers massive gains for CNNs (+14.67 pp), whereas geometric TTA incurs severe degradation (−10.17 pp).
- **Spatial and Morphological Shifts:** On microscopic histology datasets (Camelyon17, EpiStr), geometric TTA outperforms STyle-TTA on CNN backbones. In these settings, spatial perturbations (rotations, flips) preserve cellular morphology more effectively than synthetic style operations, producing the strongest overall performance.
- **Foundation Model Behavior:** While foundation models (CLIP ViT-B/16, DINOv2 ViT-B/14) favor geometric TTA on ImageNet-R, they effectively leverage STyle-TTA on specialized medical domains like Camelyon17 and MIDOG2022. This demonstrates that pretraining objectives alter how representations ingest stylized views rather than rendering foundation models inherently immune or averse to style transformations.

Ablation analyses further clarify the operational boundaries of STyle-TTA. First, semantic matching matters: feature-based retrieval consistently outperforms random style selection, confirming that STyle-TTA relies on semantically meaningful

style transfer rather than arbitrary visual noise. Second, view-scaling behavior diverges across architectures; conventional supervised models benefit monotonically from larger view budgets ($N = 16$ or $N = 64$), whereas foundation models peak early ($N = 2$) on natural image shifts, after which additional stylized views introduce unhelpful variance. Third, aggregation strategies reveal a fundamental trade-off between raw classification accuracy and probabilistic calibration. While confidence-weighted aggregation schemes can maximize raw accuracy for specific backbones, simple probability averaging (vanilla aggregation) consistently preserves superior calibration, making it the more defensible default in safety-critical applications like medical diagnosis.

In summary, test-time augmentation should not be treated as a static, model-agnostic post-processing step. Effective TTA requires aligning the augmentation modality with both the feature reliance of the classifier and the geometric or appearance-based nature of the domain shift. Where spatial orientation is paramount or foundation models face natural images, geometric TTA remains a reliable default. However, when domain discrepancy is driven by appearance variations or when convolutional models must generalize to novel visual environments—STyle-TTA offers a powerful, parameter-free mechanism to enhance model robustness at inference time.

10 Appendix

Table A.1: Detailed architectural specifications of the evaluated classifiers.

Model	**Parameters (M)**	**Feature Dim. (d)**	**Input Resolution**	**Patch / Window Size**
ResNet-18	11.7M	512	224×224	N/A
DenseNet-121	8.0M	1024	224×224	N/A
ViT-B/16	86.6M	768	224×224	16×16
Swin-B	88.0M	1024	224×224	4×4 (window 7)
DINOv2 ViT-B/14	86.6M	768	224×224	14×14
CLIP ViT-B/16	86.2M	512	224×224	16×16

Table A.2: Hyperparameter settings for fine-tuning and linear probing.

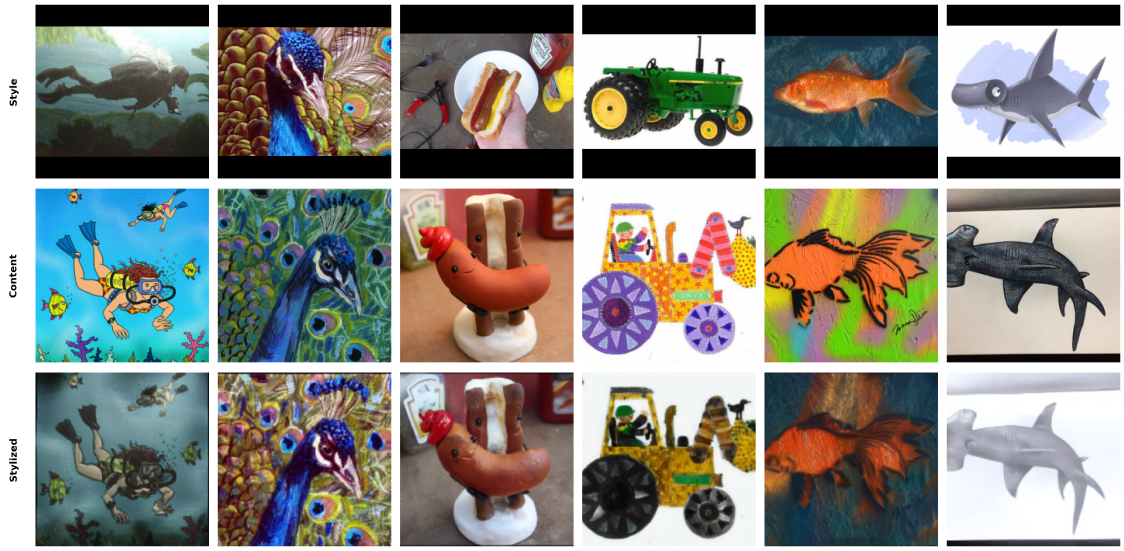
Setting	**Fine-tuning**	**Linear probing**
Optimizer	SGD	SGD
Momentum	0.9	0.9
Weight decay	0	0
Initial LR	3×10^{-4}	1×10^{-2}
Batch size	32	32
Maximum epochs	20	50
LR schedule	Cosine annealing	Cosine annealing
Early stopping	Validation loss	Validation loss
Seed	42	42
Augmentation	Random crop + horizontal flip	Random crop + horizontal flip

