Table 6. Classification error rate (%) of various methods when the domain difficulty with respect to the initial source error at severity level 5 either increases from low-to-high (*easy-to-hard*) or decreases from high-to-low (*hard-to-easy*).

| Method | Source-free | Source error | *CIFAR10C* | *CIFAR100C* | *ImageNet-C* |
|---|---|---|---|---|---|
| BN–1 | ✓ | - | 20.4 | 35.4 | 68.6 |
| TENT | ✓ | high → low | 19.6 | 66.9 | 62.8 |
| TENT | ✓ | low → high | 20.2 | 52.1 | 60.2 |
| AdaContrast | ✓ | high → low | 18.8 | 33.7 | 66.3 |
| AdaContrast | ✓ | low → high | 17.9 | 32.6 | 60.2 |
| GTTA-MIX | ✗ | high → low | 17.6 | 30.5 | 60.2 |
| GTTA-MIX | ✗ | low → high | 17.4 | 30.1 | 58.8 |
| MT + $\mathcal{L}_{\text{CE}}$ | ✓ | high → low | 19.2 | 32.4 | 66.4 |
| MT + $\mathcal{L}_{\text{CE}}$ | ✓ | low → high | 17.1 | 30.5 | 61.2 |
| MT + $\mathcal{L}_{\text{SCE}}$ | ✓ | high → low | 18.9 | 31.9 | 63.5 |
| MT + $\mathcal{L}_{\text{SCE}}$ | ✓ | low → high | 16.7 | 29.4 | 51.5 |
| RMT (ours) | ✗ | high → low | 14.2 | 27.7 | 57.5 |
| RMT (ours) | ✗ | low → high | 12.9 | 26.3 | 50.3 |

## A. Adaptation with increasing difficulty

Since mean teacher based approaches have shown tremendous performance improvements in the gradual setting compared to the continual setting, we now consider the case, where the domain difficulty changes from easy-to-hard and hard-to-easy. Specifically, we sort the order of the corruptions with respect to the error at severity level 5 of the initial source model from low-to-high and high-to-low. While Tab. 6 shows the results, Tab. 7 illustrates the specific corruption orders for an increasing source error. Clearly, all methods improve when the domain difficulty increases compared to the other way round. Notably, mean teachers using a symmetric cross-entropy loss (MT + $\mathcal{L}_{\text{SCE}}$) demonstrate the highest error reductions with up to 12% on the ImageNet-C sequence. In contrast, a mean teacher with a cross-entropy loss (MT + $\mathcal{L}_{\text{CE}}$) only decreases the error by 5.2%, achieving an error rate of 61.2%. This is absolutely 9.7% worse compared to the error rate of a mean teacher using a symmetric cross-entropy loss.

## B. Ablation studies

For the following ablation studies, we investigate our non-source-free variant with 1 update step.

**More updates decrease the error for non-source-free methods** Since for some applications computational efficiency may be more important than a high accuracy and vice versa, we now investigate the effect of different numbers of update steps. As shown by Tab. 9 (a), all datasets profit when more update steps are applied, with 2 and 4 steps providing a good balance between performance and computational complexity. However, the best results are achieved with 6 updates. Note that the performance of source-free approaches like CoTTA deteriorates for multiple update steps due to over-adaptation.

**1% of the source data improves the performance** In Tab. 9 (b), we illustrate the error rate for different amounts of randomly sampled source data. This is especially relevant for applications with either a limited memory or when source data cannot be stored on the device due to privacy issues (0%). While the error increases slightly on CIFAR100C and DomainNet-126, the other datasets are only marginally affected by the amount of available source data. Although even our source-free variant already achieves state-of-the-art performance on all benchmarks, storing only 1% of the source data is enough to further boost the performance.

