# Consistency-Guided Temperature Scaling Using Style and Content Information for Out-of-Domain Calibration

**Wonjeong Choi**[1], **Jungwuk Park**[1], **Dong-Jun Han**[2], **Younghyun Park**[1], **Jaekyun Moon**[1]
[1]Korea Advanced Institute of Science and Technology (KAIST)
[2]Purdue University
{dnjswjd5457, savertm, dnffkf369}@kaist.ac.kr, han762@purdue.edu, jmoon@kaist.edu

## Abstract

Research interests in the robustness of deep neural networks against domain shifts have been rapidly increasing in recent years. Most existing works, however, focus on improving the accuracy of the model, not the calibration performance which is another important requirement for trustworthy AI systems. Temperature scaling (TS), an accuracy-preserving post-hoc calibration method, has been proven to be effective in in-domain settings, but not in out-of-domain (OOD) due to the difficulty in obtaining a validation set for the unseen domain beforehand. In this paper, we propose consistency-guided temperature scaling (CTS), a new temperature scaling strategy that can significantly enhance the OOD calibration performance by providing mutual supervision among data samples in the source domains. Motivated by our observation that overconfidence stemming from inconsistent sample predictions is the main obstacle to OOD calibration, we propose to guide the scaling process by taking consistencies into account in terms of two different aspects - style and content - which are the key components that can well-represent data samples in multi-domain settings. Experimental results demonstrate that our proposed strategy outperforms existing works, achieving superior OOD calibration performance on various datasets. This can be accomplished by employing only the source domains without compromising accuracy, making our scheme directly applicable to various trustworthy AI systems.

## Introduction

Despite the huge success of deep neural networks (DNNs) in various fields such as computer vision (Krizhevsky, Sutskever, and Hinton 2017) and natural language processing (Mikolov et al. 2013), it is still difficult to actively utilize DNNs in safety-critical or high-risk applications including medical engineering and defect detection. One of the main reasons is that there is a huge risk when the prediction of the model becomes incorrect. In such applications, it is important for the neural network to reliably indicate whether its prediction is likely to be correct or not. In other words, the model should have a reliable confidence about its prediction. Given the model that has a reliable confidence, we can use the prediction of the model (e.g., for disease diagnosis) if the prediction is likely to be correct (i.e., high confidence), but may rely more on other factors (e.g., decision of the doctor) if the prediction is likely to be incorrect (i.e., low confidence). Therefore, promoting the quality of the prediction confidence of DNNs is a key mission for trustworthy AI (Tomani and Buettner 2021; Guo et al. 2017). However, the problem of modern DNNs is that they often produce over-confident predictions, which means that the confidence of predictions is not reliable.

In this context, model calibration has been developed as an important research direction to improve the reliability of confidence of DNNs. In general, we say that the model is well-calibrated if the confidence approximates the probability of being correct well. The authors of (Guo et al. 2017) have shown that modern neural networks have poor calibration performance due to over-confident predictions, and suggested a simple yet effective method termed temperature scaling (TS). TS is a post-hoc calibration method that calibrates the confidence of the trained model to make the confidence similar to a true probability using a validation set, without affecting the performance of the trained model. Various other methods such as train-time regularization (Thulasidasan et al. 2019; Krishnan and Tickoo 2020; Mukhoti et al. 2020; Hebbalaguppe et al. 2022) and probabilistic approaches (Gal and Ghahramani 2016; Lakshminarayanan, Pritzel, and Blundell 2017) have been also proposed in the literature improve the model's calibration performance.

However, most prior works focus on the in-domain setting, under the assumption that the domain distributions of train data and test data are the same. This assumption often does not hold in practice having domain shifts between training and testing: in out-of-domain (OOD) settings, the neural network needs to make predictions for samples from the unseen domain (target domain) that has not been observed in the training set (source domains). Existing calibration strategies face great challenges in the OOD settings due to the difficulty of knowing the unseen domain beforehand. Although several works (Tomani et al. 2021; Gong et al. 2021; Yu et al. 2022) have recently suggested some TS-based calibration methods under domain shift, their calibration performance is still limited when the disparity between the source and target domains becomes severe, as we will see in Section .

