Supplement Material

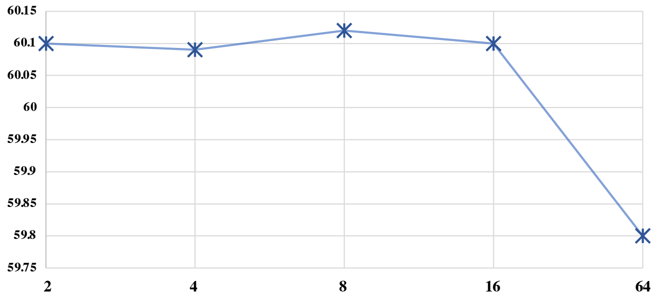

Figure 1. Averaged image classification accuracy of SoTa-DiT with different batch sizes on ImageNet-C.

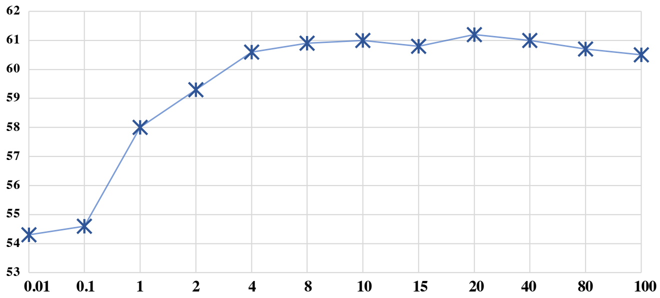

Figure 2. Averaged image classification accuracy of SoTa-DiT with different prompt extra learning rates μ on ImageNet-C.

1. More Hyper-parameters Ablation

In this section, we examine the effect of several other key hyper-parameters in SoTa-DiT.

1.1. Batch Size

We investigate the influence of batch size on accuracy. The results are depicted in Fig.1. We draw the following observation:

As batch size increases, accuracy initially remains stable and decreases. We infer that the reason behind this phenomenon is that the large batch size influences TP's target contrastive loss. When the batch size is too large, contrastive learning for TP becomes more challenging as the total number of test samples is limited, with minor inter-batch diversity. Consequently, the target contrastive loss cannot effectively supervise TP, leading to a decline in performance.

1.2. Prompt Extra Learning Rate μ

We investigate the influence of prompt extra learning rate μ. The results are depicted in Fig.2. We draw the following observation.

As the extra learning rate μ increases, the accuracy increases sharply and decreases slightly. The reason behind this is twofold. On one hand, when the μ is lower than 1, TP and SP learn slower than other parts of the network. As a result, they fail to extract the necessary source and target knowledge, resulting in a performance decline. On the other hand, as μ becomes too large, the learning rate of TP and SP is large. As a result, they learn too fast and may miss the actual optimum points due to the fast learning speed.

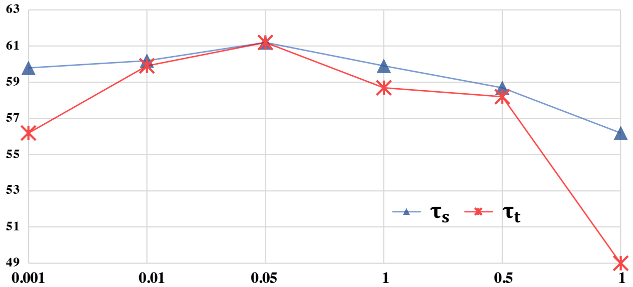

Figure 3. Averaged classification accuracy of SoTa-DiT with different contrastive temperatures τ_s and τ_t on ImageNet-C.

1.3. Contrastive Temperature τ

We investigate the influence of contrastive temperature τ. The results are depicted in Fig.3. The average accuracy under different contrastive temperatures for source contrastive loss (τ_s)and target contrastive loss (τ_t) are depicted with blue and red lines, respectively. Based on the result, we draw the following observations.

