A. Additional analysis

A.1. Loss of generalization

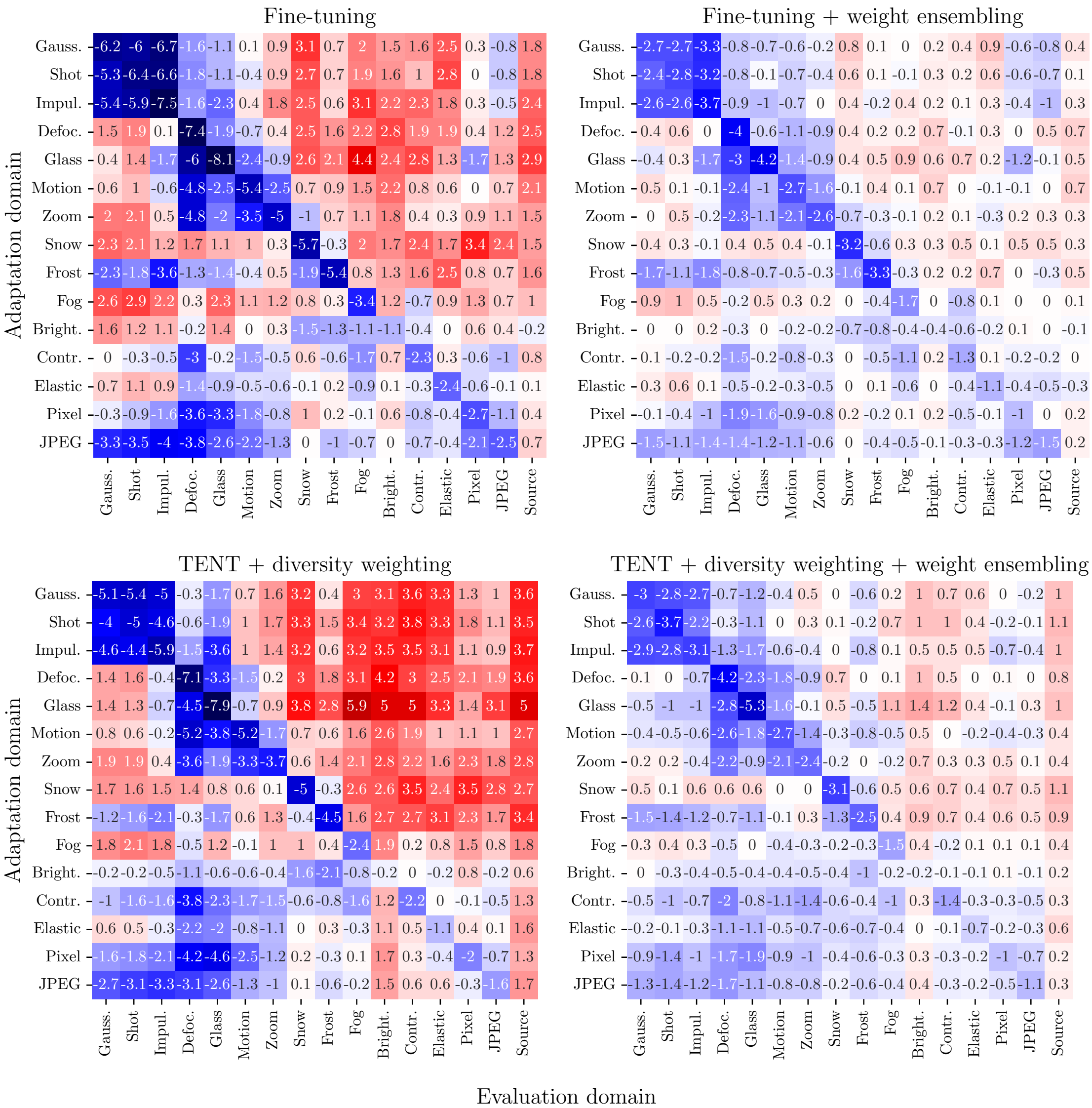

Figure 5. Difference of error for a moderate and a stronger model adaptation corresponding to a learning rate of 10^{-4} and 10^{-3}, respectively. The first row examines supervised fine-tuning, while the second row considers diversity-regularized self-training. The right column further illustrates the effect of adding our weight ensembling approach. All experiments are conducted using an ImageNet pre-trained ResNet-50 that is adapted using 40,000 samples of one of the corruptions from ImageNet-C. The model is then evaluated on the remaining 10,000 samples for all corruptions as well as the source domain. Adapting the model on a potentially narrow distribution can clearly degrade its generalization capabilities. Adding weight ensembling helps to mitigate the loss of generalization as well as catastrophic forgetting.

Since adapting a model to a target domain effectively means moving the model from its initial source parameterization to a parameterization that better models the current target distribution, this should trigger a loss of generalization when the target distribution is narrow. While we have already shown in Section 3 that a generalization loss occurs when performing self-training in the form of entropy minimization, this should also hold when our certainty and diversity weighting from Section 4.1 is further added, or when fine-tuning the model in a supervised manner.

To demonstrate the previous points, we adopt the same setup as before, i.e., we use an ImageNet pre-trained ResNet-50 and adapt the model with 40,000 samples of one of the corruptions from ImageNet-C. Afterwards, the model is evaluated for each corruption and the source domain on the remaining 10,000 samples. Figure 5 illustrates the difference of error for a moderate and a stronger adapted model, corresponding to a learning rate of 10^{-4} and 10^{-3}, respectively. Depending on the investigated corruption, not only fine-tuning but also diversity-regularized self-training result in an increased error on other corruptions, indicating a loss of generalization. This demonstrates the risks of model adaptation in a potentially unknown environment. Using our proposed weight ensembling, a loss of generalization and catastrophic forgetting can mostly be mitigated.

A.2. Model bias and trivial solutions

As stated in Section 3, a critical factor for successful TTA is stability. Current methods for online TTA mostly leverage self-training to adapt the model to the current domain shift, showing great performance on short test sequences [5, 34, 51, 53]. However, if self-training is utilized without any proper regularization, the model is likely to become biased after a while. In the worst case, the bias can even evolve into a trivial solution, where the model only predicts a small subset of classes. In this section, we first demonstrate the aforementioned points for TENT, which exploits entropy minimization for model adaptation. Then, we investigate the behaviour of current state-of-the-art methods, revealing some inefficiencies to effectively counter model bias during test-time.

Long test sequences promote model bias and domain shifts can trigger trivial solutions To investigate whether the model is becoming biased or degrades to a trivial solution during the adaptation, we consider the total variation distance (TVD). It measures the deviation between the actual class prior and the predicted prior. The TVD is defined as

$$d_{\mathrm{TV}}(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \frac{1}{2} \sum_{i=1}^{C} |p_i - \hat{p}_i|, \tag{7}$$

where p_i and $\hat{p}_i$ are the true and predicted prior probability for class i, respectively. If the TVD is calculated along the test sequence, it can also indirectly show the occurrence of error accumulation, since it is a lower bound of the error of the pseudo-labels [21]. Since TENT reports good results for adapting a model to a single domain, we begin our analysis with the same setting and only vary the length of the test sequence by repeating each domain several times. Specifically, we use ImageNet-C with 50,000 samples per corruption and CIFAR100-C with 10,000 samples per corruption (both at severity level 5). Following TENT, we utilize a ResNet-50 with a learning rate $\mathrm{lr} = 2.5\mathrm{e}^{-4}$ for ImageNet-C and a ResNeXt-29 with $\mathrm{lr} = 0.001$ for CIFAR100-C. As shown on the left side of Figure 6, TENT quickly deteriorates to a trivial solution for half of the corruptions of ImageNet-C, while developing a growing bias for the other half. In case of CIFAR100-C, TENT initially deteriorates slightly but then remains stable for most of the corruptions. To study the impact of multiple domain shifts, which is a quite common setting in practice, we leverage all 15 corruption types and create 15 randomly ordered domain sequences. The results for this setting, including different learning rates, are depicted in the middle of Figure 6. Since the TVD now steadily increases in all settings, it becomes clear that domain shifts can explicitly enhance model bias and lead to trivial solutions. If the *domain non-stationarity* is further increased to its maximum, where consecutive test samples are likely to originate from different domains, the TVD increases even more rapidly (right side of Figure 6). Now, by equipping TENT with our certainty and diversity based loss weighting, stable adaptation across all previously considered settings and a wider range of learning rates is possible. The only exception to this is ImageNet-C in the *mixed domains* TTA setting with a learning rate four times higher than the default. This clearly demonstrates that maintaining diversity is crucial in TTA.

Many state-of-the-art methods lack diversity In Figures 7 and 8, we investigate existing TTA methods and our proposed method, namely ROID, in terms of diversity on the continual ImageNet-C benchmark with 50,000 samples per corruption. Figure 7 provides a visual representation of online batch predictions across the entire continual sequence, illustrating the impact on diversity over time and the influence by different domain shifts. Figure 8 depicts the histogram over the predicted classes for the last corruption (JPEG) after adapting the model on the complete continual sequence.

Beginning with BN–1, we observe variations in the degree of model bias induced by different domain shifts. Corruptions where the performance of BN–1 is relatively bad, tend to show a higher model bias. Looking at TENT, a collapse can be seen after a few corruptions, resulting in predicting only a small subset of the 1,000 classes. AdaContrast also strongly lacks diversity after few corruptions. Since LAME solely corrects the model output without updating the model's parameters, the diversity of its predictions heavily relies on the specific type of domain shift. Although LAME maintains diversity for certain corruptions, such as brightness, it collapses for the majority. RoTTA shows the behavior whereby diversity temporarily diminishes for specific domain shifts, such as the transition from *impulse noise* to *defocus blur* and *brightness* to *contrast*. This behavior can likely be attributed to its robust batch normalization, which incorporates past statistics, resulting in bad statistics when past statistics differ from current ones. While SAR demonstrates better diversity than CoTTA and RMT, it still manifests a deficiency in diversity, evident, for example, in the predictions for the final corruption, where a strong bias towards a few classes exists. On the other hand, EATA and ROID with their diversity weighting effectively preserve diversity throughout the adaptation process.

