Dual Strategies for Test-Time Adaptation

Supplementary Material

A. Theoretical Justification for DualTTA

In the main paper, we introduced **DualTTA**, whose core idea is to (i) construct two sample sets, $\mathcal{D}^+$ and $\mathcal{D}^-$, and (ii) adapt the model using different objectives on these sets. For $\mathcal{D}^+$, we minimize prediction entropy to reinforce confident decisions, whereas for $\mathcal{D}^-$, we maximize entropy to suppress overly confident but incorrect predictions. The intuition is that samples in $\mathcal{D}^+$ are those the classifier is *likely* to classify correctly, while those in $\mathcal{D}^-$ are *likely* to be misclassified; hence we referred to them as "likely-correct" and "likely-incorrect" samples, respectively. However, we have not yet formally shown that $\mathcal{D}^+$ and $\mathcal{D}^-$ indeed correspond to samples with high and low correctness likelihoods. In this section, we state a theorem establishing this property and then provide its detailed proof.
As:

Theorem 1. Consider a sample $\mathbf{x}$ with style statistics $(\mathbf{U}, \mathbf{S})$. Let $\mathbf{x}^{sp}$ denote its semantic-preserving variant obtained by modifying $(\mathbf{U}, \mathbf{S})$ to $(\mathbf{U}^{sp}, \mathbf{S}^{sp})$, and let $\mathbf{x}^{sa}$ denote its semantic-altering variant produced via a strong augmentation $\mathbf{x}^{sa} = \mathbf{x} + \zeta$ as described in the main paper.
Let $\hat{y}(\mathbf{x})$ be the prediction probability of $\mathbf{x}$ and $f(\mathbf{x})$ be the one-hot label for $\mathbf{x}$, e.g. $f(\mathbf{x}) = \arg\max_c \hat{y}(\mathbf{x})$, and let y be its (unknown) ground-truth label. Then the probability of x being misclassified is $\text{error}(\mathbf{x}) = \Pr\big(f(\mathbf{x}) \neq y\big)$, and:

$$\Pr\big(f(\mathbf{x}) \neq y\big) \;\uparrow\uparrow\; \frac{\Pr\big(f(\mathbf{x}^{sp}) \neq f(\mathbf{x})\big)}{\Pr\big(f(\mathbf{x}^{sa}) \neq f(\mathbf{x})\big)}. \tag{1}$$

Here, the symbol $\uparrow\uparrow$ is used to denote a monotonic increasing relationship between the two variables, and $\uparrow\downarrow$ represents the monotonic decreasing relationship.

Theorem 2. Consider a sample $\mathbf{x}$ and its variant $\mathbf{x}' = \mathcal{T}(\mathbf{x})$ (either a semantic-preserving $\mathbf{x}^{sp}$ or semantic-altering variant $\mathbf{x}^{sa}$ or other transformation).
The probability that the model assigns different labels to them is given by $\Pr\big(f(\mathbf{x}') \neq f(\mathbf{x})\big)$, and it is proportional to $\text{Diff}(\hat{y}(\mathbf{x}'), \hat{y}(\mathbf{x}))$, where $\hat{y}(\mathbf{x})$ is the probability prediction and Diff(.) is the function defined in Equation 2 of the main paper. That is,

$$\Pr\big(f(\mathbf{x}') \neq f(\mathbf{x})\big) \propto [\text{Diff}(\hat{y}(\mathbf{x}'), \hat{y}(\mathbf{x}))]. \tag{2}$$

Corollary. For a sample $\mathbf{x}^+ \in \mathcal{D}^+$ and $\mathbf{x}^- \in \mathcal{D}^-$, the classification error of $\mathbf{x}^+$ is lower than that of $\mathbf{x}^-$, and there exists a factor ρ such that:

$$\text{error}(\mathbf{x}^+) < \rho\frac{\tau^{sp}}{\tau^{sa}} < \text{error}(\mathbf{x}^-), \tag{3}$$

This corollary follows directly from Theorem 1 and Theorem 2, and from the definitions of $\mathcal{D}^+$ and $\mathcal{D}^-$ as presented in the main paper.

$$\mathcal{D}^+ = \{\mathbf{x} \mid \text{Diff}(\hat{y}, \hat{y}^{sa}) > \tau^{sa},\ \text{Diff}(\hat{y}, \hat{y}^{sp}) < \tau^{sp}\},$$
$$\mathcal{D}^- = \{\mathbf{x} \mid \text{Diff}(\hat{y}, \hat{y}^{sa}) < \tau^{sa},\ \text{Diff}(\hat{y}, \hat{y}^{sp}) > \tau^{sp}\}, \tag{4}$$

This corollary justifies the construction of $\mathcal{D}^+$ and $\mathcal{D}^-$ and explains why we minimize the entropy for samples in $\mathcal{D}^+$ while maximizing the entropy for samples in $\mathcal{D}^-$, as described in the paper.