**Main contributions.** To handle the fundamental challenges of existing calibration methods in OOD settings, we propose consistency-guided temperature scaling (CTS), a new TS-based post-hoc strategy that can achieve superior OOD cali-

bration performance. Motivated by the intuition that keeping the consistency of sample predictions (regardless of domain shifts) can enhance the reliability of the model prediction in the unseen domain, our CTS trains a scaling temperature tailored to OOD scenarios, which enables the calibrated model to produce domain-invariant predictions. Our core idea is to take consistencies into account during temperature scaling, in terms of two different aspects: style and content, which are the key components representing data samples in multi-domain settings. To optimize a temperature, we compose a new TS objective function that includes two auxiliary losses for improving 1) inter-domain style consistency and 2) intra-class content consistency, thereby considering both aspects to achieve a better OOD calibration.

To gain insights into our idea and to support our claim, we start by analyzing the correlation between the two consistencies and OOD calibration, and show that the model's inconsistent predictions (on style and content variations) cause over-confident predictions on the target domain, leading to poor OOD calibration. Our CTS is designed to tackle this issue in OOD settings by optimizing the temperature to make predictions that are invariant to style and content shifts in the target domain. Surprisingly, our approach can significantly improve the OOD calibration performance (while preserving the accuracy) by strategically optimizing the consistency-guided temperature using style and content information only from the source domains. This is a key advantage of CTS as the temperature can be successfully optimized by utilizing the useful attributes in the source domains, without requiring any target domain information.

Experimental results on various multi-domain datasets show that CTS can achieve remarkable OOD calibration performance compared to existing TS-based methods. Notably, the performance gain of CTS becomes larger when the domain disparity is large, supporting the effectiveness of our method in practical OOD settings.

## Related Works

**Calibration methods.** A large body of literature has been devoted to improving the calibration performance of deep neural networks in in-domain settings. One of research directions is a train-time calibration (Thulasidasan et al. 2019; Krishnan and Tickoo 2020; Mukhoti et al. 2020; Hebbalaguppe et al. 2022; Liu et al. 2022), which prevents over-confident predictions by regularizing the model during training to enhance calibration performance. Another line of works is a post-hoc method (Guo et al. 2017; Vovk, Gammerman, and Shafer 2005; Zadrozny and Elkan 2002, 2001), which adjusts the confidence of the model output after training. Especially, temperature scaling (TS) (Guo et al. 2017) has been widely used since it can be easily combined with any trained model without compromising the original accuracy. TS prevents over-confident or under-confident predictions for test data by scaling the output with a single parameter (temperature) optimized on the validation set. However, although TS has been proven to be effective in in-domain settings, achieving a good calibration on OOD samples remains a great challenge due to the difficulty in obtaining a validation set for the unseen domain beforehand.

**Calibration for out-of-domain (OOD) scenarios.** Only a few works have proposed TS-based methods for OOD calibration. (Tomani et al. 2021) generates an augmented validation set which can simulate domain shift by injecting perturbations into validation samples before the post-hoc calibration. However, this scheme highly depends on the perturbation degree of the noise. The authors of (Gong et al. 2021) utilize multiple domains to reduce the distributional gap between the target and calibration domains for improved calibration transfer. Specifically, after clustering the calibration domains, the model is calibrated with the temperature of a specific group, which is most similar to the unseen domain encountered during the OOD inference. Similarly, (Yu et al. 2022) leverages multiple domains (with domain labels) to train a linear regression model which predicts sample-wise temperature for the target domain. However, the performances of these works (Gong et al. 2021; Yu et al. 2022) are still limited especially when the domain discrepancy between the calibration domains and the target domain is large.