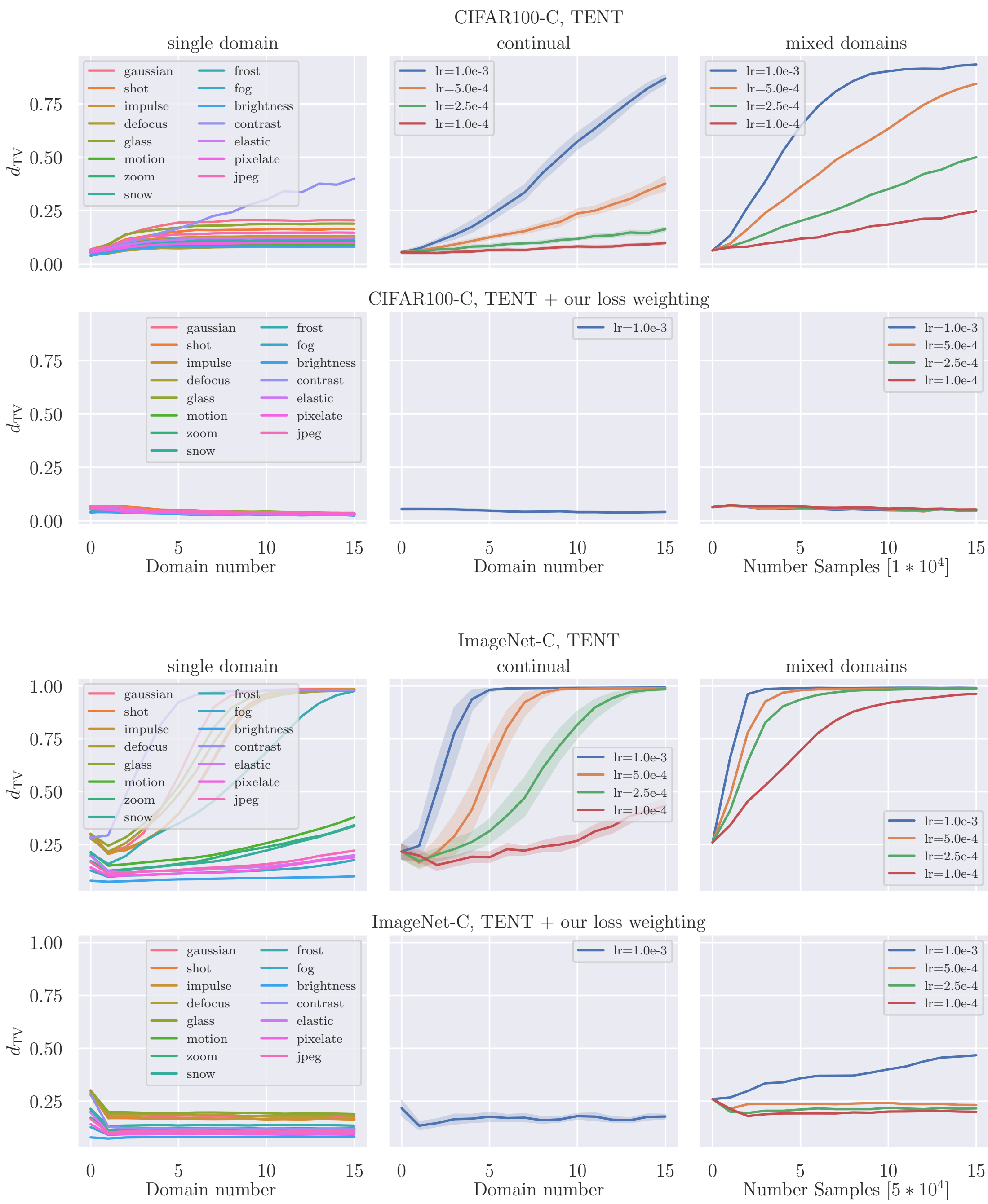

Figure 6. Illustration of the total variation distance of TENT on CIFAR100-C and ImageNet-C at severity level 5 without (first row) and with (second row) our loss weighting. The model is adapted to a single domain (left), in the continual setting (middle) using 15 randomly ordered domain sequences, and the mixed domains setting (right). Unless otherwise stated, TENT's default learning rates of $1.0\mathrm{e}^{-3}$ and $2.5\mathrm{e}^{-4}$ are used. Comparing the left and middle column of CIFAR100-C, it becomes obvious that domain shifts can promote the occurrence of trivial solutions. In case of mixed domains, model bias and trivial solutions occur even faster for both datasets. In contrast, using TENT with our loss weighting prevents the model from becoming biased in almost all settings.

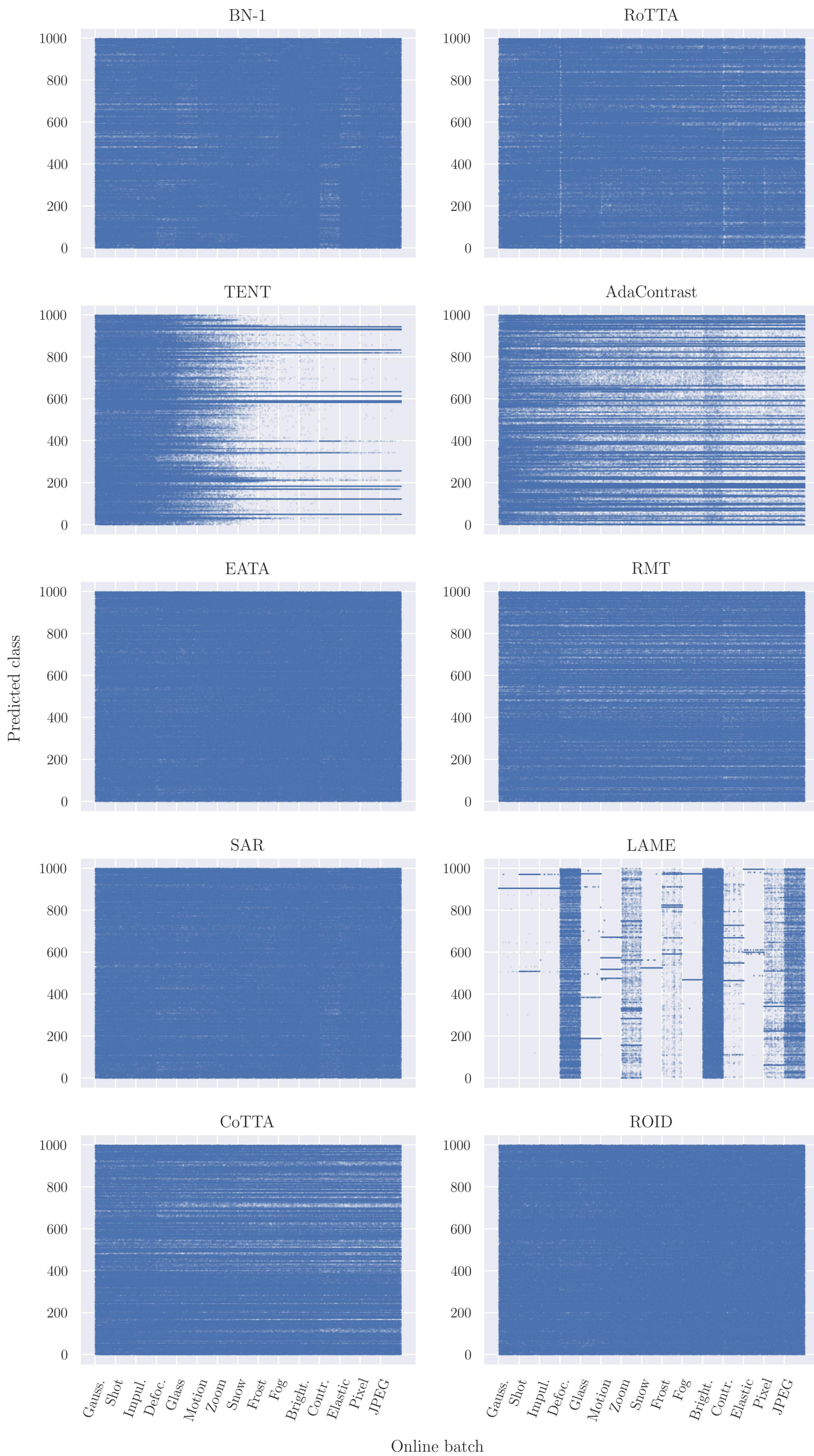

Figure 7. Illustration of the batch-wise predictions in the *continual* TTA setting using a ResNet-50 and ImageNet-C with 50,000 samples per corruption.