B. Theorem Proof

In this section we present the detailed proofs of the two theorems stated above. Since our theoretical analysis focuses solely on two types of classes—the correct class and the incorrect class for each sample—for clarity and ease of presentation, we restrict the analysis to the binary classification setting with class labels $\{+1, -1\}$

B.1. Proof of theorem 1

With the label set $\{+1, -1\}$. The decision rule can then be written compactly as:

$$f(\mathbf{x}) = \text{sign}\left(w^\top \phi(\mathbf{x})\right).$$

With the assumption above the decision boundary is located at 0. We define $M(\mathbf{x})$ as the random variable that represents distance between sample $\mathbf{x}$ and the decision boundary 0, then we have:

$$M(\mathbf{x}) := |w^\top \phi(\mathbf{x})|.$$

Let F_M denote the CDF of M, $F_M(t) = Pr(M \leq t)$ We consider two types of perturbations applied to $\mathbf{x}$:

- **Style perturbation.** Perturbing the feature statistics $(\mathbf{U}, \mathbf{S})$ to $(\mathbf{U}^{sp}, \mathbf{S}^{sp})$ ($\|\mathbf{U}^{sp} - \mathbf{U}\| + \|\mathbf{S}^{sp} - \mathbf{S}\| < \varepsilon$) induces a representation shift

  $$\phi(\mathbf{x}^{sp}) \approx \phi(\mathbf{x}) + \Delta_s,$$

  whose projection along w has magnitude $s := |w^\top \Delta_s|$. The probability of label flip is the probability that the distance to the decision boundary is smaller than the magnitude of the perturbation.

  $$\alpha := \Pr\left(f(\mathbf{x}^{sp}) \neq f(\mathbf{x})\right) = Pr(M \leq s) = F_M(s).$$

- **Semantic perturbation.** Applying a large semantic modification $\mathbf{x}^{sa} = \mathbf{x} + \zeta$ induces a representation shift

  $$\phi(\mathbf{x}^{sa}) \approx \phi(\mathbf{x}) + \Delta_{sem},$$

with projected magnitude $S := |w^\top \Delta_{sem}| \gg s$. The corresponding label-flip probability is the probability that the distance to the decision boundary is smaller than the magnitude of the perturbation.

$$\beta := \Pr\big(f(\mathbf{x}^{sa}) \neq f(\mathbf{x})\big) = Pr(M \leq S) = F_M(S).$$

Let the model's prediction error probability for $\mathbf{x}$ be

$$\text{error}(\mathbf{x}) := \Pr(f(\mathbf{x}) \neq y) = Pr(M < m_0) = F_M(m_0),$$

where m_0 is the threshold under which the classifier fails on $\mathbf{x}$.

The threshold m_0 following structural relationships:
1. The threshold m_0 increases with the sensitivity to style perturbation:

$$\exists g(.); m_0 = g(s), \qquad g' > 0.$$

2. The threshold m_0 decreases with the sensitivity to semantic perturbation:

$$\exists \psi(.); m_0 = \psi(S), \qquad \psi' < 0.$$

We prove that error$(\mathbf{x})$ is (i) increasing in α and (ii) decreasing in β.

(i) error$(\mathbf{x})$ increases with α. Since $\alpha = F_M(s)$ and F_M is strictly increasing, we have $s = F_M^{-1}(\alpha)$. Thus,

$$\text{error} = F_M(m_0) = F_M(g(s)) = F_M\big(g(F_M^{-1}(\alpha))\big).$$

The composition of three monotone increasing functions (F_M, g, and F_M^{-1}) is increasing; hence

$$\frac{\partial \text{error}}{\partial \alpha} > 0.$$

Equivalently,

$$\text{error}(\mathbf{x}) \uparrow\uparrow \alpha \tag{5}$$

(ii) error$(\mathbf{x})$ decreases with β. Similarly, since $\beta = F_M(S)$ is increasing in S, we have $S = F_M^{-1}(\beta)$, and

$$\text{error} = F_M(m_0) = F_M(\psi(S)) = F_M\big(\psi(F_M^{-1}(\beta))\big).$$

Here F_M and F_M^{-1} are increasing while ψ is strictly decreasing, so the composition is decreasing:

$$\frac{\partial \text{error}}{\partial \beta} < 0.$$

Equivalently,

$$\text{error}(\mathbf{x}) \uparrow\downarrow \beta \tag{6}$$

From 5 and 6, we have:

$$\text{error} \uparrow\uparrow \frac{\alpha}{\beta}$$

Theorem 1 is proven.

B.2. Proof of theorem 2

Let $\mathbf{x}$ be a fixed input and let $\mathbf{x}' = \mathcal{T}(\mathbf{x})$ denote a variant of $\mathbf{x}$. For binary classification we write the model's probability prediction vectors as Tie-breaking at $1/2$ is immaterial for the inequalities below.