Table A.3: Classifier performance metrics across different view counts for DINO and Balanced Random retrieval methods.

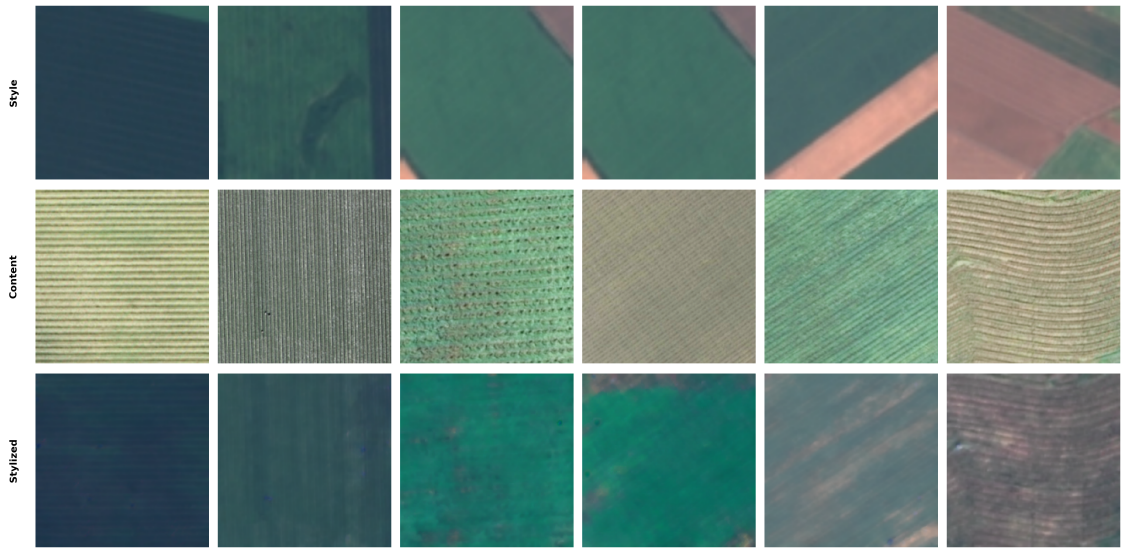
Retrieval Method	Views	True-Class Presence (%)	Unique Classes	True-Class Dominance (%)
DINO	1	45.29	1.00	45.29
	3	57.17	2.07	43.63
	7	65.24	3.81	42.24
	15	71.75	6.67	41.03
	31	77.70	11.35	39.75
	63	82.70	18.58	38.28
Balanced Random	1	0.60	1.00	0.60
	3	1.73	3.00	0.58
	7	4.05	7.00	0.58
	15	8.07	15.00	0.54
	31	15.81	31.00	0.51
	63	31.32	63.00	0.50

Table A.4: Performance of the AdaIN test-time augmentation strategy ($N = 64$) against the no-TTA baseline on ImageNet-1K $\rightarrow$ ImageNet-R. Reported are balanced accuracy, absolute change from baseline with 95% CI, and significance (paired bootstrap test).