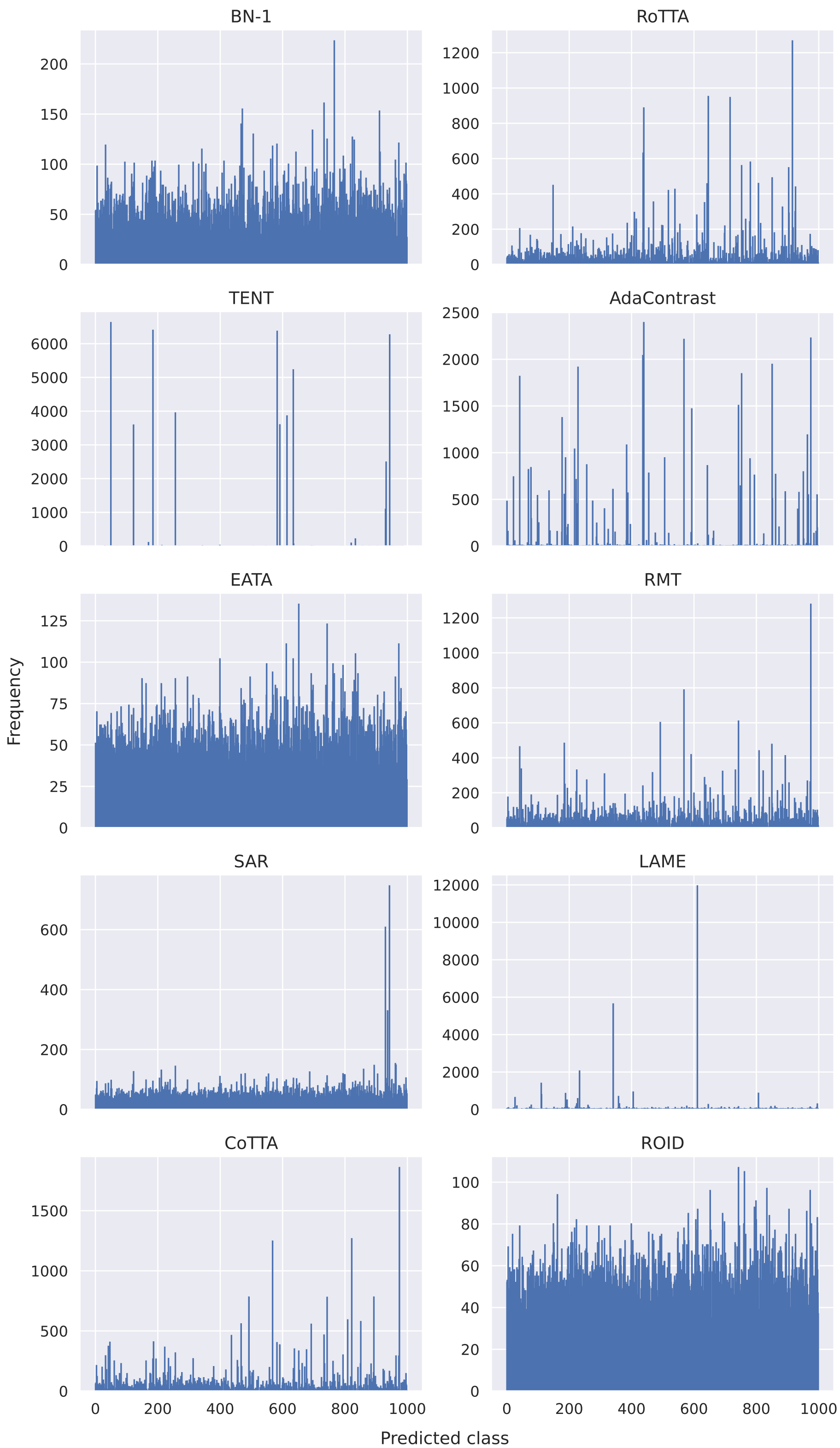

Figure 8. Frequency of the ResNet-50's predictions of the last corruption (JPEG) over the continual TTA sequence using 50,000 samples per corruption.

B. Ablation studies

B.1. Architectures

Table 5. Online classification error rate (%) for the ImageNet benchmarks in the *continual* TTA setting. Common architectures and their variations are considered.

		Inception		ResNet				ResNeXt		WideResNet		DenseNet			RegNetY	
		v1	v3	18	50	101	152	50-32x4d	101-32x8d	50	101	121	169	201	8gf	32gf
	GFLOPs	2	6	1.8	3.8	7.6	11.3	4.2	8.0	-	23	5.7	6.8	8.6	8.0	32.3
	MParams	6.6	27	12	25	44	60	25	44	69	127	8.0	14	20	39	145
IN	Source	30.2	22.7	30.2	23.9	22.6	21.7	22.4	20.7	21.5	21.2	25.6	24.4	23.1	20.0	19.1
IN-C	Source	81.7	76.5	85.3	82.0	77.4	77.6	78.9	75.2	78.9	75.3	78.6	75.7	75.5	78.4	75.9
	BN–1	70.3	69.6	72.7	68.6	66.3	65.9	67.1	64.1	66.0	65.6	68.2	63.7	63.5	67.7	64.4
	ROID	67.8	64.2	62.2	54.5	50.4	49.2	50.9	46.6	50.7	49.0	55.8	51.2	50.6	50.9	46.4
IN-R	Source	63.8	62.2	67.0	63.8	60.7	58.7	62.3	57.4	61.4	59.6	62.8	60.4	59.2	60.0	57.9
	BN–1	61.5	63.9	65.1	60.3	57.7	56.1	59.3	56.2	59.1	58.3	59.8	57.0	57.3	59.5	57.0
	ROID	59.9	59.9	59.6	51.2	46.4	43.9	48.3	42.0	46.9	44.9	50.9	47.7	46.7	49.2	42.9
IN-Sk.	Source	76.9	73.4	79.8	75.9	73.0	71.5	74.5	70.6	74.7	71.9	75.8	72.7	72.3	73.2	71.6
	BN–1	74.6	75.0	77.8	73.6	72.3	70.9	73.4	69.2	75.3	74.7	75.1	71.9	72.1	74.8	69.3
	ROID	73.3	71.1	71.5	64.0	61.2	59.2	62.1	57.3	61.9	60.6	66.0	62.2	61.4	62.3	56.6
IN-D109	Source	60.7	58.5	61.8	58.8	56.1	55.1	57.4	54.1	57.2	55.3	58.3	56.2	55.5	55.4	53.7
	BN–1	58.0	60.5	59.4	55.1	53.7	52.4	54.7	51.7	56.2	55.6	56.0	53.7	54.2	55.6	52.6
	ROID	56.5	57.4	54.6	47.9	46.1	44.0	46.3	43.6	46.5	45.0	48.5	46.5	46.1	47.3	43.2

Table 6. Online classification error rate (%) for the ImageNet benchmarks in the *continual* TTA setting. Mobile and transformer architectures and their variations are considered (**t**iny, **s**mall, **b**ase).

		MobileNet			RegNetX		RegNetY		Swin			Swin v2			ViT		MaxViT
		v2	v3-s	v3-l	400mf	800mf	400mf	800mf	t	s	b	t	s	b	b-16	b-32	t
	GFLOPs	0.30	0.06	0.22	0.40	0.80	0.40	0.80	4.5	8.7	15.4	5.9	11.5	20.3	16.9	-	5.6
	MParams	3.4	2.5	5.4	5.2	7.3	4.3	6.3	29	50	88	28	50	88	86	88	31
IN	Source	28.1	32.3	26.0	27.2	24.8	26.0	23.6	18.5	16.8	16.4	17.9	16.3	15.9	18.9	24.1	16.3
IN-C	Source	86.7	83.5	82.5	84.5	84.0	83.3	80.6	70.5	63.7	64.0	71.7	65.2	64.2	60.2	61.6	54.9
	BN–1	77.2	74.7	73.0	73.7	72.4	73.2	70.0	-	-	-	-	-	-	-	-	53.4
	ROID	66.0	67.7	64.2	63.8	61.6	63.9	59.6	52.9	48.8	46.8	54.8	47.8	47.5	44.9	52.0	40.0
IN-R	Source	69.0	70.7	65.4	66.4	65.9	67.0	64.5	58.7	55.3	54.3	60.0	55.9	54.8	56.0	58.2	50.6
	BN–1	67.8	71.7	66.5	65.9	64.4	67.1	63.9	-	-	-	-	-	-	-	-	49.0
	ROID	62.0	69.0	63.3	60.8	58.9	62.9	59.1	50.7	46.6	45.8	50.0	44.4	44.4	44.2	46.8	38.5
IN-Sk.	Source	80.9	81.6	76.4	78.7	78.1	79.6	77.7	72.8	69.0	68.5	74.0	69.4	69.3	70.6	72.2	65.1
	BN–1	81.4	86.9	82.0	80.4	79.2	81.8	79.5	-	-	-	-	-	-	-	-	67.0
	ROID	74.2	83.8	77.6	75.6	73.1	76.5	74.0	63.5	59.6	58.6	64.0	58.9	58.7	58.6	59.9	55.2
IN-D109	Source	62.5	63.5	59.5	60.0	59.4	60.1	58.6	54.3	51.8	51.4	55.2	51.8	51.5	53.6	55.9	49.4
	BN–1	60.6	66.0	61.8	60.8	59.6	62.1	59.9	-	-	-	-	-	-	-	-	48.8
	ROID	55.1	62.6	58.6	55.6	54.4	57.8	55.1	48.1	45.6	45.0	48.6	45.0	44.3	45.0	47.1	41.9

To demonstrate that our proposed method ROID is largely model-agnostic, we evaluate our method in the continual TTA setting on 31 different architectures. In Table 5, we report our results on regular architectures. In Table 6, mobile architectures and transformers are considered on the left and right, respectively. All results worse than the source performance are highlighted in red. While test-time normalization (BN–1) can decrease the error for corruptions (IN-C) on all considered architectures, this is not the case for natural shifts (IN-R, IN-Sketch, IN-D109). Especially for mobile architectures, Inception-v3, and RegNets, the error rate even increases. Since ROID applies test-time normalization, it works particularly well when a good estimation of the batch statistics is possible during test-time. ROID always outperforms BN–1, but due to the bad estimation of the batch statistics of MobileNet-v3 on ImageNet-Sketch, improvement upon the source performance is not possible. Nevertheless, in general, ROID can significantly outperform the source model, demonstrating its applicability to a wide range of different architectures. Among all networks, MaxViT-tiny, a hybrid (CNN + ViT) model, performs best on all ImageNet benchmarks. Regarding the considered CNN architectures, ResNeXt-101-32x8d and RegNetY-32gf show the best overall results.

B.2. Catastrophic Forgetting

In Figure 9, we investigate the occurrence of catastrophic forgetting [27] for CoTTA [53], EATA [33], and ROID on the long continual ImageNet-C sequence (50,000 samples per corruption). Following [33], we adapt the model on an alternating sequence of corrupted data and source data, i.e., [*Gaussian*, *Source*, *Shot*, *Source*, ...], using the complete ImageNet validation set (50,000 samples) as *Source*. Note that this procedure is different compared to how catastrophic forgetting is measured within the field of continual learning. However, in TTA, where the model is continually adapted to an unknown domain, this is the more realistic setting. Clearly, CoTTA suffers from major catastrophic forgetting, as the source error steadily increases after each corruption. By using elastic weight consolidation, EATA can largely prevent forgetting. However, to perform elastic weight consolidation, EATA requires data from the initial source domain, which may be unavailable in practice. Our proposed method ROID, which utilizes weight ensembling, is even more effective than EATA and only requires the initial parameters of the normalization layers. ROID is capable of nearly recovering the performance of the initial source model on the source domain.

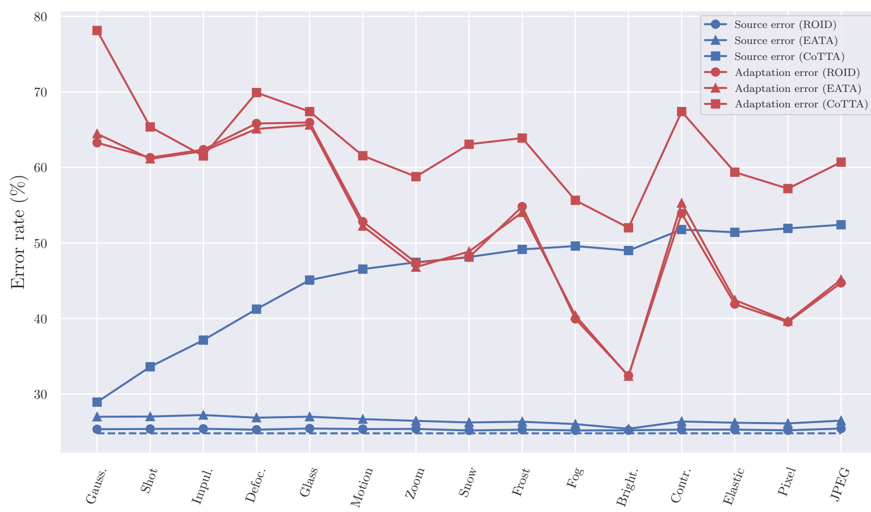

Figure 9. Source and adaptation error of ROID, EATA, and CoTTA for ImageNet-C (50,000 samples per domain) in the continual TTA setting with an alternating domain sequence. The dashed line indicates the lower bound (source error of the source model).