First, as τ_s and τ_t increases, the accuracy initially increases then decreases. The reason is twofold. When τ is small, the contrastive loss converges to 0 easily. For instance, if $\tau \to 0$, the positive pair is only slightly more similar than other samples, the contrastive loss is close to 0 without proper training. Consequently, the effectiveness of source and target contrastive loss diminishes. When τ

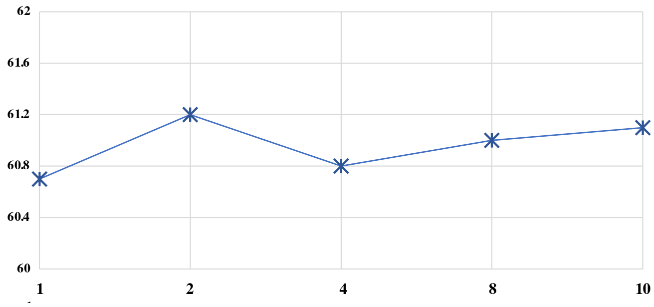

Figure 4. Averaged classification accuracy of SoTa-DiT with different augmentation group numbers M on ImageNet-C.

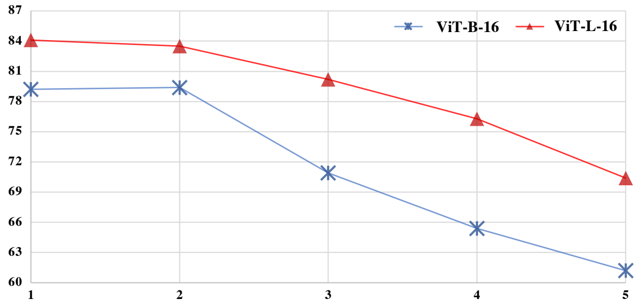

Figure 5. Averaged classification accuracy of SoTa-DiT with different corruption severity levels on ImageNet-C.

surpasses 0.5, a larger τ makes it challenging to converge. As a result, TP and SP are prone to overfitting. Specifically, when τ is set to 1, if the positive pair similarity is 1 and the negative pair similarity is -1, the contrastive loss still deviates significantly from 0.

Second, the effect of τ_t is stronger than τ_s. We infer the reason is that the SP embedding is more stable with the source similarity loss, even when it is not properly supervised with the contrastive loss.

1.4. Augmentation Group Number M

We investigate the influence of the augmentation groups M for the target contrastive learning. The results are depicted in Fig.4. We draw the following observations.

First, when M is 1, we observe a higher error rate. We infer that when M equals 1, the data is insufficient for contrastive learning. Consequently, TP cannot absorb the target knowledge effectively.

Second, the error rate remains relatively stable as M exceeds 2. We infer that 2 groups of augmented data are enough for contrastive learning. More augmented groups do not enhance the efficiency of target knowledge extraction; instead, they introduce unnecessary computational load. As a result, we set M as 2 by default.

2. Corruption Severity

We test SoTa-DiT on ImageNet-C using five different corruption severities. Notice that all other experiments on ImageNet-C are conducted on severity level 5, the strongest corruption level. The average accuracies of SoTa-DiT with two types of baseline, namely ViT-B-16 and ViT-L-16, are depicted in Fig.5 with blue and red lines, respectively.

The result shows that, as the corruption severity increases, the average classification accuracy of SoTa-DiT with two backbones decreases. The reason is that as the corruption severity increases, the test images become more dissimilar from the original images. As a result, the domain gap gets wider, making classification more challenging for the model.

3. More Key Component Ablation

We examine all the components in SoTa-DiT, as shown in Tab.1. In the ‘SoTa-DiT*’ group, we let $\mathcal{L}_{SAL}^{S}$ tune SP. Based on the results, we draw the following observations:

First, all the components in SoTa-DiT are effective in increasing classification accuracy. As shown in the table, adding the source prompt alone with source contrastive loss, source similarity loss, and source adaptation loss increases the accuracy by $+0.3\%$, $+0.2\%$, and $+0.4\%$. Adding the target prompt alone with target contrastive loss and target guiding loss increases the accuracy by $+4.0\%$ and $+8.0\%$, respectively.

