Back to Source: Open-Set Continual Test-Time Adaptation via Domain Compensation

Supplementary Material

In this appendix, we provide detailed supplementary materials to further clarify and support our framework. We begin with additional analysis of DOCO, where we present the full algorithmic procedure, discuss its connection to domain compensation and feature disentanglement, and examine how the learned prompts generalize to unseen domains with extended visualizations. We then describe implementation details, including the construction of corrupted datasets, the configurations of all baselines, and practical considerations such as batch-size stabilizers and the use of source-domain samples. Finally, we report extended experimental results, covering computational efficiency, robustness under different OOD ratios and corruption severities, and comprehensive comparisons across multiple OOD score measurements to validate the stability and effectiveness of DOCO in the OCTTA setting.

A. Additional Analysis of DOCO

A.1. Algorithm

As mentioned in *Implementation Details*, the prompts conduct a one-time self-supervised update for 50 iterations to refine their initial state. For all subsequent batches we reuse the prompt and perform only a single gradient step.

A.2. End-to-End Domain Compensation

Two feature-level routes. Pixel-space restoration (g^{-1}) could in principle clean inputs before feature extraction, but its ill-posedness risks artifacts propagating to features; we therefore focus on *feature-level* compensation. A representative *explicit separation* route is DICS [28]: during *training*, it learns domain vectors and subtracts them while enforcing same-class cross-domain consistency (DIT), and further promotes class specificity via a memory-driven soft labeling (CST), then *deploys a fixed model* without using the target stream. In contrast, our route performs *test-time, in-process* correction: within each batch t, we estimate the shared factor δ_t using only likely ID samples, and immediately propagate the learned prompt p_{t+1} to the whole batch during the forward pass, yielding $\phi(x; p_{t+1}) \approx \phi(x) - \delta_t \approx s(x)$ for both ID and OOD candidates from the same batch. This *batch-consistent* compensation leverages the live stream *inside* the feature extractor and avoids back-propagating through likely OOD samples.

Relation to DICS. Both routes aim to expose $s(x)$ by attenuating domain factors. Empirically, DICS realizes this via *training-time* explicit subtraction plus class-specific constraints, whereas we realize a *test-time* compensation conditioned on the current batch and updated online *without* backprop on likely-OOD data—thereby preventing OOD semantics from contaminating alignment and stabilizing the decision boundary under a frozen head.

Algorithm 1 DOmain COmpensation (DOCO)

Require: Model $f_\theta = h \circ \phi$; source labels $\mathcal{Y}^S$; cached source statistics (μ_S, σ_S); frozen classifier weights $\{w_c\}_{c\in\mathcal{Y}^S}$; test stream $\{\mathcal{B}_t\}_{t=1}^T$; learning rate η; regularization weight β.
Ensure: Predictions on all batches and updated prompts $\{p_t\}$.

Initialize the first prompt p_1 with Xavier-uniform initialization.
First-batch initialization ($t = 1$)**:**
Compute raw features $Z_{1,\text{raw}} = \{\phi(x)\}_{x\in\mathcal{B}_1}$.
Compute raw prototypical distances by Eq. (9) on $Z_{1,\text{raw}}$.
Run K-Means ($K = 2$) on the raw score set $\mathcal{S}_1^{\text{raw}} = \{d_{\text{proto}}(\phi(x)) \mid x \in \mathcal{B}_1\}$, and split $\mathcal{B}_1$ into $\hat{\mathcal{B}}_1^{\text{ID}}$ and $\hat{\mathcal{B}}_1^{\text{OOD}}$ analogously to Eq. (11).
Predict samples in $\hat{\mathcal{B}}_1^{\text{ID}}$ using the raw model $h(\phi(x))$.
for $k = 1$ to 50 **do**
    Update p_1 on $\hat{\mathcal{B}}_1^{\text{ID}}$ by minimizing $\mathcal{L}_{\text{DOCO}}$ in Eq. (6).
end for
Set $p_2 \leftarrow p_1$.
Predict samples in $\hat{\mathcal{B}}_1^{\text{OOD}}$ by Eq. (7) using p_2.
for $t = 2$ to T **do**
    Compute prompted features $Z_{t,p} = \{\phi(x; p_t)\}_{x\in\mathcal{B}_t}$.
    Compute prototypical distances by Eq. (9) on $Z_{t,p}$.
    Run K-Means ($K = 2$) on $\mathcal{S}_t = \{d_{\text{proto}}(z) \mid z \in Z_{t,p}\}$, and obtain $\hat{\mathcal{B}}_t^{\text{ID}}$ and $\hat{\mathcal{B}}_t^{\text{OOD}}$ by Eq. (11).
    Predict samples in $\hat{\mathcal{B}}_t^{\text{ID}}$ using the current prompt p_t.
    Update the prompt on $\hat{\mathcal{B}}_t^{\text{ID}}$ by one gradient step on Eq. (6) to obtain p_{t+1}.
    Predict samples in $\hat{\mathcal{B}}_t^{\text{OOD}}$ by Eq. (7) using p_{t+1}.
end for

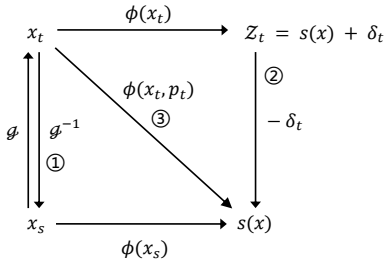


Figure 9. Three pathways from a corrupted image x_t to a domain-invariant feature $s(x)$: (1) pixel restoration (briefly noted), (2) *training-time* explicit separation (DICS), and (3) our *test-time* in-process correction that learns a batch-conditioned prompt inside $\phi(\cdot, p)$.