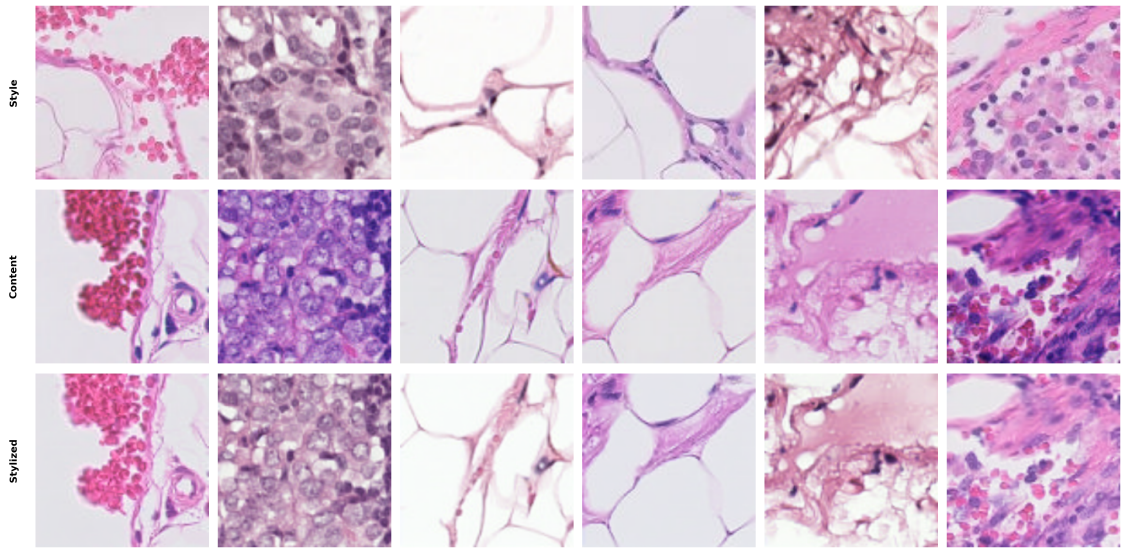
Backbone	**Baseline (%)**	**TTA (% ↑)**	**Δ (95% CI)**	p
ResNet-18	35.44%	39.46%	+4.02% [+3.53%, +4.52%]	<0.001
DenseNet-121	40.60%	42.34%	+1.75% [+1.24%, +2.26%]	<0.001
ViT-B/16 (224)	47.89%	52.95%	+5.06% [+4.55%, +5.58%]	<0.001
Swin-B (224)	60.41%	59.31%	-1.10% [-1.59%, -0.60%]	0.342
CLIP ViT-B/16	64.23%	59.93%	-4.31% [-4.85%, -3.76%]	<0.001
DINOv2 ViT-B/14	69.04%	65.02%	-4.02% [-4.49%, -3.55%]	<0.001



(a) ImageNet-R

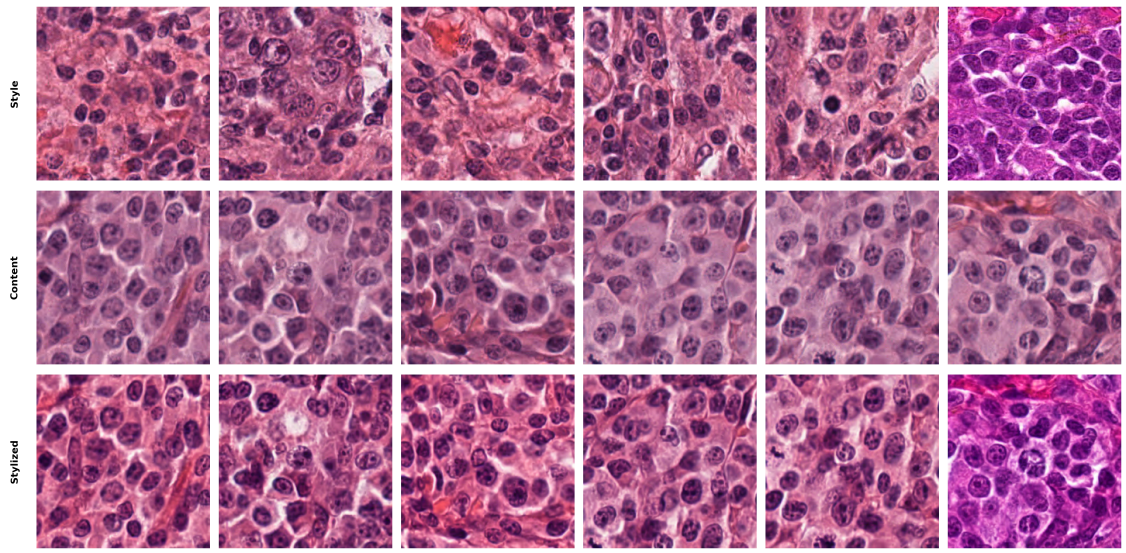


(b) Satellite (EuroSAT → UCMerced)