B.3. Momentum for Weight Ensembling

In Table 7, we analyze the sensitivity with respect to the momentum α used for our weight ensembling. ResNet-50, Swin-b, and ViT-b-16 are evaluated on the continual ImageNet-C benchmark. Choosing a relatively low momentum $\alpha = 0.9$, corresponding to only "keeping" 90% of the current model and adding 10% of the weights of the initial source model, limits adaptation. In the interval $\alpha \in [0.99, 0.9975]$, a decent compromise between allowing adaptation and remaining good generalization from the source model is possible. For large momentum values $\alpha \geq 0.999$ the advantages of weight ensembling vanish, resulting in an increase of adaptation error for all architectures.

B.4. Computational Efficiency

Since efficiency is also of great importance for a method performing its adaptation during test-time, we study in Table 8 the efficiency of each method with respect to the number of required forward and backward propagations, as well as the number of trainable parameters. We conduct the analysis on ImageNet-R using a ResNet-50. Clearly, the most inefficient methods are CoTTA, RMT, and AdaContrast which do not only require three and four times as many forward passes, but also calculate the gradients with respect to all parameters. While RoTTA also performs three forward passes per test sample,

Table 7. Online classification error rate (%) for ImageNet-C at the highest severity level 5 in the *continual* TTA setting. Different momentum values used for weight ensembling are considered for our approach. Note that we omitted prior correction and $\mathcal{L}_{\text{SCE}}$ for a clearer analysis.

Model \ α	0.99975	0.9995	0.999	0.9975	0.995	0.99	0.95	0.9
ResNet-50	60.1	58.9	58.4	57.0	56.3	**56.1**	60.0	63.0
Swin-b	53.4	52.4	51.9	50.8	50.1	**49.7**	53.4	57.0
ViT-b-16	47.9	47.8	47.4	46.7	**46.7**	47.0	51.8	55.3

significantly less parameters are trained and the number of backward propagations is not increased. The most efficient method during adaptation is EATA. Compared to the second most efficient method, TENT, fewer backward passes are required as some samples are filtered out. Due to performing consistency regularization, ROID is slightly less efficient than TENT and EATA, but comparable to SAR. Note that the additional 2000 forward and backward passes required to calculate the Fisher information matrix in EATA are not included in Table 8.

Table 8. Efficiency analysis for adapting a ResNet-50 on ImageNet-R.

Method	Error (%)	#Forwards	#Backwards	Train. Params (%)
Source	63.8	30,000	-	-
BN–1	60.3	30,000	-	-
LAME	99.4	30,000	-	-
TENT-cont.	57.4	30,000	30,000	0.21
EATA	54.2	30,000	5,440	0.21
SAR	57.2	46,279	30,111	0.12
CoTTA	57.4	90,000	30,000	100
RoTTA	60.8	90,000	30,000	0.21
AdaContrast	59.1	120,000	60,000	100
RMT	55.9	90,000	60,000	100
ROID (ours)	51.3	48,610	37,220	0.21

B.5. Memory Efficiency

Another huge advantage of architectures based on group or layer normalization is their potential to recover the batch TTA setting from a single sample scenario by leveraging gradient accumulation. This approach has the additional benefit that it significantly reduces the amount of required memory, which can be a scarce when TTA is performed on an edge device. In Table 9, the allocated memory for the batch and single sample setting is compared. Using gradient accumulation with TENT and ViT-b-16 reduces the maximum GPU memory consumption by 14.5 times while providing the same results. In case of ROID, the reduction factor is 15.8. If Swin-b is used as a model, the memory reduction factors are even larger.

Table 9. Memory efficiency analysis for TENT-cont. and ROID when adapting either Swin-b or ViT-b-16 on ImageNet-R.

Method	Architecture	Batch Size	Error (%)	Max. GPU mem. allocated
TENT-cont.	Swin-b	64	54.2	9.20 GB
TENT-cont.	Swin-b	1	54.3	0.50 GB
TENT-cont.	ViT-b-16	64	53.3	6.36 GB
TENT-cont.	ViT-b-16	1	53.3	0.44 GB
ROID (ours)	Swin-b	64	45.8	15.92 GB
ROID (ours)	Swin-b	1	45.8	0.71 GB
ROID (ours)	ViT-b-16	64	44.2	10.90 GB
ROID (ours)	ViT-b-16	1	44.1	0.69 GB

B.6. Component Analysis

In the following we elaborate and extend the component analysis from the main paper. Detailed results for the continual and mixed-domains setting are presented in Table 17 and for the correlated and mixed-domains correlated setting in Table 18. To adapt a model to the entire spectrum of Universal TTA, the most important aspect is to have a stable method. This factor isn't solely crucial for a specific scenario in TTA but resonates across all settings. As our analysis in Sec. 3 and Appendix A.2 suggests, even in the easiest setting (continual) it is essential to prevent the model from developing a bias or worse, collapsing to a trivial solution during test-time. A non-stationary setting, such as mixed-domains, can further enhance a model bias and degrade performance. To circumvent this, diversity weighting is essential. This is also supported by our component analysis which demonstrates that the driving factor in the continual and mixed-domains setting is diversity and certainty weighting.

To effectively address the challenge of dealing with multiple domain shifts over time, we employ weight ensembling (WE). WE retains generalization and still enables a good adaptation, as demonstrated in Sec. 3. It should be underscored that this is only necessary when a model adapts to a narrow distribution, potentially leading to overfitting on the current domain. In the context of mixed-domains, where samples from different domains are encountered within a single batch, adapting to such a broad distribution is also possible without WE. This is demonstrated by our component analysis, where WE improves the performance where multiple domain shifts are encountered, but actually slightly degrades the performance in the mixed-domains setting (broad distribution). Note that also for ImageNet-R and ImageNet-Sketch the best performance in the continual setting is achieved for configuration B, since here we only adapt to a single domain and do not encounter any additional domain shifts where generalization would be of importance. Nevertheless, the concept of WE carries the added benefit of enhancing overall stability. It serves as a corrective measure, capable of rectifying suboptimal adaptations over time, by continually incorporating a small percentage of the source weights. This becomes visible for the difficult adaptation in correlated settings, where highly imbalanced data can hinder a stable adaptation process. Here, WE ensembling ensures a stable adaptation process.

Shifting our focus to the correlated setting, the role of prior correction is substantial. Weighting the network's outputs with a smoothed estimate of the label prior benefits in settings with highly imbalanced data. Uncertain data points can be corrected by taking prior label information into account, while not degrading performance when a uniform label distribution is present.

Taking a look at employing consistency through data augmentation, the component analysis shows that it is beneficial across all settings and datasets. Compared to the other components, encouraging the invariance to small changes in the input space, has a moderate benefit.

C. Detailed Results

Table 10. Online classification error rate (%) for different settings using the ImageNet-D109 dataset. We report the performance of each method averaged over 5 runs. We do not report the results for ResNet-50 in the correlated setting, since BN–1 already achieves an error of 92.8%.