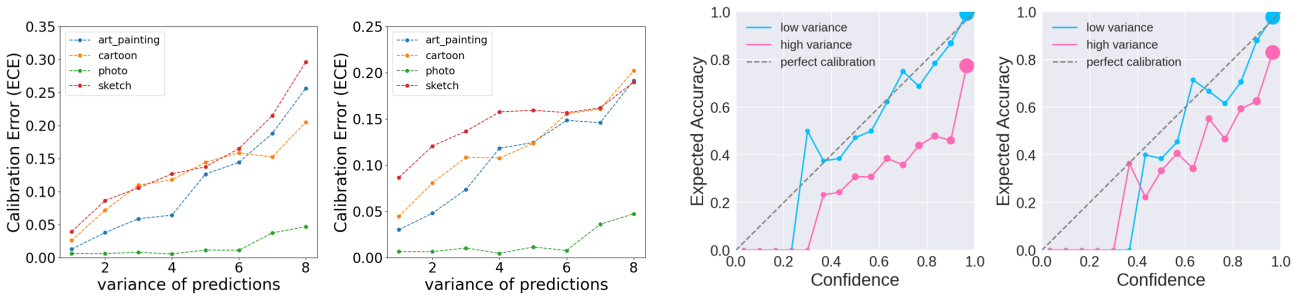
Compared to prior works, we take advantage of style and content information in the source domains and effectively optimize a temperature to enable the calibrated model to make domain-invariant predictions regardless of domain shifts, achieving superior OOD calibration performance even with a large domain gap between training and inference.

**Domain generalization (DG).** The goal of DG (Gulrajani and Lopez-Paz 2020; Zhou et al. 2021a) is to improve the generalization ability of DNNs, that is, models trained from the source domain work well even in the unseen target domain. The domain-alignment-based DG (Muandet, Balduzzi, and Schölkopf 2013; Motiian et al. 2017; Li et al. 2018) aims to learn features invariant to the domain shifts in the feature space through a domain-invariant learning. The augmentation-based method (Zhou et al. 2021c; Li et al. 2022; Shankar et al. 2018) prevents the model from overfitting to the source domains by using augmented data which simulate domain shift during training. Although existing DG methods have recently achieved high accuracy in unseen domains, one major bottleneck hindering their practical usability is a lack of calibration capability (Gong et al. 2021). In this paper, we present an effective calibration method that facilitates trustworthy AI systems in multi-domain settings by dealing with the calibration aspect for OOD scenarios.

## Background

### Temperature Scaling (TS)

As an accuracy-preserving approach that calibrates a pre-trained model in a post-hoc manner, TS (Guo et al. 2017) is widely applied in real applications due to its simplicity and effectiveness. TS aims to prevent the model from making over- or under-confident predictions by adjusting a single temperature $T$, without updating the parameters of the pre-trained model. The temperature is optimized by utilizing a validation dataset to minimize the negative log-likelihood (NLL) loss $L_{NLL}$ in the Eq. (1), using temperature scaled confidence score $\mathbb{P}(y = y_i|f_i, T) = \frac{\exp(f_{i,y_i}/T)}{\sum_{k=1}^{K}\exp(f_{i,k}/T)}$ for ground truth $y_i$, where $f_i$ is a logit vector of the sample $x_i$




(a) After style shifts (b) After content shifts (c) Style perspective (d) Content perspective

Figure 1: (a-b) Correlations between variance of predictions and OOD calibration performance for test samples from different target domains. Samples with high variance are more likely to show poor OOD calibration performance in both cases of style and content shifts on PACS dataset. (c-d) Reliability diagrams for comparing calibration tendency depending on the variance of predictions under style and content variations. It can be confirmed that the poor calibration performance of high variance samples (pink line) arises from the over-confident predictions. We note that the size of points indicates the relative counts of samples in each corresponding confidence interval.

and $f_{i,k}$ is $k$-th element of $f_i$ for class $k$. The optimized temperature $T^*$ can be defined as

$$T^* = \arg\min_T \left[ \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} L_{NLL} \right] = \arg\min_T \left[ \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} -y_i \log(\mathbb{P}(y = y_i | f_i, T)) \right], \quad (1)$$

where $N_{val}$ is the number of samples in the validation set. Conventional TS works well in in-domain scenarios where the data distributions between the validation set (for calibration) and test set are identical. However, in OOD scenarios, the model should be guided to make calibrated predictions on target domains unseen during training or calibration, which is a great challenge in the presence of domain shift (Tomani et al. 2021; Gong et al. 2021; Yu et al. 2022).