Moreover, extracting the source and target knowledge with two prompts is more effective than using a single prompt. For instance, comparing ‘TP only+$\mathcal{L}_{SCL}^{S}$’ and ‘TP only+ $\mathcal{L}_{SSL}^{S}$’, we see that the accuracy increases by $+2.0\%$ and $+2.1\%$. This proves that disentangling the source and target knowledge with two prompts benefits the CoTTA task.

4. More Knowledge Disentangle Observation

In this section, we conduct additional experiments to further substantiate our knowledge-disentangling claims. First, we demonstrate that the source knowledge is well-preserved within the source prompt. Next, we demonstrate that the target knowledge is efficiently extracted by the target prompt. All the experiments are conducted on the ImageNet-C dataset with the ViT-B-16 backbone.

4.1. Source Knowledge Preservation

We evaluate the model directly on the original ImageNet test set at different time steps to demonstrate that the source knowledge is preserved within the source prompt and adapted to other parts of the model. The classification accuracies at different time steps are depicted in Fig.6. Specifically, the blue, red, and green lines represent the prediction accuracy of 1) SP output, 2) TP output, and 3) TP output from a model without SP (denoted as TP$*$), respec-

Table 1. Evaluation of key components on ImageNet-C. We list the classification accuracy (%) in the table.

Method	SP	TP	$\mathcal{L}_{SCL}^{S}$	$\mathcal{L}_{SSL}^{S}$	$\mathcal{L}_{SAL}^{S}$	$\mathcal{L}_{TCL}^{T}$	$\mathcal{L}_{TGL}^{T}$	Average Acc.
Baseline (CoTTA)	✗	✗	✗	✗	✗	✗	✗	54.6
SP only	✓	✗	✓	✓	✓	✗	✗	55.1
SP only-$\mathcal{L}_{SCL}^{S}$	✓	✗	✗	✓	✓	✗	✗	54.8
SP only-$\mathcal{L}_{SSL}^{S}$	✓	✗	✓	✗	✓	✗	✗	54.9
SP only-$\mathcal{L}_{SAL}^{S}$	✓	✗	✓	✓	✗	✗	✗	54.7
TP only	✗	✓	✗	✗	✗	✓	✓	58.9
TP only-$\mathcal{L}_{TCL}^{T}$	✗	✓	✗	✗	✗	✗	✓	54.9
TP only-$\mathcal{L}_{TGL}^{T}$	✗	✓	✗	✗	✗	✓	✗	50.9
TP only+$\mathcal{L}_{SCL}^{S}$	✗	✓	✓	✗	✗	✓	✓	59.2
TP only+ $\mathcal{L}_{SSL}^{S}$	✗	✓	✗	✓	✗	✓	✓	59.1
SoTa-DiT*	✓	✓	✓	✓	✓	✗	✗	60.4
SoTa-DiT	✓	✓	✓	✓	✓	✓	✓	**61.2**