For short, we denote

$$\Delta := y(\mathbf{x}') - y(\mathbf{x}) = \text{Diff}\big(\hat{y}(\mathbf{x}'), \hat{y}(\mathbf{x})\big), \qquad \delta := \big|y(\mathbf{x}) - \tfrac{1}{2}\big|.$$

Let $\mathbb{E}_{\mathbf{x}' \sim \mathcal{T}(\mathbf{x})}[|\Delta|]$ denote expectation with respect to the distribution generating $\mathbf{x}' \sim \mathcal{T}(\mathbf{x})$, for short we denote as $\mathbb{E}[|\Delta|]$.

A label change is that the absolute shift be at least δ. In other words,

$$\{\mathbf{x}' \mid f(\mathbf{x}') \neq f(\mathbf{x})\} \subseteq \{\mathbf{x}' \mid |\Delta| \geq \delta\}. \tag{7}$$

Similarly, if the probability shift is smaller than δ, the label of $\mathbf{x}$ does not change:

$$\{\mathbf{x}' \mid f(\mathbf{x}') = f(\mathbf{x})\} \supseteq \{\mathbf{x}' \mid |\Delta| \leq \delta\}. \tag{8}$$

We prove that for each $\mathbf{x}$, there exist positive constants C_1, C_2 such that:

$$C_1.\mathbb{E}[|\Delta|] \;\leq\; \Pr\big(f(\mathbf{x}') \neq f(\mathbf{x})\big) \;\leq\; C_2.\mathbb{E}[|\Delta|].$$

Upper bound. Take expectation over the randomness of the variant $\mathbf{x}'$. Using the inclusion in 7 we obtain:

$$\begin{aligned}\mathbb{E}\big[|\Delta|\big] &= \mathbb{E}\big[|\Delta| \cdot \mathbf{1}_{\{f(\mathbf{x}') \neq f(\mathbf{x})\}}\big] + \mathbb{E}\big[|\Delta| \cdot \mathbf{1}_{\{f(\mathbf{x}') = f(\mathbf{x})\}}\big] \\ &\geq \mathbb{E}\big[|\Delta| \cdot \mathbf{1}_{\{f(\mathbf{x}') \neq f(\mathbf{x})\}}\big] \geq \delta \Pr\big(f(\mathbf{x}') \neq f(\mathbf{x})\big).\end{aligned}$$

Rearranging gives:

$$\Pr\big(f(\mathbf{x}') \neq f(\mathbf{x})\big) \leq \frac{\mathbb{E}\big[|\Delta|\big]}{\delta} \tag{9}$$

Lower bound. We can decompose the expectation into the following:

$$\begin{aligned}\mathbb{E}\big[|\Delta|\big] =& \mathbb{E}\big[|\Delta| \mid f(\mathbf{x}') \neq f(\mathbf{x})\big] \Pr(f(\mathbf{x}') \neq f(\mathbf{x})) \\ &+ \mathbb{E}\big[|\Delta| \mid f(\mathbf{x}') = f(\mathbf{x})\big]\big(1 - \Pr(f(\mathbf{x}') \neq f(\mathbf{x}))\big).\end{aligned}$$

From 8 and the trivial upper bound $|\Delta| \leq 1$, we obtain:

$$\begin{aligned}\mathbb{E}\big[|\Delta|\big] \leq& 1 \cdot \Pr(f(\mathbf{x}') \neq f(\mathbf{x})) + \delta\big(1 - \Pr(f(\mathbf{x}') \neq f(\mathbf{x}))\big) \\ =& \delta + (1 - \delta)\Pr(f(\mathbf{x}') \neq f(\mathbf{x})).\end{aligned}$$

Hence,

$$\Pr\big(f(\mathbf{x}') \neq f(\mathbf{x})\big) \geq \frac{\mathbb{E}\big[|\Delta|\big] - \delta}{1 - \delta} \tag{10}$$

Combining 9 and 10, for fixed $\mathbf{x}$ (hence fixed δ), we have:

$$\frac{\mathbb{E}[|\Delta|] - \delta}{1-\delta} \leq \Pr\big(f(\mathbf{x}') \neq f(\mathbf{x})\big) \leq \frac{\mathbb{E}\big[|\Delta|\big]}{\delta}. \quad (11)$$

In particular, whenever δ is bounded away from zero there exist positive constants C_1, C_2 depend on fixed δ such that:

$$C_1.\mathbb{E}_{\mathbf{x}'\sim\mathcal{T}}[|\Delta|] \leq \Pr\big(f(\mathbf{x}') \neq f(\mathbf{x})\big) \leq C_2\,\mathbb{E}_{\mathbf{x}'\sim\mathcal{T}}[|\Delta|].$$

Then, for a fixed transformation $\mathcal{T}(.)$, there exists a constant $C_\mathcal{T}$, $C_1 < C_\mathcal{T} < C_2$, such that:

$$\Pr\big(f(\mathbf{x}') \neq f(\mathbf{x})\big) = C_\mathcal{T}.|\Delta|.$$

Equivalently, with $\Delta = \text{Diff}(\hat{y}(\mathbf{x}'), \hat{y}(\mathbf{x}))$,

$$\Pr\big(f(\mathbf{x}') \neq f(\mathbf{x})\big) \propto \text{Diff}(\hat{y}(\mathbf{x}'), \hat{y}(\mathbf{x})). \quad (12)$$

Theorem 2 is proven.