### Style Shifting

It is well known that feature statistics at early CNN layers can represent domain/style information of each data sample (Huang and Belongie 2017). Specifically, given a sample $x_i$, the intermediate CNN feature $z_i \in \mathbb{R}^{C \times H \times W}$ of $x_i$ at a specific layer can be interpreted as

$$z_i = \underbrace{\sigma(z_i)}_{\text{style}} \cdot \underbrace{\frac{z_i - \mu(z_i)}{\sigma(z_i)}}_{\text{content}} + \underbrace{\mu(z_i)}_{\text{style}}, \quad (2)$$

where $\mu(z_i) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} z_{i,(c,h,w)}$ and $\sigma^2(z_i) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} (z_{i,(c,h,w)} - \mu(z_i))^2$. Here, $C$, $H$, and $W$ denote the channel, height, and width of the feature maps, respectively. The feature statistics $\mu(z_i), \sigma(z_i)$ (also called style statistics) contain characteristics of each domain (i.e., style information) while content information is preserved in the style-normalized feature in Eq. 2. For notational simplicity, we define style $s_i$ and content $c_i$ of sample $x_i$ as $s_i := (\mu(z_i), \sigma(z_i))$ and $c_i := \frac{z_i - \mu(z_i)}{\sigma(z_i)}$. Based on this observation, adaptive instance normalization (AdaIN) (Huang and Belongie 2017) was introduced to transfer the style of $x_i$ to a different style of another sample $x_j$ by replacing the feature statistics from $s_i$ to $s_j$.

## Consistency-Guided Temperature Scaling

### Problem Setup

We consider a multi-class classification task with $K$ classes. Let $x_i$, $y_i$ denote the $i$-th image sample and the corresponding label drawn from a joint data distribution $P(x, y)$. For the multi-domain scenario, we let $P^{\mathcal{S}_j}$ denote the distribution of the $j$-th source domain and we have a total of $J$ domains $(P^{\mathcal{S}_1}, ..., P^{\mathcal{S}_J})$. Data from the source domains are separated into (i) *training set* that is used for training the base model and (ii) *validation set* utilized for post-hoc calibration. Given the model pre-trained on the training set, our goal is to learn a temperature $T$ that can well calibrate the model on an arbitrary target domain (with distribution $P^{\mathcal{T}}$) using the validation set, without changing the parameters of the pre-trained model (i.e., in a post-hoc manner). Note that if the model is perfectly calibrated in OOD scenarios, the confidence score should perfectly reflect the accuracy regardless of domains so that $\mathbb{P}(\hat{y} = y | \hat{p}_x) = \hat{p}_x$ holds for each sample $(x, y)$ drawn from any $P^{\mathcal{T}}$, where $\hat{y}$ and $\hat{p}_x$ are the predicted class and the confidence for sample $x$, respectively.

### Key Insights: Correlations between Consistency and OOD Calibration

Our intuition is that keeping the consistency of predictions under style and content perturbations can improve the reliability of the model prediction on the unseen domain, resulting in good OOD calibration. To gain insights into our idea based on this intuition, we provide an analysis of the correlation between OOD calibration and style/content consistencies using PACS dataset with 4 domains (Photo, Art, Cartoon, Sketch).