**Sensitivity** To investigate the sensitivity of our proposed method with respect to the momentum value $\alpha$ of the mean teacher, as well as the temperature term $\tau$, we conduct two ablation studies. As illustrated in Tab. 9 (c), which shows different values for the temperature term, RMT performs not only stable for all the common default values in contrastive learning (0.07, 0.1 and 0.2), but also for much larger values like 1.0. As shown by Tab. 9 (c), updating the mean teacher too slow or too fast can slightly degrade the results on average. Nevertheless, for the most common range of momentum values, the performance is stable.

In Tab. 8, we analyze the influence of source replay and the contrastive loss by considering different values for $\lambda_{\text{CE}}$ and $\lambda_{\text{CL}}$, respectively. For values close to the ones used by our approach $\lambda \in [0.5, 1.0]$, we observe a stable performance. As expected, for small values $\lambda \leq 0.1$, we observe a drop in performance, underlining that source replay and contrastive learning is indeed beneficial.

Table 8. Classification error rate (%) when using different loss weights. The results are averaged over 3 runs and all datasets.

| $\lambda$ | 0.0 | 0.1 | 0.5 | 1.0 |
|---|---|---|---|---|
| $\lambda_{\text{CE}}$ (source replay) | 38.7 | 38.1 | 37.6 | 37.8 |
| $\lambda_{\text{CL}}$ (contrastive learning) | 38.6 | 38.4 | 38.0 | 37.8 |

Table 7. The corruption types are ordered with respect to the error at severity level 5 of the initial source model from low-to-high.

| | low ⟶ high | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CIFAR10C | brightness | snow | fog | elastic | jpeg | motion | frost | zoom | contrast | defocus | glass | pixelate | shot | Gaussian | impulse |
| CIFAR100C | zoom | defocus | brightness | motion | elastic | impulse | snow | jpeg | frost | fog | glass | contrast | shot | Gaussian | pixelate |
| ImageNet-C | brightness | jpeg | fog | frost | zoom | pixelate | defocus | elastic | snow | motion | glass | contrast | shot | Gaussian | impulse |

Table 9. Classification error rate (%) for different: (a) numbers of update steps; (b) amounts of available source samples during test-time; (c) temperatures $\tau$ for the contrastive loss; (d) momentum values $\alpha$ used to update the mean teacher.

(a)

| Updates | 1 | 2 | 4 | 6 | 8 |
|---|---|---|---|---|---|
| CIFAR10C | 13.9 | 13.2 | 12.5 | 12.0 | **11.8** |
| CIFAR100C | 27.6 | 26.9 | **26.8** | **26.8** | 27.0 |
| ImageNet-C | 57.9 | 56.9 | 56.4 | **56.1** | 56.4 |
| ImageNet-R | 55.5 | 54.6 | 53.5 | **53.1** | **53.1** |
| DomainNet-126 | 33.9 | 33.1 | 32.8 | **32.7** | **32.7** |

(b)

| | 100% | 50% | 25% | 10% | 5% | 1% | 0% |
|---|---|---|---|---|---|---|---|
| CIFAR10C | **13.9** | **13.9** | 14.0 | 14.1 | 14.3 | 14.3 | 14.5 |
| CIFAR100C | **27.6** | 27.8 | 27.8 | 28.2 | 28.3 | 28.9 | 29.0 |
| ImageNet-C | 57.9 | **57.8** | 57.9 | **57.8** | 58.0 | 58.4 | 59.8 |
| ImageNet-R | 55.5 | 55.4 | **55.2** | 55.4 | 55.7 | 55.6 | 55.7 |
| DomainNet-126 | **33.9** | 34.2 | 34.3 | 34.4 | 34.6 | 34.6 | 34.7 |

(c)

| temperature $\tau$ | 0.01 | 0.07 | 0.1 | 0.2 | 1.0 |
|---|---|---|---|---|---|
| CIFAR10C | **13.9** | **13.9** | **13.9** | **13.9** | 14.3 |
| CIFAR100C | **27.5** | **27.5** | 27.6 | 27.7 | 28.1 |
| ImageNet-C | **57.7** | 57.8 | 57.9 | **57.7** | 58.1 |
| ImageNet-R | 55.6 | **55.3** | 55.5 | 55.5 | 55.8 |
| DomainNet-126 | 91.8 | 34.1 | **33.9** | 34.4 | 35.6 |