C. Algorithm Pseudo Code

Algorithm 1 presents the pseudocode for our proposed DualTTA, which details the selection of low-entropy samples, partitioning into aligned and misaligned sets, and the subsequent parameter update.

D. Devices Detail and Latency Analysis

D.1. Devices Detail.

All experiments in this study were conducted on a system equipped with an NVIDIA GeForce RTX 4090 GPU with 24,564 MiB of GPU memory, of which up to 13,664 MiB was utilized during training. The computing environment was configured with NVIDIA driver version 535.183.01, CUDA Toolkit version 12.2, and ran on a Linux operating system. All models were implemented using Pytorch.

D.2. Latency analysis.

Table A presents the results of measuring the time required for adaptation of 50000 samples in our proposed DualTTA and baselines under ImageNet-C, Gaussian noise, and severity level 5 environments. DualTTA achieves the highest performance while requiring less time than SAR and only slightly slower than other baselines.

Methods	% adapt	% correct	GPU time
No adapt	x	x	1m 58s
TENT	100%	x	2m 08s
EATA	28.2%	60.1%	2m 31s
SAR	27.2%	62.4%	4m 19s
DEYO	31.7%	75.9%	3m 11s
DualTTA	35.6%	89.5%	4m 11s

Table A. Data efficiency and runtime evaluation under ImageNet-C, Gaussian noise, and severity level 5.

E. Hyperparameter Explanation

E.1. List of hyperparameters and range of values

Our proposed DualTTA relies on 4 essential hyperparameters: content-preserving threshold τ^{sp}, semantic-altering threshold τ^{sa},normalization factor Diff_0 and trade-off coefficient between the two loss terms λ, in addition to the parameters introduced in previous works: normalization factor $\text{Ent}_0 = 0.4$.

For the range of these hyperparameters, we balance the semantic-preserving and semantic-altering threshold by setting $\tau^{sp} \in [0.6, 0.9]$ and $\tau^{sa} \in [0.2, 0.5]$. Setting τ^{sp} too low or τ^{sa} too high causes model-misaligned set D^- to dominate model-aligned set D^+, leading the model to focus more on unlearning information from D^- rather than learning from D^+. To ensure both the non-negativity and numerical balance of the components in Eq.(7), we introduce normalization constants $\text{Diff}_0 \approx \tau^{\text{sp}}$ and $\text{Ent}_0 = 0.4$. These constants act as scaling factors that stabilize the relative contributions of the entropy and sensitivity-aware terms during optimization. The trade-off coefficient λ serves as a balancing factor between two competing objectives in the overall loss function. Specifically, it controls the relative contribution of each loss term—typically one promoting confident predictions (e.g., entropy minimization) and the other enforcing corrective behavior (e.g., entropy maximization or regularization).

E.2. Performance stability under changes in hyperparameters

To assess the sensitivity of our method to these hyperparameters, we conduct an ablation study whose outcomes are presented in Figure 3 of the manuscript (for semantic-preserving threshold τ^{sp} and semantic-altering threshold τ^{sa}) and Table B (for Waterbirds) and C (for Office-Home) of the Supplementary materials (for others). The results indicate that our proposed DualTTA exhibits strong robustness across a range of threshold values, demonstrating that the method does not heavily rely on precise tuning of these hyperparameters to achieve consistent performance.

For generalization, through all the results of the experiment in the main manuscript, we set $\tau^{sp} = 0.7, \text{Diff}_0 = 0.7, \tau^{ca} = 0.4$ and $\lambda = 0.5$.

F. Impact of semantic-preserving positional layer

In this section, we conduct additional experiments on the ImageNet-C dataset on the placement of the semantic-preserving (style perturbation) layer using two backbones: ResNet50 and ViT-B. In deep learning architectures with N encoder layers, the style-extraction layer i is selected from the range $\left[\frac{N}{4}, \frac{3N}{4}\right]$ to balance abstraction and semantic entanglement. Early layers often lack stylistic detail, while later layers tend to entangle style with class-specific semantics. Therefore, the appropriate range of values to extract style features is $i = [1, 3]$ for ResNet (with $N = 4$) and $i = [4, 9]$ for ViT-B/32 (with $N = 12$). As shown in Figure A of the Supplementary Material, varying i within this range has minimal impact on performance, confirming the robustness of the proposed DualTTA framework.

To enhance generalization, we set the semantic-preserving layer position to 1 for the ResNet backbone and 7 for the ViT-B backbone across all the experiments.