	Setting	continual						correlated						mixed domains					
	Time	t ⟶						t ⟶											
	Method	*clipart*	*infograph*	*painting*	*real*	*sketch*	Mean	*clipart*	*infograph*	*painting*	*real*	*sketch*	Mean	*clipart*	*infograph*	*painting*	*real*	*sketch*	Mean
ResNet-50	Source	64.2	81.0	51.5	24.2	73.2	58.8	64.2	81.0	51.5	24.2	73.2	58.8	64.2	81.0	51.5	24.2	73.2	58.8
	BN–1	55.7	80.1	50.1	25.1	64.8	55.1±0.05	92.4	93.2	91.9	92.7	94.0	92.8±0.05	58.3	78.6	51.7	26.0	66.6	56.2±0.03
	TENT-c.	53.5	78.1	47.9	24.8	60.3	52.9±0.05	-	-	-	-	-	-	57.8	81.7	50.1	25.3	65.7	56.1±0.15
	EATA	51.8	76.7	47.6	24.0	57.8	51.6±0.21	-	-	-	-	-	-	54.2	78.4	49.3	24.3	60.5	53.3±0.3
	SAR	53.9	77.6	47.3	24.4	58.1	52.2±0.07	-	-	-	-	-	-	55.2	77.5	49.1	24.8	61.9	53.7±0.05
	CoTTA	53.7	77.4	46.2	**23.1**	53.5	50.8±0.07	-	-	-	-	-	-	50.4	**75.5**	**44.7**	**23.0**	57.9	**50.3**±0.12
	RoTTA	55.3	77.7	47.6	23.5	57.5	52.3±0.04	-	-	-	-	-	-	56.9	77.7	47.5	23.6	64.3	54.0±0.18
	AdaCont.	49.7	78.0	46.2	23.8	54.4	50.4±0.17	-	-	-	-	-	-	56.2	83.2	49.0	24.6	64.0	55.4±0.12
	RMT	49.1	75.2	45.3	25.2	**52.2**	49.4±0.06	-	-	-	-	-	-	**49.9**	76.8	45.4	24.5	**56.8**	50.7±0.23
	LAME	99.1	99.4	97.8	29.6	99.2	85.0±0.12	-	-	-	-	-	-	99.0	99.6	98.7	98.8	99.2	99.1±0.02
	ROID	**45.9**	**74.2**	**44.6**	23.1	52.3	**48.0**±0.06	-	-	-	-	-	-	51.0	75.8	46.7	23.7	57.3	50.9±0.04
Swin-b	Source	53.6	73.6	44.0	20.3	65.3	51.4	53.6	73.6	44.0	20.3	65.3	51.4±0.0	53.6	73.6	44.0	20.3	65.3	51.4
	TENT-c.	53.5	80.0	59.2	42.2	95.4	66.1±0.69	53.8	80.1	60.5	49.7	98.3	68.5±0.29	66.9	83.9	55.4	24.8	76.4	61.5±0.42
	EATA	51.2	70.5	41.0	19.2	55.5	47.5±0.14	52.4	71.2	45.8	29.5	70.7	53.9±1.18	50.3	71.9	41.8	19.6	60.7	48.9±0.12
	SAR	52.2	78.5	52.0	20.4	67.7	54.2±0.62	61.3	77.7	49.3	20.2	68.9	55.5±0.25	56.7	78.3	46.4	21.2	67.3	54.0±0.14
	CoTTA	53.2	74.0	42.3	19.9	60.0	49.9±0.18	55.7	80.6	54.5	32.1	69.8	58.5±10.7	51.6	72.5	41.1	19.5	62.2	49.4±0.23
	RoTTA	52.7	72.3	41.0	19.5	57.8	48.7±0.03	53.2	73.0	42.5	20.1	63.6	50.5±0.07	49.4	70.9	40.6	19.6	60.2	48.1±0.10
	AdaCont.	48.2	73.9	40.2	**18.6**	55.8	47.3±0.08	53.2	77.5	43.8	19.9	66.4	52.1±0.11	49.8	77.3	40.4	**19.0**	60.8	49.4±0.15
	RMT	48.3	73.5	**39.4**	19.4	57.7	47.6±0.44	51.9	79.2	42.3	21.4	64.8	51.9±1.97	**46.7**	71.9	**38.1**	19.0	**56.6**	**46.5**±0.13
	LAME	98.7	99.6	96.5	37.3	99.6	86.3±0.24	**27.8**	62.3	**18.0**	**7.7**	**36.3**	**30.4**±0.27	97.3	98.7	96.8	95.8	98.1	97.3±0.07
	ROID	**46.1**	**67.7**	39.8	19.7	**52.2**	**45.1**±0.10	27.8	**53.9**	24.1	10.5	36.8	30.6±0.16	48.2	**69.9**	40.6	19.6	57.7	47.2±0.07
ViT-b-16	Source	57.5	75.9	45.1	22.0	67.5	53.6	57.5	75.9	45.1	22.0	67.5	53.6	57.5	75.9	45.1	22.0	67.5	53.6
	TENT-c.	58.1	86.5	82.0	94.5	99.2	84.0±0.09	59.0	86.4	82.3	94.7	99.2	84.3±0.03	82.1	91.0	74.0	48.0	88.5	76.7±0.22
	EATA	53.4	70.2	40.8	20.3	52.5	47.4±0.12	54.5	70.8	45.1	36.3	80.1	57.4±2.24	50.9	71.7	41.5	20.5	58.5	48.6±0.10
	SAR	57.5	83.2	50.9	21.2	74.1	57.4±0.64	64.4	81.2	53.0	21.4	73.7	58.7±0.17	67.0	83.0	54.4	26.0	76.7	61.4±0.20
	CoTTA	80.2	89.8	68.2	40.9	87.6	73.4±6.28	86.2	96.6	90.8	93.1	99.0	93.1±6.09	66.4	80.5	45.3	22.7	75.2	58.0±0.51
	RoTTA	56.7	74.4	42.8	20.8	61.2	51.2±0.03	57.4	75.5	44.6	22.4	69.0	53.8±0.04	53.2	73.1	42.1	21.0	62.9	50.5±0.06
	AdaCont.	51.5	76.8	41.4	**19.9**	59.0	49.7±0.11	59.4	81.1	47.2	21.8	74.1	56.7±0.12	53.1	79.6	41.8	**20.0**	62.5	51.4±0.12
	RMT	82.5	90.4	66.0	45.0	87.2	74.2±14.0	84.2	97.6	90.7	82.5	98.0	90.6±10.3	75.8	87.6	61.4	44.4	84.6	70.8±14.4
	LAME	99.0	99.6	96.3	45.7	99.2	88.0±0.18	31.0	75.2	**18.6**	**9.4**	43.0	35.4±0.32	98.8	99.5	98.4	98.3	99.0	98.8±0.03
	ROID	**46.2**	**68.2**	**39.9**	20.5	**50.2**	**45.0**±0.04	**30.2**	**55.7**	24.7	10.9	**36.9**	**31.7**±0.08	**48.6**	**69.7**	**40.6**	20.5	**55.2**	**46.9**±0.02

Table 11. Online classification error rate (%) for ImageNet-D109 for the *mixed domains correlated* TTA setting with Dirichlet concentration parameter $\delta = 0.1$. We report the performance of each method averaged over 5 runs.

	Method	*clipart*	*infograph*	*painting*	*real*	*sketch*	Mean
Swin-b	Source	53.6	73.6	44.0	20.3	65.3	51.4
	SAR	56.4	78.0	46.5	21.2	67.4	53.9±0.52
	LAME	26.6	**25.9**	30.8	28.2	**28.3**	**28.0**±0.39
	ROID	**25.5**	47.0	**24.0**	**12.4**	32.6	28.3±0.19
ViT-b-16	Source	57.5	75.9	45.1	22.0	67.5	53.6
	SAR	66.3	82.4	53.6	25.5	76.4	60.8±0.48
	LAME	**27.6**	**27.3**	32.2	29.6	**29.4**	**29.2**±0.55
	ROID	28.0	47.8	**24.7**	**12.8**	33.8	29.4±0.13

Table 12. Online classification error rate (%) for the corruption benchmarks at the highest severity level 5 for the *continual* TTA setting. For CIFAR10-C the results are evaluated on WideResNet-28, for CIFAR100-C on ResNeXt-29, and for Imagenet-C, ResNet-50, Swin-b and ViT-b-16 are used. We report the performance of each method averaged over 5 runs.