A.3. Generalization to Unseen Domains

We evaluate whether the learned prompt generalizes across unseen domains *before* any update on the new domain. For each domain transition in a sequence, we take the very first target batch (except the first domain for which the prompt is initialized) and compute the statistical misalignment $\mathcal{L}_{stat}$ against pre-cached source statistics. We compare (i) the static **Source** model and (ii) **DOCO** carrying the prompt updated on *previous* domains but untouched on the current one. On both ImageNet-C and LAION-C streams (see Fig. 11 and Fig. 12; six random orders are examined for each), DOCO consistently exhibits a lower initial $\mathcal{L}_{stat}$, indicating a zero-backprop corrective effect that transfers to novel domains. In a few difficult transitions the initial gap is small, yet the loss still decreases rapidly without degradation, suggesting the prompt provides a beneficial starting point rather than causing negative transfer.

In short, the prompt functions as a batch-wise domain compensator that generalizes to new domains at encounter time, aligning features toward the source geometry and enabling stable adaptation in OCTTA.

A.4. Extended visualization results.

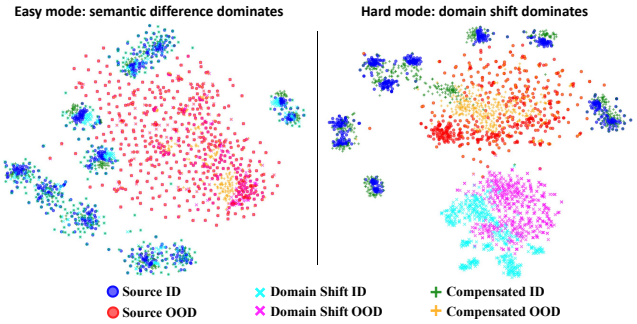


Figure 10. Semantic and domain shift antagonism

To further visualize this internal mechanism, we present a t-SNE [26] visualization of the feature space in Fig. 10, contrasting a mild domain shift (brightness) with a severe one (fog). In the **Hard mode**, the severe domain shift overwhelms the semantic differences. This causes the features of both shifted ID (×) and OOD (×) samples to drift significantly from their origins and mix together in the feature space. Crucially, DOCO effectively reverses this effect, pulling the compensated features (+ and +) back to align with their corresponding source ID (●) and OOD (●) clusters. Conversely, in the **Easy mode**, the intrinsic semantic differences dominate the mild domain shift. Here, aided by our pairwise structural regularizer, DOCO demonstrates its precision by ensuring the compensated features (+ and +) remain tightly anchored to their respective clusters (● and ●) without introducing distortion. This confirms DOCO's dual ability to robustly correct large shifts while delicately preserving feature structures under smaller ones.

B. Implementation Details

B.1. Dataset Corruption Settings.

The LAION-C benchmark features six highly challenging domains: *Mosaic, Glitched, Vertical Lines, Geometric Shapes, Stickers,* and *Luminance Checkerboard*. Example corruptions are shown in Fig. 13. To evaluate covariate-shifted OOD robustness, we applied these synthetic corruptions to OOD datasets. The specific generation settings for `mosaic` and `sticker` are as follows:

- **Tile Pool Source**: We used the ImageNet-1K (`ILSVRC2012`) validation set as the tile pool for corruption generation due to its diversity.
- **Tile Pool Subsampling**: To manage memory constraints, we subsampled 5000 images from the 50,000-image validation set, sequentially selected and packaged into a `.tar` archive using the `WebDataset` format, as required by the LAION-C data loader.
- **Corruption Generation**: Corruptions were applied using the curated 5000-image tile pool. All parameters, such as `intensity_level`, and generation protocols followed the default behavior of the LAION-C codebase, except for the specified sub-sampling strategy.

We use a fixed random seed when subsampling the 5,000 validation images and generating corruptions to ensure that LAION-C benchmarks are fully reproducible.

B.2. Baselines.

For a fair and reproducible comparison, we implement all baseline methods using their official, publicly available codebases. We initialize all hyperparameters and learning rates for each algorithm strictly according to the configurations recommended by the original authors. The implementation of OSTTA is taken from the official UniEnt repository. For the ViDA baseline, we evaluate two backbone settings: (i) a standard pre-trained model from the `timm` library, and (ii) the pre-trained model released in the official ViDA repository. In the latter case, the low-rank and high-rank ViDA modules are pre-trained, providing a much better initialization than random parameters. In our final results, we report the performance of the ViDA variant that achieves the higher score between these two configurations. We consider both backbone settings to avoid penalizing ViDA due to implementation differences, and always report the better one, while all other baselines are evaluated with their official configurations. For STAMP, we follow its experimental setup on the ImageNet benchmark and remove the consistency filtering mechanism to avoid discarding too many samples. For E-COME, UniEnt, EATA, and DPCore, which require collecting information from the

source domain beforehand, we also adhere to their default settings. Specifically, for E-COME, UniEnt, and EATA, the number of source samples used to compute the Fisher information matrix is set to 2000, while DPCore uses 300 source samples by default. To ensure a fair comparison, DOCO is likewise restricted to only 300 source-domain samples. Moreover, to keep the number of parameters comparable, although DOCO can obtain better performance with a larger prompt number (Fig. 7a), in all main experiments we fix the prompt length to $L = 8$, which matches the configuration used by DPCore.