G. Impact of sample weighting factor

To study the effectiveness of using the reliability weight $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$, which consists of three main components: Ent, Semantic-Preserving weight (SP), and Semantic-Altering weight (SA), we analyze the impact of these parameters when tested on the Waterbirds and Office-Home datasets with varying weight configurations. The results presented in Table D indicate that the model achieves the best performance when $\alpha(\mathbf{x})$ incorporates entropy weight, semantic-preserving weight (SP), and semantic-altering weight (SA), while $\beta(\mathbf{x})$ utilizes only the entropy weight. The reason is that model-aligned samples are filtered based on three factors: entropy, semantic-altering, and semantic-preserving, meaning their reliability depends on all three criteria. Meanwhile, $\beta(\mathbf{x})$ serves as a coefficient that helps the model forget the knowledge from model-misaligned samples. However, if $\beta(\mathbf{x})$ is too large, the model may prioritize forgetting over learning from model-aligned samples.

H. Further experiment results

H.1. Real and unknown domain shift

DualTTA as well as other TTA methods are designed to perform test-time adaptation (TTA) to reduce the domain discrepancy between training and testing data. In the main paper, we demonstrated its advantages over existing TTA methods through controlled experiments across multiple datasets exhibiting spurious correlations, domain shifts, or simulated corruptions such as noise, blur, or adverse weather conditions. In this experiment, we evaluate DualTTA in a more realistic setting where domain shift is assumed but neither its nature nor its extent is known a priori. To do so, we seek a dataset with many classes, each containing multiple images with varying numbers of object instances. Our goal is to assess whether TTA can still improve performance in such a challenging scenario, where the number of classes is large and each class exhibits significant intra-class variation. Unlike previous controlled settings, here we simply test whether a pretrained classifier can be effectively adapted at test time to improve accuracy.

Given this purpose, we use FSC147 [1], a dataset originally developed for visual object counting. This dataset is ideal for our evaluation because it contains 147 object categories, each with a diverse number of annotated instances. We construct a new classification benchmark by cropping each annotated object (using the provided bounding boxes) to form individual image samples.

Because this benchmark involves a large number of categories, it is natural to adopt an open-vocabulary classifier. To examine whether DualTTA is compatible with such models, we use the CLIP ViT-B/32 model (public weights) for all experiments. For each cropped image, we extract an image embedding using CLIP's vision encoder. For each class C in the 147 categories, we construct a text embedding using the prompt "A photo of <C>". Classification is then performed by assigning each test image to the class whose text embedding has the highest correlation with the image embedding. The text embeddings remain frozen throughout, while the image embeddings are adapted at test time.

Table F reports the performance of DualTTA and competing TTA baselines. Most TTA methods (except TENT) improve the accuracy of the base classifier, demonstrating the feasibility and usefulness of applying TTA to open-vocabulary classification under unknown domain shift. DualTTA achieves more than a 2% improvement over the no-adaptation baseline—a substantial gain given the difficulty of the task, which requires selecting a correct label among 147 classes (with extremely low chance performance). Among all evaluated TTA methods, DualTTA achieves the best performance, surpassing other baselines by a significant margin.

Methods	Acc
Pre-trained CLIP-based classifier (no adaptation)	17.38
TENT	14.13
EATA	18.55
SAR	17.91
DEYO	18.95
DualTTA	**19.92**

$\text{Ent}_0 =$	0.1	0.3	0.4	0.5
DualTTA	86.33	87.12	**88.44**	88.25
$\text{Diff}_0 =$	0.6	0.7	0.8	0.9
DualTTA	88.02	88.44	**88.50**	88.22
$\lambda =$	0.3	0.5	0.7	0.9
DualTTA	88.08	**88.44**	88.02	87.12

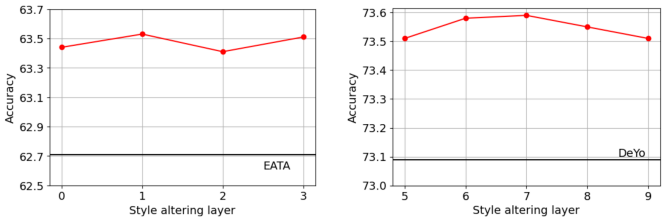
Table B. Impact of different hyper-parameters on Waterbirds dataset.

$\text{Ent}_0 =$	0.1	0.3	0.4	0.5
DualTTA	59.64	61.28	**61.51**	61.34
$\text{Diff}_0 =$	0.6	0.7	0.8	0.9
DualTTA	61.36	**61.51**	61.44	61.47
$\lambda =$	0.3	0.5	0.7	0.9
DualTTA	61.35	**61.51**	61.14	59.69

Table C. Impact of different hyper-parameters on Office-Home dataset.