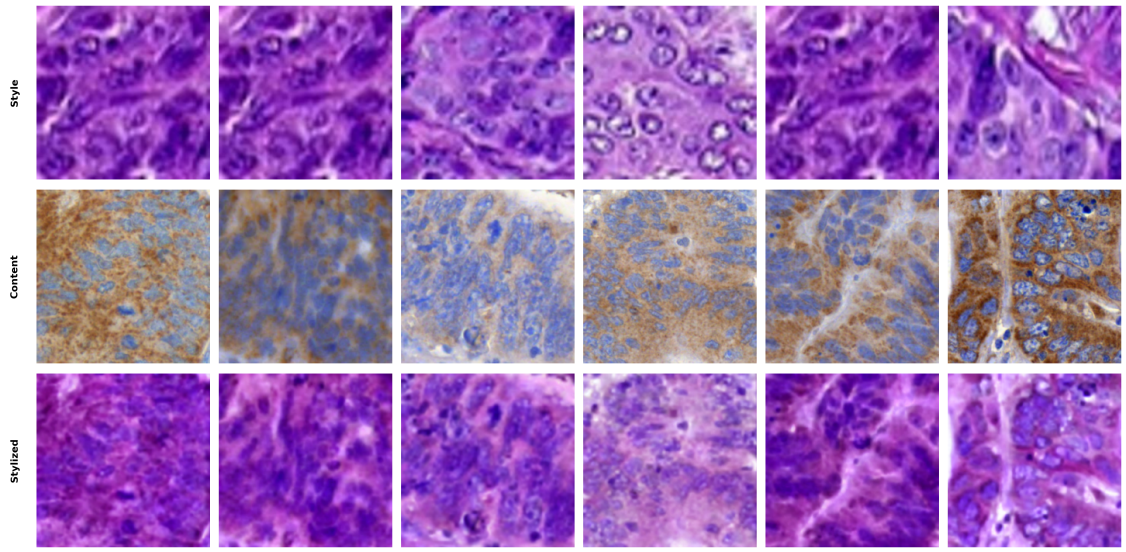


(c) Camelyon17-WILDS

Figure 16: Qualitative style transfer examples across datasets (Part 1).



(a) MIDOG



(b) EpiStr

Figure 17: Qualitative style transfer examples across datasets (Part 2).

Table A.5: **Impact of augmented view count (N) on TTA performance across backbones.** Evaluation of balanced accuracy (%) on ImageNet-R across varying numbers of generated views ($N \in \{2, 4, 8, 16, 32, 64\}$) for AdaIN, StyleID, and Geometric TTA. The unadapted baseline ($N = 1$) corresponds to single-image inference. $\Delta(N_{\text{min}} \rightarrow N_{\text{max}})$ indicates the absolute percentage point change from the minimal evaluated view count ($N = 2$) to the maximum ($N = 64$ for AdaIN/Geometric; $N = 16$ for StyleID). Bold text denotes the peak balanced accuracy for each backbone.