B.3. Details for Batch Size and Source Number

Note on omitted points. We provide the data integrity explanation of Fig. 7b here for further understanding. **DPCore** at small batches (BS=2/4/8) on mixed data exhibits core-set blowup due to unstable per-batch statistics, rendering runs infeasible. **EATA-based** variants (EATA, E-COME, UniEnt) at BS=128 are omitted because their GPU memory cost is prohibitive.

Small-batch stabilizers in DOCO. In the small-batch regime (test batch size ≤ 8) discussed in §*Effect on batch size and source number*, DOCO enables two lightweight stabilizers that are only activated in this analysis and are disabled in all main results. (1) A FIFO buffer $\mathcal{R}$ of a fixed size (we use 64 recent values in our experiments) stores recent proto-distances $d_{\text{proto}}(z)$; we run k-means (K=2) over scores in $\mathcal{R}$ and use the resulting clusters to assign the current batch to ID/OOD, reducing the variance of the split when batches are tiny. (2) We enforce a minimum of one ID sample to update p_t; otherwise, we skip adaptation and only perform forward prediction. The structure-preserving regularizer $\mathcal{L}_{\text{reg}}$ is the Frobenius norm between the pairwise cosine-similarity matrices of prompted and raw CLS features, and is evaluated only when the ID subset has at least two samples.

More details on source number ablation. As a supplementary view of Fig. 7c, we additionally provide a 2D source number comparison visualization on Fig. 14.

C. Extended Experimental Results

C.1. Computational efficiency

Tab. 5 reports runtime and memory under the same protocol as the main results: batch size 64 on **LAION-C (sev=3)** and **ImageNet-C (sev=5)**. Numbers are shown as "LAION-C/ImageNet-C". *Time* is a relative measure normalized to Tent = 1.0 (lower is faster), and *Memory* is GPU usage (MB). All methods are measured on a single NVIDIA Quadro P6000 GPU under the same implementation, so the relative runtime is directly comparable. Overall, **DOCO** attains the strongest performance in the main tables while keeping *moderate* overhead—its prompt-based updates add little computation compared to methods that retrain normalization layers or adapters. Notably on the harder LAION-C (sev=3), DPCore's core-set rapidly grows during the stream, inflating computation and wall-clock time. These results confirm that our in-process prompt correction offers a favorable accuracy–efficiency trade-off in OCTTA.

Table 5. Comparison w.r.t computational complexity

Method	Update	Memory(MB)	Time	H-s (%)
Tent [34]	Norm	10,094/10,112	1.0/1.0	0.6/23.8
CoTTA [36]	All	20,780/21,354	2.8/4.7	18.5/54.8
EATA [29]	Norm	12,442/12,442	0.9/0.84	27.9/57.8
SAR [30]	Norm	12,542/6,800	1.5/1.6	9.6/54.3
OSTTA [20]	Norm	12,944/12,618	1.9/1.9	17.5/58.5
ViDA [24]	Adapters	11,812/11,786	8.6/8.7	3.9/48.4
UniEnt [9]	Norm	13,944/13,944	1.4/1.4	29.3/65.4
STAMP [42]	Norm	9,978/10,070	2.8/2.8	19.5/60.2
E-COME [43]	Norm	12,494/12,494	0.7/0.8	19.9/65.2
S-COME [43]	Norm	6,966/6,800	1.6/1.6	0.3/45.5
DPCore [44]	Prompts	11,424/10,700	3.6/2.0	30.3/62.6
DOCO (Ours)	Prompts	14,694/16,604	2.1/1.9	32.7/70.1

C.2. Different OOD percentage

We thoroughly analyze the impact of varying OOD sample percentages on model performance, with overview Tab. 6 and detailed results for OOD ratios $10\% - 40\%$ in Tab. 7 - Tab. 10. DOCO demonstrates strong robustness, delivering consistently high accuracy across all tested OOD ratios, and in particular achieves a $\mathbf{5}\%$ favorable improvement over the next-best method DPCore when $\kappa = 0.4$.

C.3. Different severity experiment

Similarly, we test our method on the LAION-C benchmark with a lower corruption severity level of 1, while keeping the OOD ratio at $\kappa = 0.5$. As shown in Tab. 11, DOCO continues to outperform other methods, securing the highest average metrics, surpassing the second by $\mathbf{3.4}\%$. This demonstrates that DOCO's effectiveness is not limited to extreme domain shifts but also holds in scenarios with more subtle corruptions, confirming its consistent superiority.

C.4. Different OOD Score Measurement

In the main paper, we adopt the energy-based OOD score as the default choice for computing AUC and H-score. To verify that our conclusions are not tied to a particular score, we further evaluate all methods under three additional mainstream post-hoc OOD scores, including entropy, Max Logit (MLS), and maximum softmax probability (MSP). As summarized in Tab. 12, DOCO consistently achieves the best H-score under all four score functions and exhibits only minor variation across them, whereas the strongest competing method reaches at most 65.36%. These results indicate that DOCO is insensitive to the specific OOD score used for evaluation and remains clearly ahead of existing baselines across different OOD score measurements.

Table 6. Results (%) for ImageNet-to-ImageNet-C benchmark (severity = 5) in OCTTA setting with different OOD samples percentages across six covariate-shifted OOD datasets. All the results are averaged over 15 domains.