Sample weights	$\alpha(\mathbf{x})$		$\beta(\mathbf{x})$	
	Office-Home	Waterbirds	Office-Home	Waterbirds
Ent+SA+SP	**61.51**	**88.44**	58.00	84.34
Ent+SA	60.08	87.33	60.58	87.28
Ent+SP	60.65	86.72	60.74	87.22
Ent	58.22	85.02	**61.51**	**88.44**

Table D. Impact of sample weighting.



(a) ResNet50 (b) ViT-B

Figure A. Performance of proposed method DualTTA with different semantic-preserving positional layer.

Table F. Accuracy comparison of DualTTA and other TTA methods on a new benchmark derived from FSC147 [1]. This benchmark involves classifying cropped object instances into one of 147 open-vocabulary categories using CLIP. While most methods (except TENT) improve upon the base classifier, DualTTA achieves the highest accuracy, demonstrating strong adaptation capability in a challenging open-vocabulary, domain-shifted setting.

H.2. Detail result of domain-shift datasets

We have summarized the overall results in Table 2 of the main script for the PACS and ColoredMNIST datasets, respectively. In this section, we provide a more detailed analysis by presenting the specific results. Table E demonstrates that on the Office-Home dataset, the model's performance can improve by up to 6.02% compared to the second-best method, DeYO, when the model is pretrained on Art and tested on Clipart. Similarly, Table E shows that on the PACS dataset, the model's performance can improve by up to 7.02% when pretrained on Photo and tested on Sketch.

H.3. Experiment with different levels of corruption

In this section, we present experiments conducted with a corruption level of 3 on the ImageNet-C dataset, in addition to the experiments with a corruption level of 5 previously reported in the main script. The detailed results are provided in G.

H.4. F1-Score Evaluation

While overall accuracy remains a useful indicator of model performance, it can be misleading in the presence of class imbalance or skewed error costs. To provide a more nuanced assessment of our method's ability to balance precision and recall, we additionally report the F1-score for all experiments. The F1-score—defined as the harmonic mean of precision and recall—offers a single, interpretable metric that penalizes both false positives and false negatives equally.

Table H summarizes the F1-scores obtained on three benchmark datasets: ColorMNIST, WaterBirds, and OfficeHome. For each dataset, we group results by source (train) domain and report per-shift F1-scores. Across all settings, DualTTA consistently outperforms alternative adaptation strategies, demonstrating average gains of 0.03–0.06 in F1-score over the strongest baseline. Notably, on Office-Home (Train : A $\rightarrow$ P, A $\rightarrow$ R) and WaterBirds, DualTTA achieves improvements of up to 0.05, indicating enhanced robustness under severe domain shifts.

These results confirm that the dual-threshold mechanism of DualTTA not only maintains high overall accuracy but also yields balanced and reliable predictions, as evidenced by the substantial F1-score increases. By reporting both accuracy and F1-score, we offer a comprehensive view of model behavior, ensuring that high performance is not achieved at the expense of one error type over another.

Office-Home: Art (A), Clipart (C), Product (P), Realworld (R)																	
Methods	**Train domain → Test domain**			**Train : A**	**Train domain → Test domain**			**Train : C**	**Train domain → Test domain**			**Train : P**	**Train domain → Test domain**			**Train : R**	**Avg**
	A→C	A→P	A→R	Mean	C→A	C→P	C→R	Mean	P→A	P→C	P→R	Mean	R→A	R→C	R→P	Mean	
ResNet50-BN	43.16	60.13	72.23	58.51	53.61	61.05	65.23	59.96	52.45	42.27	72.71	55.81	64.07	47.49	75.26	62.28	59.14
+TENT	47.08	62.83	72.89	60.93	55.17	63.17	**66.12**	61.49	54.39	45.18	**73.58**	57.72	65.27	51.00	76.28	64.18	61.08
+SAR	44.74	61.00	72.37	59.37	54.47	62.09	65.96	60.84	53.85	42.13	72.30	56.09	65.22	48.25	75.49	62.99	59.82
+EATA	35.60	52.29	67.98	51.96	50.43	52.83	60.16	54.47	48.95	28.55	69.38	48.96	61.52	32.33	68.12	53.99	52.34
+DEYO	43.96	61.03	71.33	58.77	53.98	61.14	65.16	60.09	53.11	41.42	71.61	55.38	64.40	46.69	75.13	62.07	59.08
+DualTTA	**49.98**	**63.12**	**72.93**	**62.01**	**55.67**	**63.17**	65.99	**61.61**	**54.35**	**46.19**	73.28	**57.94**	**65.64**	**51.30**	**76.50**	**64.48**	**61.51**
PACS: Art (A), Cartoon (C), Photo (P), Sketch (S)																	
Methods	**Train domain → Test domain**			**Train : A**	**Train domain → Test domain**			**Train : C**	**Train domain → Test domain**			**Train : P**	**Train domain → Test domain**			**Train : S**	**Avg**
	A→C	A→P	A→S	Mean	C→A	C→P	C→S	Mean	P→C	P→A	P→S	Mean	S→C	S→P	S→A	Mean	
ResNet50-BN	75.73	96.65	70.20	80.86	81.98	95.21	73.71	83.63	64.29	78.32	46.40	63.00	68.64	57.78	59.13	61.85	72.34
+TENT	77.39	96.89	72.79	82.36	84.96	95.49	78.19	86.28	67.75	78.81	48.39	64.98	70.44	58.32	60.89	63.22	74.96
+SAR	76.92	96.89	71.72	81.84	83.30	95.45	73.81	84.19	65.87	77.88	47.52	63.76	70.01	58.26	61.47	63.25	73.26
+EATA	75.60	97.25	70.09	80.98	83.35	94.85	73.43	83.88	65.36	78.37	46.70	63.48	68.13	57.72	58.69	61.52	72.46
+DEYO	**78.24**	**97.49**	67.93	81.22	85.89	95.27	78.37	86.51	**71.33**	**79.35**	52.97	67.88	71.72	59.20	**64.01**	**65.04**	75.16
+DualTTA	77.90	97.37	**72.59**	**82.62**	**86.67**	**95.51**	**78.64**	**86.94**	70.92	79.00	**59.99**	**69.90**	**72.12**	**59.22**	62.55	64.63	**76.02**