	Time	$t \longrightarrow$															
	Method	*Gaussian*	*shot*	*impulse*	*defocus*	*glass*	*motion*	*zoom*	*snow*	*frost*	*fog*	*bright.*	*contrast*	*elastic*	*pixelate*	*jpeg*	Mean
CIFAR10-C	Source	72.3	65.7	72.9	47.0	54.3	34.8	42.0	25.1	41.3	26.0	9.3	46.7	26.6	58.4	30.3	43.5
	BN–1	28.3	26.2	36.2	12.7	35.1	13.9	12.2	17.5	17.7	15.0	8.3	13.0	23.6	19.7	27.4	20.4±0.07
	TENT-cont.	25.0	20.3	29.0	13.8	31.7	16.2	14.1	18.6	17.6	17.4	10.8	15.6	24.3	19.7	25.1	20.0±1.19
	EATA	24.6	19.1	27.7	12.8	29.4	14.5	12.1	16.3	15.8	15.2	9.3	13.0	21.6	16.1	20.8	17.9±0.15
	SAR	28.3	26.2	35.6	12.7	34.7	13.9	12.2	17.5	17.7	15.0	8.3	13.0	23.6	19.7	27.4	20.4±0.06
	CoTTA	24.2	21.9	26.5	12.0	27.9	12.7	10.7	15.2	14.6	12.8	7.9	11.2	18.5	**14.0**	18.1	16.5±0.16
	RoTTA	30.3	25.4	34.6	18.3	34.0	14.7	11.0	16.4	14.6	14.0	8.0	12.4	20.3	16.8	19.4	19.3±0.07
	AdaContrast	29.2	22.5	29.9	13.9	32.8	14.2	11.8	16.6	15.0	14.3	8.0	10.0	21.7	17.7	19.9	18.5±0.04
	RMT	24.1	20.2	**25.7**	13.2	**25.5**	14.7	12.8	16.2	15.4	14.6	10.8	14.0	**18.0**	14.1	**16.6**	17.0±0.34
	LAME	86.0	83.9	88.4	83.6	88.7	64.4	82.0	28.4	71.7	37.1	9.4	74.1	41.3	79.7	46.3	64.3±0.18
	ROID (ours)	**23.7**	**18.7**	26.4	**11.5**	28.1	**12.4**	**10.1**	**14.7**	**14.3**	**12.0**	**7.5**	**9.3**	19.8	14.5	20.3	**16.2**±0.05
CIFAR100-C	Source	73.0	68.0	39.4	29.4	54.1	30.8	28.8	39.5	45.8	50.3	29.5	55.1	37.2	74.7	41.2	46.4
	BN–1	42.3	40.7	43.2	27.7	41.8	29.8	27.9	35.0	34.7	41.8	26.4	30.2	35.6	33.1	41.2	35.4±0.03
	TENT-cont.	37.3	35.6	41.6	37.9	51.3	48.1	48.9	59.8	65.3	73.6	74.2	85.7	89.1	91.1	93.7	62.2±2.17
	EATA	37.2	33.1	36.0	27.8	37.6	29.6	27.0	32.6	31.5	35.2	26.6	29.1	33.4	29.6	37.5	32.2±0.10
	SAR	40.4	34.8	37.1	26.0	37.1	28.0	25.6	31.9	30.8	35.9	25.3	28.1	32.0	29.2	37.3	32.0±0.10
	CoTTA	40.5	38.2	39.8	27.2	38.2	28.4	26.4	33.4	32.2	40.6	25.2	27.0	32.4	28.4	33.8	32.8±0.07
	RoTTA	49.1	44.9	45.5	30.2	42.7	29.5	26.1	32.2	30.7	37.5	24.7	29.1	32.6	30.4	36.7	34.8±0.15
	AdaContrast	42.5	36.9	38.5	27.7	40.4	29.3	27.4	32.8	30.7	38.0	26.1	28.4	34.1	33.4	36.1	33.5±0.08
	RMT	40.2	36.2	36.0	27.9	**33.9**	28.4	26.4	**28.7**	28.8	**31.1**	25.5	27.1	**28.0**	26.6	**29.0**	30.2±0.15
	LAME	98.9	99.0	98.2	98.1	98.8	98.1	98.0	98.2	98.8	98.9	98.0	98.9	98.1	99.0	98.4	98.5±0.05
	ROID (ours)	**36.5**	**31.9**	**33.2**	**24.9**	34.9	**26.8**	**24.3**	28.9	**28.5**	31.1	**22.8**	**24.2**	30.7	**26.5**	34.4	**29.3**±0.04
ImageNet-C (RN-50)	Source	97.8	97.1	98.2	81.7	89.8	85.2	78.0	83.5	77.0	75.9	41.3	94.5	82.5	79.3	68.5	82.0
	BN–1	84.9	84.0	84.8	84.9	84.5	73.3	61.1	65.8	68.2	51.9	35.0	83.0	56.3	51.2	60.0	68.6±0.06
	TENT-cont.	81.7	74.6	72.6	77.6	73.8	66.1	55.7	61.5	63.1	51.3	38.0	71.8	51.0	47.5	52.9	62.6±0.11
	EATA	76.3	66.5	65.0	73.1	69.1	62.1	53.5	58.9	59.3	48.1	35.9	62.8	47.5	43.9	47.5	58.0±0.18
	SAR	81.8	74.1	71.4	77.8	73.4	65.8	56.0	61.4	62.3	51.0	37.3	69.4	49.7	46.1	50.9	61.9±0.20
	CoTTA	84.5	82.0	80.4	81.8	79.5	69.2	58.8	60.8	61.1	48.5	36.5	67.5	47.8	**41.8**	45.9	63.1±0.45
	RoTTA	88.3	82.8	82.1	91.3	83.7	72.9	59.4	66.2	64.3	53.3	35.6	74.5	54.3	48.2	52.6	67.3±0.25
	AdaContrast	83.0	80.6	78.7	82.4	78.8	72.5	63.5	63.5	64.0	53.2	38.7	67.0	54.3	49.7	53.2	65.5±0.18
	RMT	79.9	76.3	73.1	75.7	72.9	64.7	56.8	56.4	58.3	49.0	40.6	**58.2**	47.8	43.7	**44.8**	59.9±0.21
	LAME	99.9	99.9	99.9	83.6	99.8	99.8	96.7	99.9	98.7	99.8	41.6	99.7	99.9	98.3	84.3	93.5±0.12
	ROID (ours)	**71.7**	**62.2**	**62.2**	**69.6**	**66.5**	**57.1**	**49.3**	**52.3**	**57.4**	**43.5**	**33.4**	59.1	**45.4**	41.8	46.2	**54.5**±0.10
ImageNet-C (Swin-b)	Source	71.1	70.0	75.4	72.8	81.6	63.8	68.2	57.9	50.8	40.7	28.6	60.5	72.1	86.6	59.3	64.0
	TENT-cont.	67.0	62.0	63.5	79.2	78.6	65.3	67.3	59.1	55.7	52.0	32.5	62.9	73.4	82.6	59.3	64.0±0.15
	EATA	63.1	55.5	54.7	67.4	64.0	54.1	54.5	52.4	46.8	44.4	26.1	47.0	55.0	61.0	46.2	52.8±0.14
	SAR	63.6	57.4	58.1	75.7	73.5	65.7	65.0	60.9	59.4	57.7	31.2	72.4	71.9	81.1	62.4	63.7±1.23
	CoTTA	63.8	58.4	58.3	76.2	73.9	65.1	69.3	62.1	52.4	50.5	35.3	51.8	61.2	60.6	50.0	59.3±1.23
	RoTTA	71.0	69.0	73.1	72.9	79.7	62.0	66.8	56.1	48.0	42.2	28.7	56.7	68.1	88.1	57.8	62.7±0.10
	AdaContrast	63.3	60.1	59.9	72.6	81.1	65.6	67.4	54.7	46.3	51.3	27.3	47.8	64.5	60.4	49.4	58.1±0.11
	RMT	60.4	52.6	52.5	74.8	68.3	58.0	61.8	52.0	48.2	42.9	33.4	49.6	50.8	**41.6**	42.9	52.6±1.00
	LAME	88.6	76.5	87.5	84.3	97.5	86.6	80.3	99.6	99.4	96.8	28.8	90.0	99.7	95.1	61.8	84.8±0.29
	ROID (ours)	**58.0**	**51.6**	**51.4**	**62.9**	**57.6**	**49.9**	**47.5**	**44.2**	**39.9**	**36.2**	**24.2**	**43.9**	**44.5**	50.4	**42.5**	**47.0**±0.26
ImageNet-C (ViT-b-16)	Source	65.8	67.3	65.3	68.8	74.4	64.3	66.6	56.8	45.2	48.6	29.2	81.8	57.1	60.8	50.2	60.2
	TENT-cont.	63.6	59.8	58.0	65.8	68.2	58.0	61.4	53.9	45.4	47.9	28.2	61.2	53.5	50.8	42.4	54.5±0.04
	EATA	61.5	55.3	53.7	60.2	58.7	52.6	54.8	51.1	43.5	42.8	28.9	49.1	48.8	46.3	39.7	49.8±0.14
	SAR	61.2	55.7	54.3	62.1	61.4	54.0	57.1	53.8	45.2	45.7	29.0	53.8	51.7	50.0	40.3	51.7±0.02
	CoTTA	72.7	79.6	75.7	82.5	80.3	75.7	75.9	79.7	68.9	74.4	70.5	96.7	74.2	74.6	74.1	77.0±13.3
	RoTTA	65.8	66.7	64.5	68.6	72.9	62.5	64.8	55.1	43.5	44.4	27.9	77.8	53.9	58.5	48.3	58.3±0.13
	AdaContrast	65.3	62.8	60.0	67.4	73.1	63.0	66.4	56.0	44.2	49.8	28.9	72.4	54.9	47.6	42.5	57.0±0.19
	RMT	75.8	74.8	69.8	78.1	73.5	66.0	69.8	80.5	71.3	73.5	68.8	80.6	73.0	68.2	69.7	72.9±12.0
	LAME	95.5	81.6	97.5	72.0	89.9	96.9	93.9	96.1	48.8	99.8	29.4	99.9	82.4	64.7	50.5	79.9±0.19
	ROID (ours)	**57.6**	**51.5**	**52.2**	**55.1**	**52.4**	**46.5**	**47.2**	**45.6**	**39.5**	**36.0**	**26.0**	**45.0**	**43.8**	**39.7**	**36.3**	**45.0**±0.09

Table 13. Online classification error rate (%) for the corruption benchmarks at the highest severity level 5 for the *mixed domains* TTA setting. For CIFAR10-C the results are evaluated on WideResNet-28, for CIFAR100-C on ResNeXt-29, and for Imagenet-C, ResNet-50, Swin-b and ViT-b-16 are used. We report the performance of each method averaged over 5 runs.