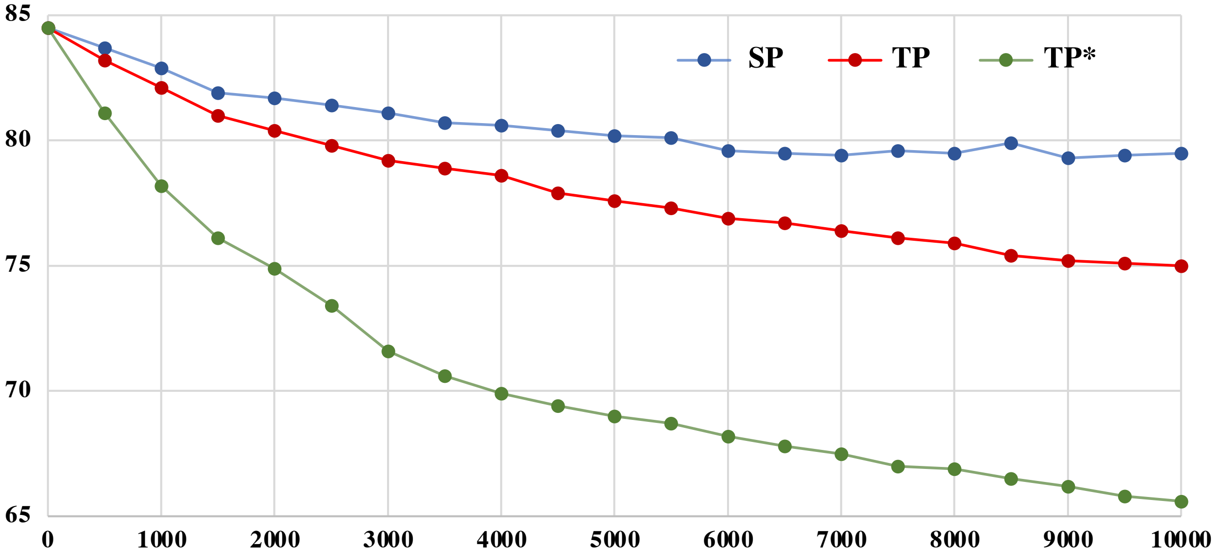

Figure 6. We evaluate the model directly on the original ImageNet test set at different time steps using 1) SP output only, 2) TP output only, and 3) TP output from a model without source prompt, denoted as TP$*$, respectively.

tively. Based on the result, we draw the following observations:

First, SP successfully preserves the source knowledge. Comparing the classification accuracy of SP output and TP output, we see that the classification accuracy of SP outputs decreases much slower than that of TP output before the time step 6000. After 6000, the classification accuracy of SP output remains stable. This phenomenon proves that SP successfully extracts and preserves the source knowledge from the source model.

Second, SP adapts the source knowledge and helps the model retain it. Comparing the classification accuracy of TP output and TP$*$ output, we see that the classification accuracy of TP output decreases much slower than TP$*$ output. This phenomenon indicates that with SP, the TP also retains the ability to classify the source data better, proving

Table 2. Average domain classification accuracy of SP output and TP output.

Method	Average Acc.
SP	12.9
TP	27.4

that SP helps the model preserve the source knowledge.

Overall, the results prove that SP successfully extracts the source knowledge and prevents the model from forgetting.

4.2. Target Knowledge Extraction

We first train an SoTa-DiT model with only the target prompt and an additional domain classification head to distinguish different test domains. Then, we incorporate a domain entropy loss to train the classification head. Note that domain entropy loss only tunes the domain classification head. Each test image, along with its domain label, is fed into the network. After training the network with all 15 corruptions on ImageNet-C, we take the domain classification head for further evaluation.

We then train a standard SoTa-DiT model. At each time step, the domain classification accuracy of the model is evaluated with the aforementioned domain classification head. The average classification accuracy is shown in Tab. 2, where we compare the domain classification accuracy of 1) SP output and 2) TP output.

The result shows that the TP output achieves 27.4% accuracy while the SP output achieves only 12.9%. This indicates that the TP embedding is more distinguishable regarding the domain, indirectly proving our knowledge disentangles claims and that TP extracts more target domain knowledge.

5. Limitation

The main limitation is that our method can only be applied to models with transformer layers. To use SoTa-DiT with the CNN backbones, one or a few transformer layers would need to be added.

Another limitation is that our work cannot be directly applied to small-scale datasets like CIFAR-10-C and CIFAR-100-C datasets. This is because directly training the ViT model on small-scale datasets without any pretraining does not yield satisfying results. The source model should be pre-trained on a large-scale dataset like ImageNet and then fine-tuned on small-scale datasets to achieve satisfying classification accuracy. As a result, we did not compare SoTa-DiT with other works on small-scale corruption datasets like CIFAR-10-C and CIFAR-100-C.

References