Table E. Detail performance of DualTTA on Office-Home and PACS datasets.

Figure B. Examples of objects cropped from images in the FSC147 dataset. The number of objects per class differs significantly, resulting in a highly imbalanced dataset. Moreover, the varying sizes of the bounding boxes require resizing them to a unified resolution, which causes some images to become blurred or degraded in quality.

I. Statistical Significance Analysis

To evaluate the statistical significance of the performance gains brought by our proposed DualTTA framework, we conducted a Wilcoxon signed-rank test against several baseline methods, including Tent, SAR, EATA, and DeYO. The test was applied across all experiments reported in Table 1, 2, 3 in the mainscript, encompassing results on ImageNet-C (15 corruptions, 3 types of normalization layers), ColoredMNIST, Waterbirds, Office-Home, and PACS.

Specifically, we compared the accuracy of DualTTA against each baseline on a per-condition basis (i.e., per corruption type or domain split).

The resulting p-values from the Wilcoxon signed-rank test are as follows:

- **DualTTA vs Tent:** $p = 5.14 \times 10^{-11}$
- **DualTTA vs SAR:** $p = 1.13 \times 10^{-6}$
- **DualTTA vs EATA:** $p = 3.19 \times 10^{-5}$
- **DualTTA vs DeYO:** $p = 1.38 \times 10^{-5}$

All comparisons yield p-values significantly below the conventional threshold of 0.05, confirming that the improvements achieved by DualTTA over all baseline methods are statistically significant. This demonstrates that the observed performance gains are unlikely to be due to random variation and validates the robustness of our approach

Methods	Noise			Blur				Weather				Digital				Avg
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright.	Contr.	Elastic	Pixel	JPEG	
ViTBase-LN	64.36	62.06	63.92	52.31	39.63	60.75	46.39	47.27	38.31	71.84	74.50	77.64	69.36	70.60	69.13	62.13
+ Tent	70.69	70.10	70.53	66.97	60.45	70.31	61.40	63.82	59.79	73.32	77.51	79.30	75.06	74.97	73.18	70.54
+ SAR	69.64	68.19	69.38	63.59	57.95	67.89	61.15	61.70	58.82	75.65	76.78	78.93	73.28	73.36	71.54	69.22
+ EATA	70.15	69.96	70.22	66.79	64.09	70.23	66.78	68.19	64.72	77.71	78.63	**80.19**	76.10	76.45	74.19	72.12
+ DeYO	70.98	71.15	71.11	68.18	65.81	**72.17**	67.08	**70.36**	66.94	76.87	**79.84**	79.47	77.15	**77.61**	**75.52**	73.09
+ DualTTA	**71.14**	**72.21**	**71.33**	**68.29**	**68.58**	72.12	**67.16**	70.23	**66.98**	**79.75**	79.77	79.35	**77.47**	77.51	75.38	**73.59**
ResNet50-GN	56.78	54.96	54.55	44.33	21.06	49.63	39.28	54.45	53.53	56.47	75.01	69.81	59.29	59.76	66.63	54.37
+ Tent	58.99	58.72	58.15	37.23	30.15	54.82	42.74	48.07	29.09	62.41	74.85	69.78	62.96	63.41	66.05	54.49
+ SAR	62.61	62.29	62.01	50.50	39.64	59.85	52.26	58.80	55.04	66.70	**76.74**	71.96	68.97	67.05	68.38	61.50
+ EATA	61.85	61.96	61.00	52.91	**45.96**	61.03	53.93	60.34	**57.15**	68.00	75.54	72.33	70.13	68.97	69.62	62.71
+ DeYO	**64.56**	**65.65**	**64.67**	**53.91**	33.94	60.89	54.41	56.91	51.42	64.04	73.25	71.10	69.54	69.23	68.50	61.47
+ DualTTA	64.41	64.77	63.91	53.30	41.74	**63.56**	**57.99**	**62.68**	48.04	**69.92**	76.47	**73.37**	**71.82**	**71.05**	**70.39**	**63.53**

Table G. Model performance on ImageNet-C with corruption level 3. The best results are colored **bold red**.

across diverse datasets and corruption settings.