Method	10%			20%			30%			40%			50%			Avg.		
	ACC	AUC	H-s	ACC	AUC	H-s	ACC	AUC	H-s	ACC	AUC	H-s	ACC	AUC	H-s	ACC	AUC	H-score
Source	49.7	67.9	56.3	49.8	68.1	56.4	49.8	68.2	56.4	49.8	68.1	56.4	49.8	68.0	56.4	49.8	68.1	56.4
Tent [34]	51.2	65.3	55.0	49.2	60.3	51.8	30.0	54.4	31.9	29.1	53.8	31.3	22.4	50.9	23.8	36.4	56.9	38.8
CoTTA [36]	49.9	67.5	56.4	49.9	67.0	56.2	49.8	66.6	56.0	49.8	65.8	55.6	49.5	64.5	54.8	49.8	66.3	55.8
EATA [29]	58.6	70.6	63.5	58.5	70.1	63.3	56.6	69.4	61.2	55.0	67.8	59.6	52.9	67.3	57.8	56.3	69.0	61.1
SAR [30]	57.1	71.9	63.2	56.1	69.2	61.4	54.0	66.8	58.7	52.8	64.5	57.2	50.4	61.5	54.3	54.1	66.8	59.0
OSTTA [20]	58.5	69.3	63.0	58.2	67.3	62.0	57.8	65.7	61.1	57.3	64.2	60.2	56.2	61.9	58.5	57.6	65.7	61.0
ViDA [24]	56.0	70.8	62.1	55.3	63.5	58.4	54.8	57.6	55.0	53.9	52.5	51.6	53.0	47.9	48.4	54.6	58.5	55.1
UniEnt [9]	57.6	75.1	64.5	57.7	75.5	64.7	57.8	76.2	65.1	56.8	76.3	64.2	57.8	77.0	65.4	57.5	76.0	64.8
STAMP [42]	51.3	71.5	59.0	51.5	71.8	59.2	51.3	72.3	59.2	51.9	73.7	60.1	52.0	73.8	60.2	51.6	72.6	59.5
E-COME [43]	58.4	75.7	65.4	58.5	76.0	65.6	57.5	75.1	64.5	53.5	73.1	59.9	58.3	75.5	65.2	57.2	75.1	64.1
S-COME [43]	53.1	69.7	58.9	53.8	70.8	60.0	50.2	67.7	56.1	50.2	68.0	56.4	40.9	63.0	45.5	49.7	67.9	55.4
DPCore [44]	59.0	79.3	67.2	58.3	80.1	66.9	56.8	78.5	65.4	56.2	78.3	64.9	54.1	76.2	62.6	56.9	78.5	65.4
DOCO (Ours)	**61.7**	**79.8**	**69.2**	**62.0**	**81.0**	**69.8**	**61.6**	**81.0**	**69.5**	**61.7**	**82.0**	**69.9**	**61.5**	**82.7**	**70.1**	**61.7**	**81.3**	**69.7**

Table 7. Results (%) for ImageNet-to-ImageNet-C (severity = 5, $\kappa = 0.1$) in OCTTA setting across six covariate-shifted OOD datasets.

Method	Places.–C		Texture–C		iNatur.–C		SUN–C		SSB-H.–C		NINCO–C		Avg.		
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	H-score
Source	49.7	66.5	49.7	70.7	49.7	78.4	49.7	71.6	49.7	56.1	49.7	64.4	49.7	67.9	56.3
Tent [34]	17.2	52.0	57.3	60.0	57.6	78.6	58.0	67.6	58.7	64.8	58.3	68.9	51.2	65.3	54.9
CoTTA [36]	49.9	65.7	49.9	69.9	49.9	77.9	49.9	70.7	50.0	56.4	50.0	64.2	49.9	67.5	56.4
EATA [29]	56.6	66.3	59.5	74.7	58.2	78.6	59.3	73.8	58.4	61.8	59.8	68.2	58.6	70.6	63.5
SAR [30]	57.2	68.0	57.0	74.1	57.0	82.3	57.3	74.5	57.0	62.2	57.1	70.5	57.1	71.9	63.2
OSTTA [20]	58.4	65.7	58.3	66.1	58.5	79.7	58.5	71.8	58.7	63.7	58.7	68.7	58.5	69.3	63.0
ViDA [24]	56.1	71.6	55.5	69.2	55.9	74.2	55.7	69.1	56.4	**66.6**	56.0	73.9	55.9	70.8	62.1
UniEnt [9]	59.1	73.3	58.7	81.3	57.0	84.3	56.8	79.6	55.0	60.7	58.8	71.5	57.6	75.1	64.5
STAMP [42]	51.2	71.0	51.1	71.3	51.4	82.9	51.4	74.8	51.5	59.7	51.2	69.2	51.3	71.5	59.0
E-COME [43]	58.4	75.9	58.2	78.4	58.4	85.3	58.4	82.1	58.2	60.5	58.7	72.0	58.4	75.7	65.4
S-COME [43]	42.4	65.8	54.3	74.4	54.0	75.7	55.3	76.6	55.4	58.5	57.1	67.4	53.1	69.7	58.9
DPCore [44]	61.5	**79.7**	**61.5**	**82.5**	57.8	**93.7**	58.0	83.4	59.1	63.2	56.3	73.4	59.0	79.3	67.2
DOCO (Ours)	**61.7**	76.8	61.1	82.5	**61.9**	92.3	**62.4**	**88.0**	**62.0**	64.4	**61.3**	**74.8**	**61.7**	**79.8**	**69.2**

Table 8. Results (%) for ImageNet-to-ImageNet-C (severity = 5, $\kappa = 0.2$) in OCTTA setting across six covariate-shifted OOD datasets.