	Method	*Gaussian*	*shot*	*impulse*	*defocus*	*glass*	*motion*	*zoom*	*snow*	*frost*	*fog*	*bright.*	*contrast*	*elastic*	*pixelate*	*jpeg*	Mean
CIFAR10-C	Source	72.3	65.7	72.9	46.9	54.3	34.8	42.0	25.1	41.3	26.0	9.3	46.7	**26.6**	58.4	30.3	43.5
	BN–1	45.5	42.8	59.7	34.2	44.3	29.8	32.0	19.8	21.1	21.5	9.3	27.9	33.1	55.5	30.8	33.8±0.04
	TENT-cont.	73.5	70.1	81.4	31.6	60.3	29.6	28.5	30.8	35.3	25.7	13.6	44.2	32.6	70.2	34.9	44.1±3.82
	EATA	**36.4**	**33.5**	51.5	24.1	38.9	23.4	21.5	19.8	19.8	21.5	11.4	32.0	27.1	42.2	25.3	28.6±0.7
	SAR	45.5	42.7	59.6	34.1	44.3	29.7	31.9	19.8	21.1	21.5	9.3	27.8	33.0	55.4	30.8	33.8±0.04
	CoTTA	38.7	36.0	56.1	36.0	**36.8**	32.3	31.0	19.9	**17.6**	27.2	11.7	52.6	30.5	**35.8**	25.7	32.5±1.35
	RoTTA	60.0	55.5	70.0	23.8	44.1	**20.7**	21.3	20.2	22.7	**16.0**	9.4	22.7	27.0	58.6	29.2	33.4±0.15
	AdaContrast	36.7	34.3	**48.8**	**18.2**	39.1	21.1	**17.7**	**18.6**	18.3	16.8	**9.0**	**17.4**	27.7	44.8	**24.9**	**26.2**±0.11
	RMT	42.8	39.7	55.0	28.5	38.6	26.5	25.9	19.6	18.9	20.6	12.2	27.3	26.9	56.9	25.9	31.0±0.75
	LAME	87.8	86.5	88.0	79.5	83.0	72.4	76.8	67.5	78.1	68.7	49.8	78.1	69.3	75.3	66.9	75.2±0.12
	ROID (ours)	37.1	34.3	50.9	24.8	38.1	22.5	22.0	18.8	18.5	18.8	9.9	25.6	27.2	45.7	26.2	28.0±0.12
CIFAR100-C	Source	73.0	68.0	39.4	29.3	54.1	30.8	28.8	39.5	45.8	50.3	29.5	55.1	37.2	74.7	41.2	46.4
	BN–1	62.7	60.7	43.1	35.5	50.3	35.7	34.4	39.9	38.8	51.5	27.5	45.5	42.3	72.8	46.4	45.8±0.04
	TENT-cont.	95.6	95.2	89.2	72.8	82.9	74.4	72.3	78.0	79.7	84.7	71.0	88.5	77.8	96.8	78.7	82.5±1.45
	EATA	42.4	40.1	34.2	30.1	42.7	31.7	29.3	35.6	35.8	43.7	30.2	42.0	36.9	38.1	40.6	36.9±0.21
	SAR	75.8	72.7	41.1	29.2	45.2	31.1	28.9	36.7	37.7	43.9	29.3	41.8	37.1	89.2	42.4	45.5±0.24
	CoTTA	54.4	52.7	49.8	36.0	45.8	36.7	33.9	38.9	35.8	52.0	30.4	60.9	40.2	**38.0**	41.1	43.1±0.05
	RoTTA	65.0	62.3	39.3	33.4	50.0	34.2	32.6	36.6	36.5	45.0	**26.4**	41.6	40.6	89.5	48.5	45.4±0.14
	AdaContrast	54.5	51.5	37.6	30.7	45.4	32.1	30.3	36.9	36.5	45.3	28.0	42.7	38.2	75.4	41.7	41.8±0.05
	RMT	52.6	49.9	32.2	31.0	**40.5**	31.8	30.4	**33.4**	33.9	**40.6**	27.8	**36.9**	35.3	65.0	**38.1**	38.6±0.15
	LAME	98.5	98.5	98.2	98.2	98.4	98.3	98.2	98.3	98.3	98.5	98.2	98.4	98.4	98.8	98.4	98.4±0.04
	ROID (ours)	**40.5**	**38.0**	**32.0**	**28.1**	40.5	**29.7**	**27.6**	34.1	**33.8**	41.3	28.7	38.7	**34.3**	39.7	38.5	**35.0**±0.04
ImageNet-C (RN-50)	Source	97.8	97.1	98.2	81.7	89.8	85.2	77.9	83.5	77.1	75.9	41.3	94.5	82.5	79.3	68.6	82.0
	BN–1	92.8	91.1	92.5	87.8	90.2	87.2	82.2	82.2	82.0	79.8	48.0	92.5	83.5	75.6	70.4	82.5±0.06
	TENT-cont.	99.2	98.7	99.0	90.5	95.1	90.5	84.6	86.6	84.0	86.5	46.7	98.1	86.1	77.7	72.9	86.4±1.35
	EATA	90.1	88.1	90.1	**76.5**	**80.9**	**73.8**	**68.5**	71.4	69.5	**63.5**	42.1	93.2	**69.7**	**52.4**	**54.8**	72.3±1.57
	SAR	98.4	97.3	98.0	84.0	87.3	82.6	77.2	77.5	76.1	72.5	43.1	96.0	78.3	61.8	60.4	79.4±0.75
	CoTTA	89.1	86.6	88.5	80.9	87.2	81.1	75.8	73.3	75.2	70.5	41.6	**85.0**	78.1	65.6	61.6	76.0±0.17
	RoTTA	89.4	88.6	89.3	83.4	89.1	86.2	80.0	78.9	76.9	74.2	**37.4**	89.6	79.5	69.0	59.6	78.1±0.07
	AdaContrast	96.2	95.5	96.2	93.2	96.4	96.3	90.5	92.7	91.9	92.4	50.8	97.0	96.6	89.7	87.1	90.8±0.11
	RMT	87.0	84.6	86.6	79.9	86.5	80.8	74.3	**70.2**	74.0	69.9	45.7	86.4	78.1	64.8	61.6	75.4±0.19
	LAME	99.4	99.3	99.5	95.2	97.3	95.9	93.9	95.5	93.9	93.8	84.3	98.5	95.3	94.2	91.3	95.1±0.39
	ROID (ours)	**76.4**	**75.3**	**76.1**	77.9	81.7	75.1	69.9	70.9	**68.8**	64.3	42.5	85.4	69.8	53.0	55.6	**69.5**±0.13
ImageNet-C (Swin-b)	Source	71.1	70.0	75.4	72.8	81.6	63.8	68.2	57.9	50.8	**40.7**	28.6	60.5	72.1	86.6	59.3	64.0
	TENT-cont.	65.8	63.9	68.2	73.4	75.3	59.1	64.5	60.0	57.9	49.1	28.8	61.4	72.2	81.6	56.9	62.6±0.12
	EATA	61.7	60.4	61.4	**65.8**	68.7	**52.8**	58.1	54.1	50.8	46.1	27.2	**51.0**	63.4	72.0	51.5	56.3±0.18
	SAR	64.1	62.3	64.9	71.4	71.8	57.5	62.0	58.8	56.0	51.0	29.0	59.5	68.4	77.3	54.3	60.6±0.62
	CoTTA	54.5	54.9	55.9	77.9	79.8	67.1	70.9	62.8	59.1	53.7	37.3	60.4	70.3	87.5	57.7	63.3±7.69
	RoTTA	67.4	65.8	70.2	72.9	78.8	62.7	67.7	53.7	48.5	43.2	28.8	58.5	70.2	87.8	62.0	62.6±0.11
	AdaContrast	62.7	61.5	63.5	75.1	83.5	74.3	71.9	67.7	71.6	72.9	29.0	53.5	79.6	69.5	53.5	66.0±0.80
	RMT	**49.0**	**48.1**	**49.2**	67.9	72.4	58.5	62.7	56.4	52.0	54.7	33.7	51.3	62.1	**63.5**	**49.0**	55.4±4.54
	LAME	71.6	70.4	75.9	73.2	82.0	64.4	68.6	58.6	51.9	42.2	29.5	61.7	72.7	86.9	59.8	64.6±0.12
	ROID (ours)	61.1	59.6	60.8	66.4	**67.3**	53.4	**57.3**	**51.0**	**45.1**	43.1	**26.2**	52.6	**59.6**	71.1	50.9	**55.0**±0.26
ImageNet-C (ViT-b-16)	Source	65.8	67.3	65.3	68.8	74.4	64.3	66.6	56.8	45.2	48.6	29.2	81.8	57.1	60.8	50.2	60.2
	TENT-cont.	60.6	60.4	59.6	63.6	67.8	57.1	61.2	55.0	48.8	47.4	28.6	66.7	53.9	50.4	44.4	55.0±0.08
	EATA	59.2	57.7	57.8	59.0	63.1	52.6	**58.2**	51.1	46.5	44.2	28.6	58.6	50.9	**47.0**	41.9	51.8±0.14
	SAR	58.9	57.6	57.6	59.4	63.6	53.0	58.5	52.3	47.1	45.4	28.3	61.6	51.4	47.4	42.0	52.3±0.11
	CoTTA	89.4	92.0	88.9	93.6	92.6	90.6	86.5	94.9	88.2	86.6	75.8	96.5	85.7	93.5	84.6	89.3±6.18
	RoTTA	64.4	65.6	63.7	67.6	71.3	59.8	64.1	52.7	**43.5**	48.6	27.9	78.5	54.3	60.4	50.1	58.2±0.06
	AdaContrast	64.8	63.4	63.3	72.8	76.6	73.7	74.6	67.7	48.0	89.6	30.2	93.2	60.8	57.3	46.3	65.5±0.15
	RMT	76.6	76.1	76.5	78.1	78.0	72.6	72.4	80.4	67.8	71.2	55.0	94.6	69.3	66.5	65.2	73.4±13.44
	LAME	67.9	69.1	67.4	70.6	75.7	66.3	68.4	59.2	48.1	53.8	33.1	84.6	59.3	62.8	52.8	62.6±0.16
	ROID (ours)	**58.3**	**57.2**	**57.3**	**57.4**	**61.6**	**52.1**	58.3	**49.7**	44.1	**42.1**	**27.2**	**55.8**	**50.6**	47.0	**41.5**	**50.7**±0.08

Table 14. Online classification error rate (%) in the *correlated* TTA setting where samples are sorted by class. The corruption datasets are evaluated at the highest severity level 5. We report the performance of each method averaged over 5 runs.

Dataset	Architecture	Source	TENT	EATA	SAR	CoTTA	RoTTA	AdaCont.	RMT	LAME	ROID (ours)
CIFAR10-C	RN-26 GN	32.7	87.6	40.8	37.1	44.5	33.7	30.5	57.5	**11.3**	15.9±0.27
IN-C	Swin-b	64.0	86.7	74.2	59.3	99.5	75.5	77.6	99.6	47.0	**18.5**±0.10
	ViT-b-16	60.2	80.6	76.2	53.9	98.8	65.1	87.4	99.6	44.1	**16.8**±0.72
IN-R	Swin-b	54.2	53.6	53.9	53.1	58.9	54.1	56.9	48.1	**13.6**	25.2±0.37
	ViT-b-16	56.0	53.4	53.6	49.9	81.0	55.8	62.1	85.8	**13.0**	25.8±0.13
IN-Sketch	Swin-b	68.4	67.4	66.3	72.3	95.3	68.1	66.9	91.8	58.2	**43.9**±0.19
	ViT-b-16	70.6	66.7	63.7	74.6	95.5	70.1	72.3	97.9	61.0	**44.0**±0.14
IN-D109	Swin-b	51.4	68.5	53.9	55.5	58.5	50.5	52.1	51.9	**30.4**	30.6±0.16
	ViT-b-16	53.6	84.3	57.4	58.7	93.1	53.8	56.7	90.6	35.4	**31.7**±0.08

Table 15. Online classification error rate (%) for the corruption benchmarks at the highest severity level 5 for the *correlated* TTA setting. For CIFAR10-C the results are evaluated on ResNet-26 with group norm (RN-26 GN). For Imagenet-C, Swin-b, and ViT-b-16 are used. We report the performance of each method averaged over 5 runs.