References

[1] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 4, 5

Method	ColorMNIST	WaterBirds	Train : A			Train : C			Train : P			Train : R		
			A → C	A → P	A → R	C → A	C → P	C → R	P → A	P → C	P → R	R → A	R → C	R → P
NoAdapt	0.4561	0.8259	0.4210	**0.5956**	**0.6579**	0.4354	0.5404	0.5645	0.4574	0.382	0.6503	0.5742	0.4322	0.7054
TENT	0.3683	0.8428	0.4108	0.5821	0.6431	0.4922	0.5648	0.5787	**0.4840**	0.4118	0.6487	0.5874	0.4676	0.6955
EATA	0.4525	0.8252	0.3916	0.4790	0.5326	0.4727	0.5521	0.5813	0.4637	0.3879	0.6441	0.5801	0.4480	0.6864
SAR	0.4213	0.8297	0.4003	0.4822	0.5349	0.4762	0.5151	0.5370	0.4642	0.3742	0.6440	0.5735	0.4358	0.6871
DEYO	0.7746	0.8746	0.3942	0.5568	0.6317	0.4783	0.5413	0.5695	0.4566	0.3384	0.6399	0.5725	0.4193	0.6790
DualTTA	**0.8147**	**0.8828**	**0.4347**	0.5920	0.6524	**0.5037**	**0.5762**	**0.5929**	0.4795	**0.4259**	**0.6589**	**0.6017**	**0.4796**	**0.7054**

Table H. Comparison of methods (F1-score) on ColorMNIST, WaterBirds and OfficeHome (grouped by train domain).

Algorithm 1 Dual Strategies Test-Time Adaptation (DualTTA)

Require: Pretrained model f_θ, test batch $\mathcal{D}^{tst} = \{x_i\}_{i=1}^{B}$, semantic-altering threshold τ^{sa}, semantic-preserving threshold τ^{sp}, normalization factors $\text{Ent}_0, \text{Diff}_0$, trade-off λ, step size η

Ensure: Adapted parameters θ'

Compute initial predictions and entropies:
for $x \in \mathcal{D}^{tst}$ **do**
  $\hat{y} \leftarrow f_\theta(x)$
  $\text{Ent}(x) = -\sum_c \hat{y}_c \log \hat{y}_c$
end for
Partition into aligned/misaligned sets:
$\mathcal{D}^+ \leftarrow \emptyset, \quad \mathcal{D}^- \leftarrow \emptyset$
for $x \in \mathcal{D}^{tst}$ **do**
  Generate content-perturbed x^{sa} and style-perturbed x^{sp}
  $\hat{y}^{\text{sa}} \leftarrow f_\theta(x_{\text{sa}}), \quad \hat{y}^{\text{sp}} \leftarrow f_\theta(x^{\text{sp}})$
  $k \leftarrow \arg\max_c \hat{y}_c$
  $\text{Diff}_{\text{sa}} = \hat{y}_k - (\hat{y}^{\text{sa}})_k$
  $\text{Diff}_{\text{sp}} = \hat{y}_k - (\hat{y}^{\text{sp}})_k$
  if $\text{Diff}_{\text{sa}} > \tau^{\text{sa}}$ **and** $\text{Diff}_{\text{sp}} < \tau^{\text{sp}}$ **then**
    $\mathcal{D}^+ \leftarrow \mathcal{D}^+ \cup \{x\}$
  else if $\text{Diff}_{\text{sa}} < \tau^{\text{sa}}$ **and** $\text{Diff}_{\text{sp}} > \tau^{\text{sp}}$ **then**
    $\mathcal{D}^- \leftarrow \mathcal{D}^- \cup \{x\}$
  end if
end for
Compute weighted entropy losses:

$$L^+ = \sum_{x \in \mathcal{D}^+} \left(\exp(\text{Ent}_0 - \text{Ent}(x)) + \exp(\text{Diff}_{\text{sa}}) + \exp(\text{Diff}_0 - \text{Diff}_{\text{sp}})\right) \text{Ent}(x)$$

$$L^- = \sum_{x \in \mathcal{D}^-} \exp(\text{Ent}_0 - \text{Ent}(x)) \, \text{Ent}(x)$$

Update model parameters:

$$L_{\text{Dual}} = L^+ - \lambda L^-, \quad \theta \leftarrow \theta - \eta \nabla_\theta L_{\text{Dual}}$$

Return $\theta' \leftarrow \theta$