Method	Places.–C		Texture–C		iNatur.–C		SUN–C		SSB-H.–C		NINCO–C		Avg.		
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	H-score
Source	49.8	67.0	49.8	70.8	49.8	78.3	49.8	71.6	49.8	56.2	49.8	64.5	49.8	68.1	56.4
Tent [34]	56.9	61.8	31.2	50.7	33.8	57.5	56.8	58.0	58.5	65.5	57.9	68.3	49.2	60.3	51.8
CoTTA [36]	49.9	65.6	49.9	69.0	49.7	77.3	50.0	69.8	49.9	56.3	49.9	64.0	49.9	67.0	56.2
EATA [29]	57.9	66.5	58.5	73.1	58.7	78.4	58.8	73.4	59.2	62.3	57.7	66.9	58.5	70.1	63.3
SAR [30]	56.9	67.5	53.3	67.8	56.0	78.9	56.3	68.1	57.3	62.9	56.8	70.0	56.1	69.2	61.4
OSTTA [20]	58.0	64.5	57.7	62.3	58.0	76.2	58.4	67.8	58.5	64.6	58.3	68.3	58.2	67.3	62.0
ViDA [24]	55.7	69.5	54.5	56.7	55.1	59.6	55.0	57.4	56.2	**66.3**	55.5	71.6	55.3	63.5	58.4
UniEnt [9]	59.1	73.9	59.5	82.1	52.2	82.6	59.4	82.7	56.7	59.8	59.2	71.8	57.7	75.5	64.7
STAMP [42]	51.6	71.6	51.3	71.6	51.4	83.0	51.4	74.7	51.6	60.0	51.5	69.6	51.5	71.7	59.2
E-COME [43]	59.6	76.7	59.1	79.1	59.1	86.4	56.9	81.8	58.7	60.2	57.9	72.1	58.6	76.1	65.6
S-COME [43]	54.6	72.1	44.9	73.1	55.5	77.2	56.8	77.2	55.8	58.4	55.5	66.9	53.8	70.8	60.0
DPCore [44]	58.4	**79.7**	58.4	82.8	57.0	93.3	58.4	87.6	60.4	63.1	57.1	74.4	58.3	80.1	66.9
DOCO (Ours)	**62.1**	78.4	**61.8**	**83.6**	**62.5**	**94.0**	**61.9**	**90.0**	**61.9**	64.7	**62.1**	**75.2**	**62.0**	**81.0**	**69.8**

Table 9. Results (%) for ImageNet-to-ImageNet-C (severity = 5, $\kappa = 0.3$) in OCTTA setting across six covariate-shifted OOD datasets.

Method	Places.–C		Texture–C		iNatur.–C		SUN–C		SSB-H.–C		NINCO–C		Avg.		
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	H-score
Source	49.8	67.0	49.8	71.1	49.8	78.7	49.8	71.7	49.8	56.0	49.8	64.5	49.8	68.2	56.4
Tent [34]	18.2	53.5	11.3	46.4	34.3	58.1	41.6	46.6	17.7	54.4	56.9	67.6	30.0	54.4	31.9
CoTTA [36]	49.8	65.2	49.7	68.3	49.8	77.5	50.0	69.1	50.1	55.9	49.6	63.5	49.8	66.6	55.9
EATA [29]	56.0	66.2	56.6	72.2	55.4	76.7	54.8	72.8	58.1	60.8	58.6	67.4	56.6	69.4	61.2
SAR [30]	54.8	66.1	50.0	63.6	54.2	78.7	52.8	61.5	57.3	62.2	54.7	68.9	54.0	66.8	58.7
OSTTA [20]	58.1	63.2	57.1	59.8	56.6	74.3	57.7	65.0	58.8	63.9	58.4	68.0	57.8	65.7	61.1
ViDA [24]	55.1	66.7	54.0	48.7	54.6	50.0	53.9	46.7	56.1	**64.5**	55.1	69.2	54.8	57.6	55.0
UniEnt [9]	56.1	73.2	59.0	82.8	59.8	89.8	57.1	81.2	56.6	58.7	58.1	71.8	57.8	76.2	65.1
STAMP [42]	51.1	71.8	51.0	72.7	51.5	83.9	51.2	75.6	51.7	60.0	51.4	70.0	51.3	72.3	59.2
E-COME [43]	58.9	76.5	57.3	78.1	54.5	82.2	58.0	82.6	58.1	59.0	58.4	72.1	57.5	75.1	64.5
S-COME [43]	29.9	57.1	55.4	75.5	54.2	75.1	55.0	76.3	53.4	56.4	53.7	66.1	50.2	67.7	56.1
DPCore [44]	57.5	76.1	55.1	80.5	57.0	89.2	57.3	86.4	58.0	63.5	55.8	75.1	56.8	78.5	65.4
DOCO (Ours)	**61.4**	**79.0**	**61.6**	**83.4**	**61.6**	**94.3**	**61.5**	**89.5**	**62.1**	64.4	**61.4**	**75.3**	**61.6**	**81.0**	**69.5**

Table 10. Results (%) for ImageNet-to-ImageNet-C (severity = 5, $\kappa = 0.4$) in OCTTA setting across six covariate-shifted OOD datasets.