			Bal. Acc (%↑) (N_{refs})						
Classifier	**Strategy**	**Baseline**	$N = 2$	$N = 4$	$N = 8$	$N = 16$	$N = 32$	$N = 64$	$\Delta(N_{\text{min}} \rightarrow N_{\text{max}})$
ResNet-18	AdaIN	35.44	37.09	38.27	37.97	39.01	39.37	39.46	+2.37
	StyleID		41.91	44.93	46.28	**46.80**	–	–	+4.89
	Geometric		37.16	38.01	39.11	39.26	39.38	39.46	+2.30
DenseNet-121	AdaIN	40.60	41.60	42.15	40.96	42.09	42.48	42.34	+0.74
	StyleID		46.31	48.68	49.84	**50.32**	–	–	+4.01
	Geometric		41.69	42.86	43.38	43.67	43.74	44.07	+2.37
ViT-B/16 (224)	AdaIN	47.89	49.55	50.50	50.45	51.84	52.45	52.95	+3.40
	StyleID		52.95	55.40	56.75	**57.82**	–	–	+4.87
	Geometric		49.57	51.07	52.00	52.43	52.66	52.80	+3.23
Swin-B (224)	AdaIN	60.41	59.07	58.89	58.51	59.28	59.25	59.31	+0.24
	StyleID		61.94	62.97	63.56	**63.88**	–	–	+1.94
	Geometric		60.96	61.33	62.23	62.47	62.45	62.45	+1.49
CLIP ViT-B/16	AdaIN	64.23	64.91	64.86	61.69	60.84	60.47	59.93	-4.99
	StyleID		65.27	64.35	63.52	62.78	–	–	-2.49
	Geometric		64.65	65.17	65.44	65.71	**65.72**	65.64	+0.99
DINOv2 ViT-B/14	AdaIN	69.04	69.10	68.66	66.78	65.76	65.32	65.02	-4.08
	StyleID		69.18	68.50	67.73	67.23	–	–	-1.94
	Geometric		69.00	69.52	70.25	70.41	70.44	**70.58**	+1.57

Table A.6: Calibration (ECE & AUC) of Hybrid approach of STyle-TTA against (Baseline → TTA), ImageNet → ImageNet-R. Ensemble size for all methods $N = 64$ except regular STyle-TTA $N = 16$

		Geo-TTA ($N = 64$)		**StyleID** ($N = 16$)		**Hybrid S1** ($N = 64$)		**Hybrid S2** ($N = 64$)	
Dataset	**Backbone**	**ECE ↓**	**AUC ↑**	**ECE ↓**	**AUC ↑**	**ECE ↓**	**AUC ↑**	**ECE ↓**	**AUC ↑**
MIDOG	CNN	0.23	0.48	**0.15**	**0.52**	0.23	0.47	0.25	0.48
	Vision Transformer	**0.17**	**0.51**	0.23	0.49	0.17	0.51	0.22	0.50
	Foundation Model	0.14	0.49	**0.11**	**0.52**	0.14	0.49	0.22	0.49
UCMerced	CNN	0.37	0.67	**0.04**	**0.83**	0.35	0.70	0.08	0.81
	Vision Transformer	**0.17**	0.93	0.34	0.89	0.17	**0.93**	0.25	0.89
	Foundation Model	0.22	**0.92**	0.27	0.91	**0.21**	0.92	0.25	0.91
Camelyon17-WILDS	CNN	0.10	0.93	0.14	0.93	0.10	0.94	**0.08**	**0.97**
	Vision Transformer	0.07	0.98	**0.02**	**0.98**	0.07	0.98	0.05	0.98
	Foundation Model	0.04	0.98	**0.02**	0.98	0.05	0.98	0.04	**0.99**
EpiStr	CNN	**0.08**	0.87	0.18	0.91	0.08	0.88	0.12	**0.94**
	Vision Transformer	0.21	**0.96**	0.12	0.83	0.21	0.96	**0.12**	0.83
	Foundation Model	0.17	**0.96**	0.20	0.87	0.17	0.95	**0.16**	0.91
ImageNet-R	CNN	**0.08**	0.93	0.09	0.95	0.09	0.94	0.15	**0.95**
	Vision Transformer	0.19	0.97	**0.13**	**0.98**	0.20	0.97	0.23	0.98
	Foundation Model	0.09	0.98	**0.05**	0.98	0.09	**0.99**	0.14	0.98

Bibliography

Marc Aubreville, Nikolas Stathonikos, Vincent Christlein, Samir Jabari, Katharina Ganz, Christof A. Bertram, Robert Klopfleisch, Mitko Veta, Katharina Breininger, et al. Domain generalization across tumor types, laboratories, and species – insights from the 2022 edition of the mitosis domain generalization challenge. *Medical Image Analysis*, 94:103155, 2024. doi: 10.1016/j.media.2024.103155.

Peter Bandi, Oscar Geessink, Qian Manson, Marcory Van Dijk, Maschenka Balkenhol, Geert Hermsen, Babak Ehteshami Bejnordi, Binyoung Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2018. doi: 10.1109/TMI.2018.2867350.