(d)

| momentum $\alpha$ | 0.99 | 0.995 | 0.999 | 0.9995 | 0.9999 |
|---|---|---|---|---|---|
| CIFAR10C | **13.8** | **13.8** | 13.9 | 14.3 | 15.5 |
| CIFAR100C | 28.6 | 27.7 | **27.6** | 28.2 | 29.2 |
| ImageNet-C | 64.1 | 60.9 | **57.9** | 58.3 | 60.9 |
| ImageNet-R | 60.1 | 58.9 | **55.5** | 56.1 | 57.2 |
| DomainNet-126 | 34.6 | 33.9 | **33.9** | 34.5 | 35.0 |

## C. DomainNet-126

While a variety of corruption benchmarks for test-time adaptation exist, benchmarks that investigate natural shifts are limited. A common dataset which contains natural shifts is ImageNet-R. However, it has the drawback that the included shifts are not separated. The continual DomainNet-126 benchmark closes this gap, consisting of four domains (real, clipart, painting, sketch) and 126 classes. It enables the investigation of continual TTA for natural shifts and further provides source models for all covered domains. Additionally, DomainNet-126 also includes shifts in label priors (imbalanced data), which corresponds to a more realistic setting than the uniform class distributions prevailing in existing benchmarks.

The exact four sequences used in the continual DomainNet-126 benchmark are shown in Tab. 10. While the left column indicates the name of the sequence and the domain on which the model was pre-trained, the order of the test domains is shown on the right.

Detailed results for the continual DomainNet-126 benchmark are shown in Tab. 11.

Table 10. Details on the test sequences used for the continual DomainNet-126 benchmark.

| Source domain | Test sequence | | | | |
|---|---|---|---|---|---|
| real | clipart | → | painting | → | sketch |
| clipart | sketch | → | real | → | painting |
| painting | real | → | sketch | → | clipart |
| sketch | painting | → | clipart | → | real |

Table 11. Classification error rate (%) for DomainNet-126 in the online continual test-time adaptation setting, where the test domains are sequentially displayed from left to right. We report the performance of our method averaged over 5 runs.

| Source domain | | | real | | | clipart | | | painting | | | sketch | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Source-free | Updates | clipart | painting | sketch | sketch | real | painting | real | sketch | clipart | painting | clipart | real | Mean |
| Source | ✓ | - | 44.9 | 37.4 | 53.6 | 52.5 | 40.1 | 55.3 | 24.7 | 53.8 | 46.7 | 49.2 | 44.8 | 40.4 | 45.3 |
| BN–1 | ✓ | - | 46.0 | 37.2 | 52.1 | 50.7 | 35.1 | 49.7 | 25.1 | 47.7 | 45.8 | 40.9 | 40.6 | 32.0 | 41.9 |
| TENT cont. | ✓ | 1 | 44.6 | 35.0 | 47.5 | 49.3 | 34.9 | 48.3 | 24.2 | 44.5 | 43.0 | 40.1 | 39.5 | 33.0 | 40.3 |
| CoTTA | ✓ | 1 | 45.3 | 35.8 | 49.2 | 50.1 | 33.4 | 45.6 | 23.4 | 43.9 | 41.8 | 40.0 | 39.2 | 29.6 | 39.8 |
| AdaContrast | ✓ | 1 | 39.3 | 32.7 | 41.4 | 45.5 | **27.5** | 39.7 | 20.5 | 39.2 | 37.2 | 35.9 | 34.7 | **25.1** | 34.9 |
| GTTA-MIX | ✗ | 4 | 40.4 | 32.7 | 42.8 | 46.4 | 34.2 | 46.7 | 23.3 | 40.2 | 37.3 | 36.9 | 36.0 | 29.8 | 37.7 |
| RMT (ours) | ✓ | 1 | 37.7 | 31.7 | 41.5 | 43.8 | 29.8 | 40.1 | 20.9 | 38.4 | 35.8 | 35.3 | 33.4 | 27.5 | 34.7 |
| RMT (ours) | ✗ | 1 | 37.9 | 31.4 | 41.1 | 43.5 | 29.1 | 38.6 | 21.1 | 37.1 | 33.8 | 35.1 | 31.7 | 26.4 | 33.9 |
| RMT (ours) | ✗ | 4 | **36.9** | **30.0** | **38.4** | **41.7** | 29.6 | **38.0** | **19.9** | **36.9** | **32.9** | **33.2** | **29.7** | 26.8 | **32.8** |