Method	Places.–C		Texture–C		iNatur.–C		SUN–C		SSB-H.–C		NINCO–C		Avg.		
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	H-score
Source	49.8	66.9	49.8	70.9	49.8	78.6	49.8	71.7	49.8	56.2	49.8	64.3	49.8	68.1	56.4
Tent [34]	26.1	51.8	6.0	53.1	14.3	47.2	17.2	41.7	58.2	64.3	52.4	64.8	29.0	53.8	31.3
CoTTA [36]	49.8	64.2	49.4	67.3	49.8	76.6	49.9	67.8	50.1	55.9	49.6	62.8	49.8	65.8	55.6
EATA [29]	55.8	65.6	52.8	69.5	54.8	75.5	54.6	72.1	56.3	58.1	55.8	66.0	55.0	67.8	59.6
SAR [30]	54.6	64.8	44.8	59.0	53.5	70.3	52.8	63.2	57.0	61.8	54.4	68.1	52.8	64.5	57.2
OSTTA [20]	58.0	62.1	56.7	57.7	54.9	71.3	57.1	62.6	58.8	63.8	58.2	67.5	57.3	64.2	60.2
ViDA [24]	54.0	63.2	52.9	40.4	53.9	42.5	52.7	38.7	55.8	63.6	54.0	66.5	53.9	52.5	51.6
UniEnt [9]	57.6	72.7	56.6	81.1	56.3	88.2	55.9	82.7	58.9	62.2	55.3	71.0	56.7	76.3	64.1
STAMP [42]	52.0	72.5	51.9	75.0	51.8	85.4	51.9	77.3	51.9	60.7	52.0	71.1	51.9	73.7	60.1
E-COME [43]	54.8	74.6	56.9	78.0	59.2	86.2	34.4	70.3	59.1	59.1	56.4	70.7	53.5	73.1	59.9
S-COME [43]	54.9	71.6	53.1	73.5	53.9	76.2	36.7	64.5	49.7	56.3	53.2	65.7	50.2	68.0	56.4
DPCore [44]	56.9	77.3	58.7	82.5	50.6	86.5	54.9	83.5	60.0	64.6	56.5	75.6	56.2	78.3	64.9
DOCO (Ours)	**62.4**	**80.1**	**62.1**	**84.6**	**61.3**	**95.5**	**60.5**	**90.8**	**61.9**	**65.2**	**62.1**	**75.8**	**61.7**	**82.0**	**69.9**

Table 11. Results (%) for LAION-C benchmark (severity = 1, $\kappa = 0.5$) in OCTTA setting across six covariate-shifted OOD datasets. All the results are averaged over 6 constantly switching domains. $-\mathtt{L}$ stands for applying LAION-C corruption to OOD dataset.

Method	Places.–L		Texture–L		iNatur.–L		SUN–L		SSB-H.–L		NINCO–L		Avg.		
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	H-score
Source	51.4	63.8	51.4	68.9	51.4	77.7	51.4	72.6	51.4	60.8	51.4	66.5	51.4	68.4	57.2
Tent (ICLR'21)	2.5	47.9	2.8	50.9	1.6	39.6	1.7	54.1	2.5	53.8	2.1	51.9	2.2	49.7	3.5
CoTTA (CVPR'22)	51.3	61.6	51.3	65.0	51.3	72.4	51.2	68.8	51.3	60.8	50.4	64.1	51.1	65.4	55.9
EATA (ICML'22)	61.7	64.5	61.6	72.4	59.9	78.5	60.7	77.0	61.9	62.6	61.9	65.2	61.3	70.0	64.9
SAR (ICLR'23)	29.1	52.2	39.6	56.3	31.6	57.5	32.4	58.0	52.2	61.8	51.5	66.7	39.4	58.7	44.4
OSTTA (ICCV'23)	56.8	58.1	54.6	55.1	52.2	60.5	54.5	60.2	57.9	63.1	57.2	64.1	55.5	60.2	57.0
ViDA (ICLR'24)	30.8	46.8	33.6	45.0	29.1	37.4	29.5	42.1	35.0	62.2	29.7	55.3	31.3	48.1	34.2
UniEnt (CVPR'24)	61.4	70.6	61.2	82.1	61.0	88.2	61.2	85.0	60.7	64.0	60.8	69.3	61.0	76.5	67.3
STAMP (ECCV'24)	51.3	68.0	51.2	67.7	51.2	75.1	51.3	70.9	51.2	62.2	51.2	67.8	51.2	68.6	57.7
E-COME (ICLR'25)	60.5	73.3	48.7	72.2	59.9	86.5	61.0	81.4	60.9	59.6	59.9	70.1	58.5	73.8	64.2
S-COME (ICLR'25)	1.8	46.3	12.4	53.9	16.2	53.2	1.5	55.1	59.6	57.0	48.1	61.9	23.3	54.6	25.1
DPCore (ICML'25)	65.2	**78.8**	62.3	80.4	52.9	86.5	62.9	87.1	64.6	67.8	62.3	**75.9**	61.7	79.4	68.9
DOCO (Ours)	**66.9**	77.6	**65.7**	**82.4**	**64.9**	**94.9**	**65.4**	**89.7**	**66.2**	**68.8**	**64.7**	75.7	**65.6**	**81.5**	**72.3**