Andrew H. Beck, Ankur R. Sangoi, Samuel Leung, Robert J. Marinelli, Torsten O. Nielsen, Marc J. van de Vijver, Robert B. West, Matt van de Rijn, and Daphne Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine*, 3(108):108ra113, 2011. doi: 10.1126/scitranslmed.3002564.

Srinadh Bhojanapalli, Aritra Chakrabarti, Daniel Glasner, D Li, A Sarna, and Sanjiv Kumar. Understanding robustness of transformers for image classification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10231–10241, 2021.

Kay H Brodersen, Cheng Soon Ong, Klaas E Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *International Conference on Pattern Recognition (ICPR)*, pages 3121–3124, 2010.

Tom Burgert, Oliver Stoll, Paolo Rota, and Begüm Demir. Imagenet-trained cnns are not biased towards texture: Revisiting feature reliance through controlled suppression, 2026. URL `https://arxiv.org/abs/2509.20234`.

Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1992–2001, 2017.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Sebastian Doerrich, Francesco Di Salvo, Jonas Alle, and Christian Ledig. Stylizing vit: Anatomy-preserving instance style transfer for domain generalization, 2026. URL `https://arxiv.org/abs/2601.17586`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1033–1038, 1999.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven test-time adaptation, 2023. URL `https://arxiv.org/abs/2207.03442`.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2022. URL `https://arxiv.org/abs/1811.12231`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

Jiayi Guo, Junhao Zhao, Chaoqun Du, Yulin Wang, Chunjiang Ge, Zanlin Ni, Shiji Song, Humphrey Shi, and Gao Huang. Everything to the synthetic: Diffusion-driven test-time adaptation via synthetic-domain alignment, 2024. URL `https://arxiv.org/abs/2406.04295`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. URL `https://arxiv.org/abs/1512.03385`.

David J Heeger and James R Bergen. Pyramid-based texture analysis/synthesis. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 229–238, 1995.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2019. URL `https://arxiv.org/abs/1709.00029`.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017. doi: 10.48550/arxiv.1610.02136.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, 2021. URL `https://arxiv.org/abs/2006.16241`.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Ariel Aly, Michael Rubinstein, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations (ICLR)*, 2023.

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 *Advances in Neural Information Processing Systems*, volume 30, pages 6626–6637, 2017. doi: 10.48550/arXiv.1706.08500.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

Yunzhong Hou and Liang Zheng. Visualizing adapted knowledge in domain transfer. pages 13824–13833, 2021.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017.

Philip T. G. Jackson, Amir Atapour Abarghouei, Stephen Shirht, and Toby P. Breckon. Style augmentation: Data augmentation via style randomization. In *CVPR Workshops*, 2019.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 694–711, 2016.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild

distribution shifts. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5637–5664. PMLR, 2021. URL `https://arxiv.org/abs/2012.07421`.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

Chenxin Li, Xin Lin, Yijin Mao, Wei Lin, Qi Qi, Xinghao Ding, Yue Huang, Dong Liang, and Yizhou Yu. Domain generalization on medical imaging classification using episodic training with task augmentation. *Computers in Biology and Medicine*, 141:105144, 2022a. doi: 10.1016/j.compbiomed.2021.105144.

Shuaifeng Li, Mao Ye, Xiatian Zhu, Lihua Zhou, and Lin Xiong. Source-free object detection by learning to overlook domain style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8014–8023, 2022b.

Yijun Li, Chen Fang Liu, Jimei Yang, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 386–396, 2017.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1): 31–64, 2024. doi: 10.1007/s11263-024-02181-w.

Nina Linder, Riku Konusti, Riku Turkki, Vinod Diwan, Mikael Lundin, Caj Haglund, and Johan Lundin. Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. *Diagnostic Pathology*, 7(1):1–11, 2012. doi: 10.1186/1746-1596-7-22.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3122–3130, 2018.

Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, Quirine F. Manson, Nikolas Stathonikos, Alexi Baidoshvili, Paul van Diest, Carla Wauters, Marcory van Dijk, and Jeroen van der Laak. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*, 7(6):giy065, 2018. doi: 10.1093/gigascience/giy065.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

Ruiqi Lyu, Alistair Turcan, and Bryan Wilder. Explaining concept shift with interpretable feature attribution, 2026. URL `https://arxiv.org/abs/2505.20634`.