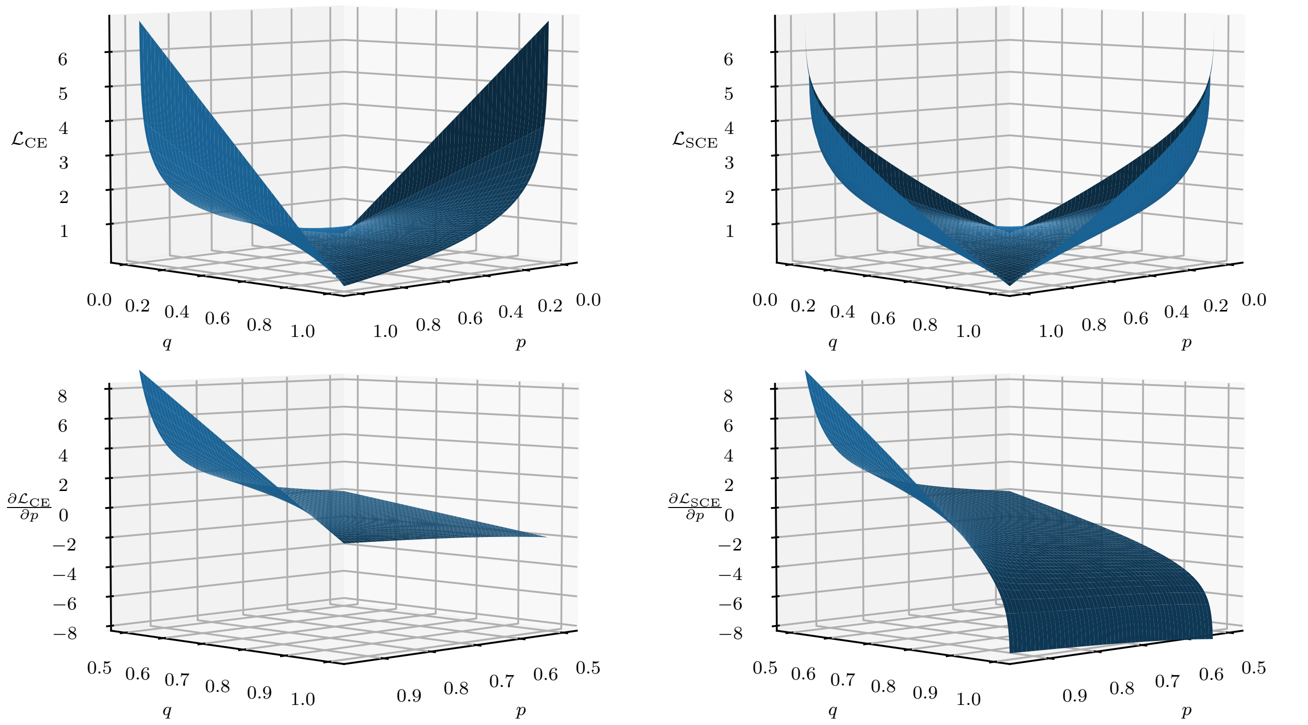

Figure 3. Loss (top) and gradient (bottom) surface illustrated for the binary case of the cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$ and the symmetric cross-entropy $\mathcal{L}_{\mathrm{SCE}}$ in dependence of the confidences $p$ and $q$ of the student and teacher, respectively.