	Time	$t\longrightarrow$															
	Method	*Gaussian*	*shot*	*impulse*	*defocus*	*glass*	*motion*	*zoom*	*snow*	*frost*	*fog*	*bright.*	*contrast*	*elastic*	*pixelate*	*jpeg*	Mean
CIFAR10-C (RN-26 GN)	Source	48.4	44.8	50.3	24.1	47.8	24.5	24.1	24.1	33.1	28.0	14.1	29.7	25.6	43.7	28.3	32.7
	TENT-cont.	62.3	82.6	89.9	89.4	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	87.6±0.98
	EATA	39.3	30.9	42.6	30.3	45.4	24.9	33.6	36.7	36.1	41.4	36.0	47.5	52.6	62.9	51.0	40.8±1.62
	SAR	48.6	55.6	63.8	23.4	55.8	29.1	24.9	30.6	33.1	26.1	14.5	27.8	31.6	58.5	33.0	37.1±1.12
	CoTTA	35.3	33.9	39.3	39.3	50.1	43.9	44.0	37.8	44.0	60.6	22.3	62.1	57.6	48.0	49.5	44.5±1.39
	RoTTA	49.2	47.2	55.0	21.9	50.5	23.1	20.0	27.9	36.6	29.4	15.2	27.9	26.4	43.8	31.0	33.7±0.19
	AdaContrast	42.5	33.0	46.3	23.1	50.0	24.2	23.0	27.2	29.8	23.1	19.0	22.7	26.9	41.4	25.5	30.5±0.14
	RMT	57.3	57.4	66.6	26.8	64.8	40.8	42.0	54.7	63.0	66.7	56.2	67.4	70.5	62.4	66.3	57.5±5.30
	LAME	**26.0**	23.8	**25.2**	**5.3**	**12.7**	**4.3**	**4.9**	**5.2**	**6.9**	**6.2**	**4.8**	11.5	**4.0**	**24.6**	**4.4**	**11.3**±0.21
	ROID (ours)	26.6	**13.9**	28.5	8.9	38.1	6.1	6.1	18.3	10.8	7.7	5.6	**9.2**	13.6	33.5	11.0	15.9±0.27
ImageNet-C (Swin-b)	Source	70.4	69.9	75.5	72.8	81.9	64.4	68.6	57.9	50.5	40.7	29.2	59.8	72.6	87.0	58.8	64.0
	TENT-cont.	61.4	57.5	59.9	76.6	76.3	80.1	93.0	97.6	99.7	99.9	98.9	99.8	99.8	99.8	99.7	86.7±0.90
	EATA	64.7	71.6	77.1	81.3	78.9	75.9	73.9	71.7	71.7	71.1	54.4	81.6	78.3	83.4	77.3	74.2±2.42
	SAR	62.3	58.7	59.7	80.9	79.5	60.5	65.5	66.8	59.8	52.5	27.4	46.5	69.2	53.0	47.5	59.3±0.57
	CoTTA	94.2	99.9	99.9	99.6	99.9	99.9	99.9	99.9	99.9	99.9	99.8	99.9	99.8	99.9	99.7	99.5±0.17
	RoTTA	69.4	66.6	70.8	82.5	79.8	76.4	76.8	62.8	58.9	76.2	47.9	95.2	77.3	98.3	94.2	75.5±0.29
	AdaContrast	61.8	61.2	66.2	78.9	84.1	81.7	82.1	75.5	70.7	82.4	62.0	85.8	89.0	92.5	90.0	77.6±0.14
	RMT	95.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.6±0.06
	LAME	43.0	42.0	46.2	52.8	69.7	43.4	51.1	44.5	36.4	33.9	20.4	41.3	64.4	72.7	44.0	47.0±0.10
	ROID (ours)	**25.8**	**22.9**	**22.7**	**33.2**	**31.3**	**18.8**	**21.0**	**14.0**	**12.0**	**11.2**	**6.9**	**15.5**	**13.6**	**14.7**	**14.6**	**18.5**±0.10
ImageNet-C (ViT-b-16)	Source	66.0	66.8	64.9	68.5	74.7	64.0	66.9	57.3	45.0	49.4	28.7	81.8	57.8	60.8	49.9	60.2
	TENT-cont.	58.7	53.9	54.3	58.4	58.6	52.7	73.6	99.4	99.8	99.9	99.9	99.9	99.9	99.9	99.9	80.6±0.05
	EATA	59.6	63.5	68.9	77.0	75.4	77.4	75.4	72.8	70.2	77.2	65.0	97.9	88.0	87.7	87.0	76.2±4.53
	SAR	55.8	51.7	55.0	57.5	56.9	50.3	58.3	64.6	55.0	48.7	41.0	55.3	59.1	50.2	48.9	53.9±11.5
	CoTTA	95.8	99.5	99.5	98.9	98.2	97.6	96.0	99.7	99.7	99.0	99.5	99.8	99.6	99.4	99.7	98.8±0.68
	RoTTA	66.3	67.0	69.9	70.5	70.2	58.9	64.8	60.4	55.0	56.0	34.3	79.6	61.2	87.5	74.3	65.1±0.15
	AdaContrast	66.2	70.4	78.7	81.7	87.3	88.5	91.7	89.8	90.6	94.4	87.3	94.5	96.6	97.1	96.9	87.4±0.10
	RMT	95.8	99.9	99.9	99.8	99.8	99.8	99.8	99.8	99.8	99.8	99.8	99.9	99.8	99.8	99.8	99.6±0.14
	LAME	40.1	39.1	39.3	48.6	58.6	43.1	48.4	39.5	34.1	42.3	23.7	84.8	44.5	40.1	35.8	44.1±0.02
	ROID (ours)	**25.7**	**23.0**	**23.8**	**29.0**	**21.9**	**19.0**	**18.1**	**14.8**	**12.2**	**10.2**	**6.4**	**14.9**	**12.3**	**9.9**	**10.1**	**16.8**±0.72

Table 16. Online classification error rate (%) for ImageNet-C at the highest severity level 5 for the *mixed domains correlated* TTA setting with the Dirichlet concentration parameter $\delta = 0.01$. We report the performance of each method averaged over 5 runs.

	Method	*Gaussian*	*shot*	*impulse*	*defocus*	*glass*	*motion*	*zoom*	*snow*	*frost*	*fog*	*bright.*	*contrast*	*elastic*	*pixelate*	*jpeg*	Mean
Swin-b	Source	70.4	69.9	75.5	72.8	81.9	64.4	68.6	57.9	50.5	40.7	29.2	59.8	72.6	87.0	58.8	64.0
	SAR	70.8	68.8	71.2	73.4	76.2	61.4	66.0	65.3	61.1	56.7	32.8	62.0	73.2	79.6	55.7	64.9±0.81
	LAME	37.4	37.4	37.4	37.6	37.8	37.4	37.7	37.2	37.0	37.3	36.5	37.5	37.6	37.8	37.2	37.4±0.12
	ROID (ours)	**30.9**	**30.1**	**31.0**	**34.8**	**37.2**	**26.4**	**31.1**	**26.8**	**22.8**	**23.2**	**12.3**	**27.0**	**33.6**	**36.4**	**25.2**	**28.6**±0.16
ViT-b-16	Source	66.0	66.8	64.9	68.5	74.7	64.0	66.9	57.3	45.0	49.4	28.7	81.8	57.8	60.8	49.9	60.2
	SAR	64.5	62.9	63.2	59.0	64.1	52.8	60.6	54.8	49.6	47.4	30.6	59.1	53.5	49.0	42.9	54.3±0.59
	LAME	36.2	36.1	36.1	36.3	36.3	36.1	36.4	36.1	35.8	36.1	35.4	36.6	35.9	36.0	35.7	36.1±0.15
	ROID (ours)	**27.0**	**26.2**	**26.1**	**26.1**	**32.0**	**23.4**	**29.2**	**23.4**	**19.4**	**18.6**	**10.8**	**26.6**	**25.5**	**21.5**	**17.7**	**23.6**±0.05

Table 17. Average online classification error rate (%) over 5 runs for different configurations for a) the *continual* TTA setting and b) the *mixed domains* TTA setting. For the ImageNet variants, a ResNet-50 is used. For CIFAR10-C and CIFAR100-C, the results are evaluated utilizing a WideResNet-28 and a ResNeXt-29, respectively.

	a) continual							b) mixed domains				
Method	*CIFAR10-C*	*CIFAR100-C*	*ImageNet-C*	*ImageNet-R*	*ImageNet-Sk.*	*ImageNet-D109*	Mean	*CIFAR10-C*	*CIFAR100-C*	*ImageNet-C*	*ImageNet-D109*	Mean
Source	43.5	46.4	82.0	63.8	75.9	58.8	61.7	43.5	46.4	82.0	58.8	57.7
TENT	20.0	62.2	62.6	57.6	69.5	52.9	54.1	44.1	82.5	86.4	56.1	67.3
SLR	20.1	57.7	61.5	55.6	67.8	52.7	52.6	42.8	78.2	87.4	58.2	66.7
+ Loss weighting	17.7	31.1	60.8	**51.1**	**64.1**	52.0	46.1	**26.9**	35.2	72.1	51.6	46.4
+ Weight ensembling	17.7	29.5	56.2	52.3	65.5	48.9	45.0	29.1	35.4	71.4	51.5	46.9
+ Consistency	16.3	**29.3**	**54.4**	51.2	64.2	48.1	**43.9**	28.4	35.1	69.6	51.0	46.0
+ Prior correction	**16.2**	**29.3**	54.5	51.2	64.3	**48.0**	**43.9**	28.0	**35.0**	**69.5**	**50.9**	**45.9**

Table 18. Average online classification error rate (%) over 5 runs for different configurations for a) the *correlated* TTA setting and b) the *mixed domains correlated* TTA setting. For the ImageNet variants, a ViT-b-16 is used, while for CIFAR10-C a ResNet26-GN is applied.

	a) correlated						b) mixed + correlated			
Method	*CIFAR10-C*	*ImageNet-C*	*ImageNet-R*	*ImageNet-Sk.*	*ImageNet-D109*	Mean	*CIFAR10-C*	*ImageNet-C*	*ImageNet-D109*	Mean
Source	32.7	60.2	56.0	70.6	53.6	54.6	32.7	60.2	53.6	48.8
TENT	87.6	80.6	53.4	66.7	84.3	74.5	88.2	81.3	77.3	82.3
SLR	89.0	90.3	52.3	78.0	90.4	80.0	88.3	88.7	87.4	88.1
+ Loss weighting	29.1	91.6	50.4	63.1	67.6	60.4	41.9	88.9	52.6	61.1
+ Weight ensembling	28.1	44.7	49.8	61.7	49.2	46.7	31.0	53.9	49.8	44.9
+ Consistency	29.5	42.5	48.0	60.5	48.1	45.7	31.0	51.5	48.8	43.8
+ Prior correction	**15.9**	**16.8**	**25.8**	**44.0**	**31.7**	**26.8**	**17.4**	**23.6**	**29.4**	**23.5**

D. Comparison to Related Work

Comparison with CoTTA While both CoTTA and our proposed method utilize source weights, CoTTA uses stochastic restoring, where with a small probability current weights are restored with the corresponding weights from the source model. The idea behind stochastic restoring is that the network avoids drifting too far away from the initial source model. But, as discussed in Section B.2, CoTTA first of all cannot prevent catastrophic forgetting on the continual ImageNet-C benchmark with 50,000 samples per corruption and, second, shows instabilities for certain domain shifts or settings. Instead of performing a stochastic restore, our proposed weight ensembling, which continually ensembles the weights of the initial source model and the weights of the current model, prevents catastrophic forgetting and mostly preserves the generalization capabilities of the initial source model.

Comparison with EATA EATA, like our proposed method, utilizes certainty and diversity weighting. However, their weighting scheme relies on dataset-specific hyperparameters, such as an entropy threshold and a cosine similarity threshold. While the entropy threshold is determined heuristically, the cosine similarity threshold needs to be manually specified for each dataset. Choosing an inappropriate cosine similarity threshold can lead to a significant decrease in performance. For example, switching the cosine similarity threshold of CIFAR10-C and CIFAR100-C reduces the performance by absolutely 2.7% and 10.8%, respectively. In contrast, our proposed diversity weighting scheme does not necessitate dataset-specific hyperparameters and has demonstrated success across a wide range of different datasets, models, and domain shifts, as validated by our experiments. To address catastrophic forgetting, EATA incorporates elastic weight consolidation, which requires access to source samples for computing the Fisher information matrix. As discussed in Appendix B.2, our proposed weight ensembling approach also effectively mitigates catastrophic forgetting without the need for source data availability. Furthermore, EATA does not only exhibit instabilities when dealing with correlated data, but also demonstrates impractical performance outcomes in this setting due to not employing any prior correction.