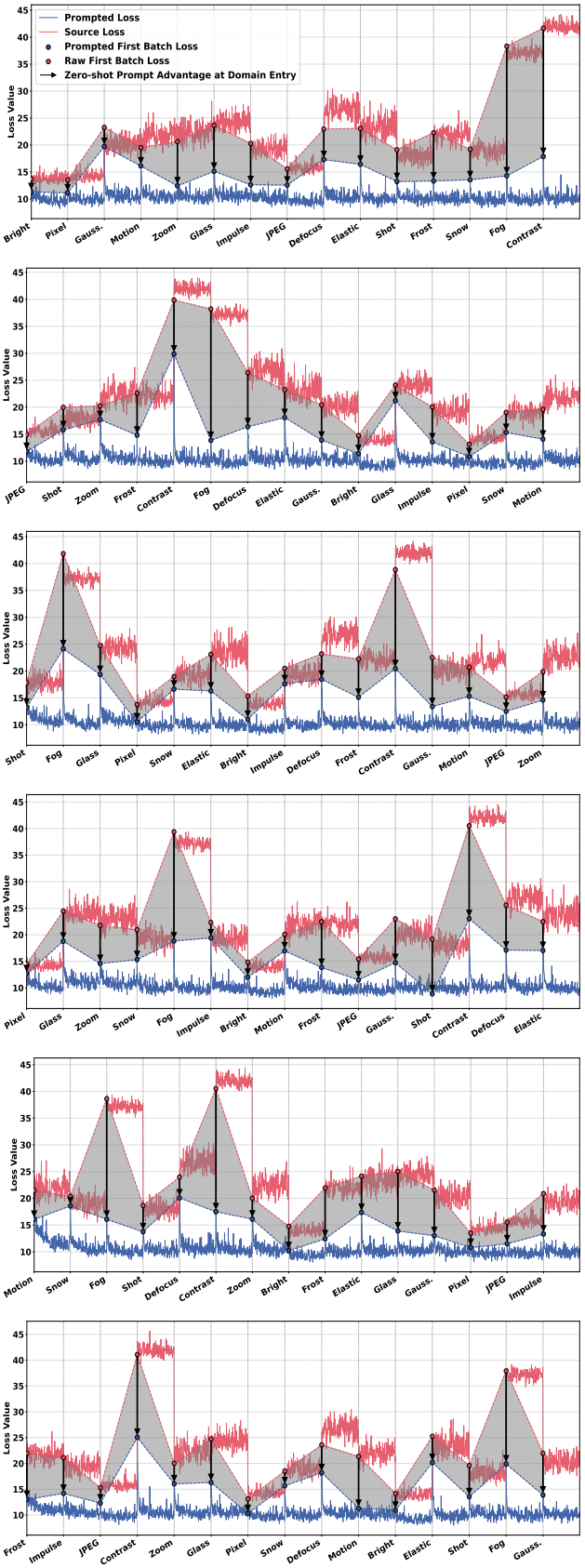


Figure 11. First-batch statistical loss per domain in six different OCTTA orders (ImageNet-C, $\kappa = 0.5$).

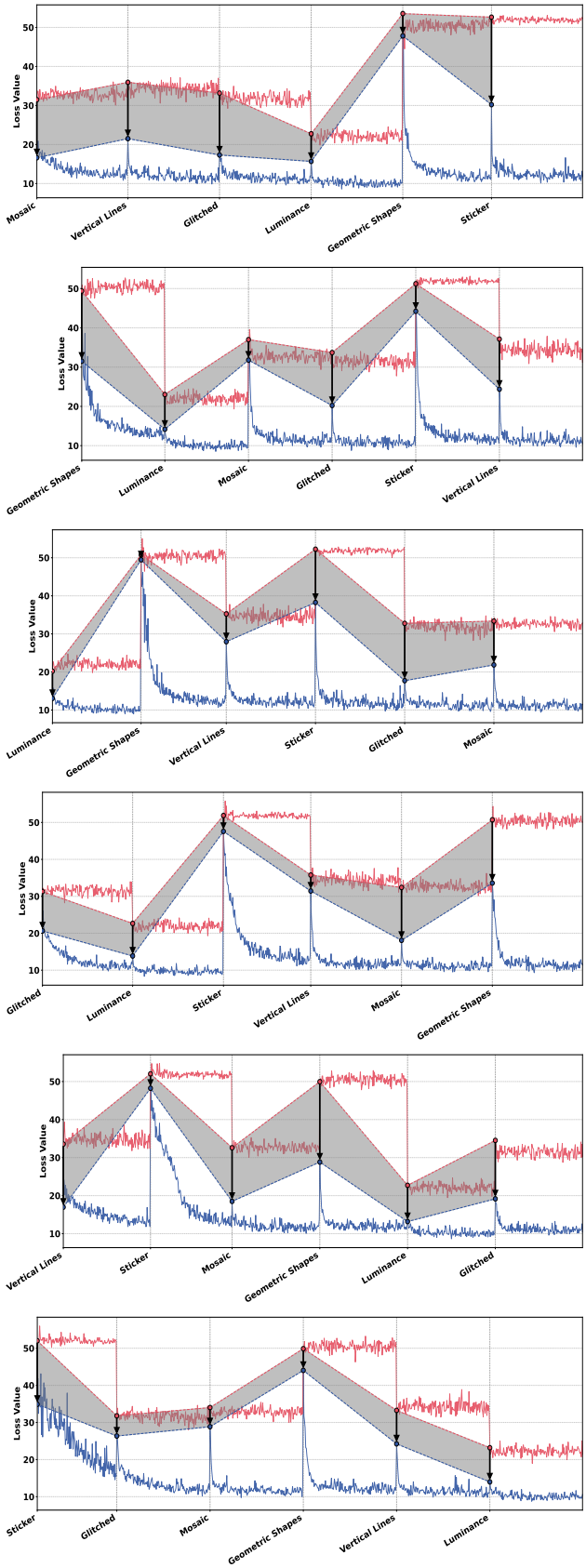


Figure 12. First-batch statistical loss per domain in six different OCTTA orders (LAION-C, $\kappa = 0.5$).