Biwen Meng, Xi Long, Wanrong Yang, Ruochen Liu, Yi Tian, Yalin Zheng, and Jingxin Liu. Advancing cross-organ domain generalization with test-time style transfer and diversity enhancement. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2025. doi: 10.1109/ISBI60581.2025.10981139.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Zachary Nado, Shreyas Padhy, D. Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. In *ICML Workshop on Uncertainty & Robustness in Deep Learning*, 2020.

Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI Conference on Artificial Intelligence*, volume 29, pages 2901–2907, 2015.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16805–16827. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/nie22a.html`.

Aude Oliva, Antonio Torralba, and Philippe G Schyns. Hybrid images. *ACM Transactions on Graphics (TOG)*, 25(3):527–532, 2006.

Maxime Oquab, Timothée Darcourt, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu Xu, Vasu Sharma, Shang-Wen Li, Tomer Sharma, Mengzhou Xu, David Girard, Priya Goyal, Pascal Singhal, Pierre-Emmanuel Mazaré, Armand Joulin, Piotr Bojanowski, and Ishan Misra. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2023.

Jungwuk Park, Dong-Jun Han, Soyeong Kim, and Jaekyun Moon. Test-time style shifting: Handling arbitrary styles in domain generalization, 2023. URL `https://arxiv.org/abs/2306.04911`.

Mihir Prabhudesai, Tsung-Wei Ke, Alexander C. Li, Deepak Pathak, and Katerina Fragkiadaki. Diffusion-tta: Test-time adaptation of discriminative models via generative feedback, 2023. URL `https://arxiv.org/abs/2311.16102`.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2008.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation, 2020. URL `https://arxiv.org/abs/2006.16971`.

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1214–1223, 2021.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: 10.1016/S0378-3758(00)00115-4.

Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019. doi: 10.1186/s40537-019-0197-0.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 14274–14287, 2022. URL `https://arxiv.org/abs/2209.07511`.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL `https://arxiv.org/abs/2508.10104`.

Prabhu Teja Sivaprasad and François Fleuret. Test time adaptation through perturbation robustness. *arXiv preprint arXiv:2110.10232*, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2256–2265, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Jan-Aike Termöhlen, Marvin Klingner, Leon J Brettin, Nico M Schmidt, and Tim Fingscheidt. Continual unsupervised domain adaptation for semantic segmentation by online frequency domain style transfer. In *Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2881–2888, 2021.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection, 2019. URL `https://arxiv.org/abs/1902.11097`.

Kangxian Xie, Siyu Huang, Sebastian Andres Cajas Ordonez, Hanspeter Pfister, and Donglai Wei. S^3-tta: Scale-style selection for test-time augmentation in biomedical image segmentation, 2024. URL `https://arxiv.org/abs/2310.16783`.

Rui Xie, Hongxin Wei, Lei Feng, Yidong Cao, and Bo An. On the importance of feature separability in predicting out-of-distribution error. *arXiv preprint arXiv:2303.15488*, 2023. doi: 10.48550/arxiv.2303.15488.

Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4085–4095, 2020.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 270–279, New York, NY, USA, 2010. Association for Computing Machinery. doi: 10.1145/1869790.1869829.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pages 3320–3328, 2014. URL `https://arxiv.org/abs/1411.1792`.

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 38629–38642, 2022.

Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning (ICML)*, pages 7301–7311, 2019.

Xingchen Zhao, Chang Liu, Anthony Sicilia, Seong Jae Hwang, and Yun Fu. Test-time Fourier style calibration for domain generalization. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1721–1727, 2022.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations (ICLR)*, 2021.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022. doi: 10.1109/TPAMI.2022.3195549.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.

Declaration of Authorship

Ich erkläre hiermit gemäß §9 Abs. 12 APO, dass ich die vorstehende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Des Weiteren erkläre ich, dass die digitale Fassung der gedruckten Ausfertigung der Abschlussarbeit ausnahmslos in Inhalt und Wortlaut entspricht und zur Kenntnis genommen wurde, dass diese digitale Fassung einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

Bamberg, 15.09.2026
Place, Date

N. Emmler
Signature