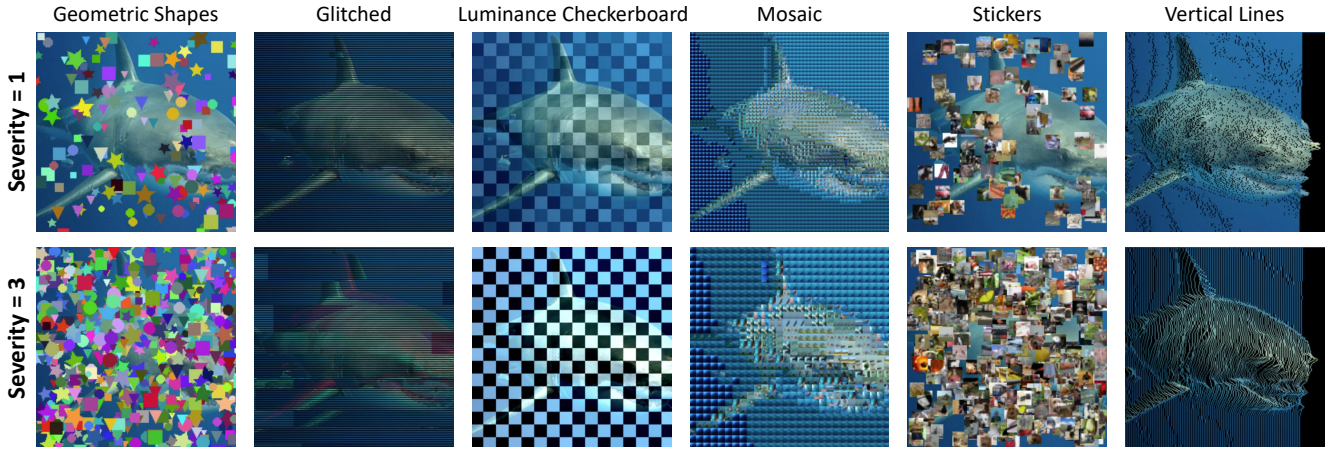
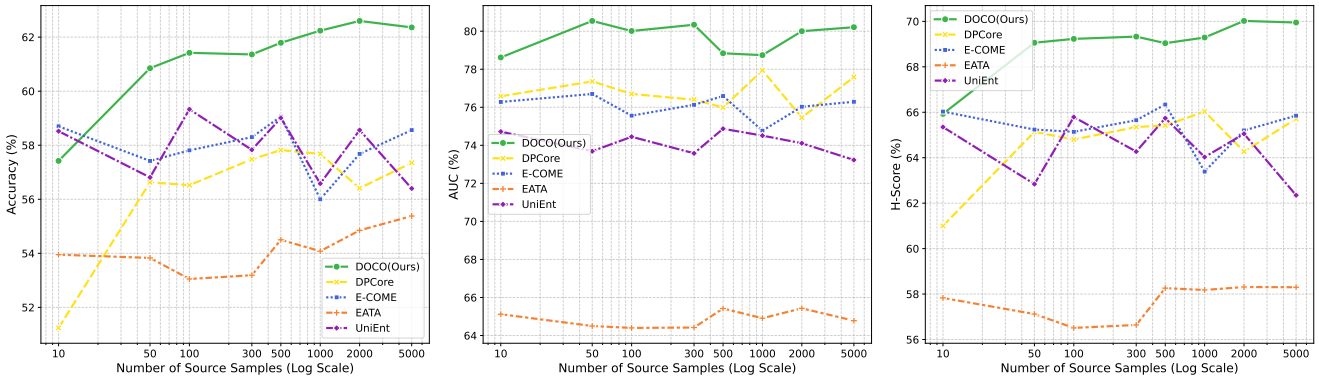


Figure 13. Examples of six LAION-C corruption types.



(a) Accuracy(%) (b) AUC(%) (c) H-score(%)

Figure 14. **Source-number and small-batch ablations.** (a)–(c): Accuracy, AUC, and H-score vs. number of source samples (log scale).

Method	OOD score				Mean±Std
	Ent	MLS	Energy	MSP	
Source	57.29	57.02	56.37	56.96	56.91±0.39
Tent [34]	24.25	23.99	23.82	24.11	24.04±0.18
CoTTA [36]	56.25	55.42	54.79	55.82	55.57±0.62
EATA [29]	57.98	57.84	57.77	57.29	57.72±0.30
SAR [30]	55.10	54.51	54.31	54.63	54.64±0.34
OSTTA [20]	60.39	58.92	58.51	59.82	59.41±0.85
ViDA [24]	48.03	48.22	48.40	47.92	48.14±0.21
UniEnt [9]	64.95	<u>65.25</u>	<u>65.39</u>	64.02	64.90±0.62
STAMP [42]	60.16	59.94	60.18	59.94	60.05±0.13
E-COME [43]	<u>65.36</u>	65.00	65.22	<u>64.73</u>	<u>65.08±0.27</u>
S-COME [43]	45.87	45.53	45.47	45.37	45.56±0.22
DPCore [44]	62.12	61.76	62.62	61.05	61.89±0.66
DOCO (Ours)	**69.57**	**69.38**	**70.10**	**68.45**	**69.38±0.69**

Table 12. H-score results on ImageNet-C with $\kappa = 0.5$, $sev = 5$.