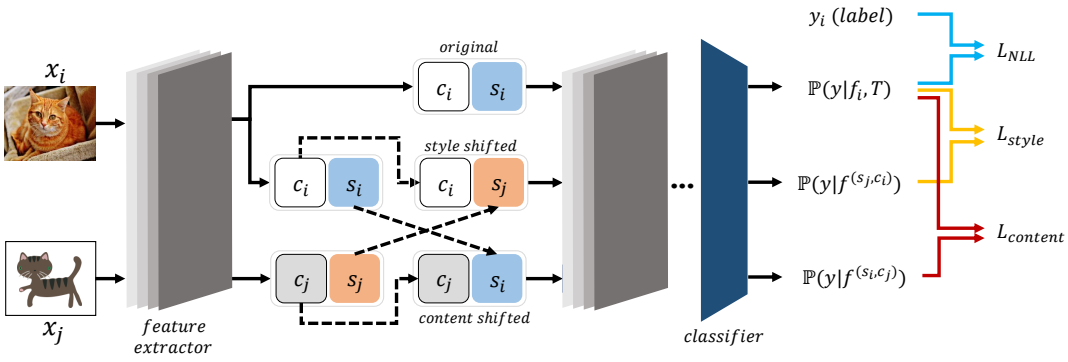



Figure 2: Overview of our consistency-guided temperature scaling (CTS). Samples from the same class on the validation set are fed into the model in a pair-wise manner, and three different intermediate features (original, style shifted, content shifted) are generated. Then, style/content shifted logits $\mathbb{P}(y|f^{(s_j,c_i)})$/$\mathbb{P}(y|f^{(s_i,c_j)})$ are created, and TS is performed with consistency losses described in Section .

**Setup.** To simulate the OOD scenarios, the base model is trained on three source domains and evaluated on the remaining one domain (target domain). Regarding the style consistency, given the base model pre-trained on the training set, we first obtain 4 different predictions for each test sample (in the target domain) by changing its style to 4 different styles (3 randomly chosen styles from each domain in the validation set, and 1 for the original style) while maintaining the content of each sample via Eq. 2. For the content consistency, we generate 5 different predictions by injecting different levels of zero-mean Gaussian noises with different variance [0, 0.1, 0.2, 0.3, 0.4] into the content $c_i$ in Eq. 2 of each test sample. After obtaining multiple predictions in each case, an indicator for each sample's style/content consistency is measured as follows: the variance of each element in the logit vector is first computed with different predictions, and then we average these results to obtain a scalar value. Finally, we sort test data according to their measured style/content consistency (i.e., variance of predictions), and measure the average expected calibration error (ECE) among samples with the same consistency value.

**Key observations.** Based on these measurements, Figs. 1a, 1b show the trend of ECE according to style/content consistency. It can be confirmed that the variance of predictions under style/content shifts ($x$-axis) is highly related to the calibration error ($y$-axis), which means that the higher the variance (i.e., the lower the consistency) under the style/content shifts, the higher the calibration error.

**Causality analysis.** To explore the cause-effect relation between calibration and the variance of predictions, we plot reliability diagrams (Niculescu-Mizil and Caruana 2005) of the top 5% samples and the bottom 5% samples in terms of variances of predictions after applying style/content shifts. Note that the perfect calibration is the identity function ($y = x$) while lines below the diagonal indicate that the model tends to output over-confident predictions, and vice versa. In Figs. 1c, 1d, it can be observed that inconsistent predictions with high variance (pink) tend to be over-confident compared to the consistent predictions (blue) after both shifts. These results indicate that samples that are sensitive to style and content variations cause unstable model predictions, and lead to over-confidence as shown in Figs. 1c, 1d. Inspired by these empirical observations, we propose CTS to improve OOD calibration on arbitrary target domains by optimizing the temperature such that the predictions are more consistent across variations in style/content.

## Proposed CTS for OOD Calibration

In this section, we introduce our method called consistency-guided temperature scaling (CTS). CTS strategically optimizes the temperature such that predictions are more consistent across the source domains. The core idea of CTS is to incorporate consistencies in two different aspects, style and content, which are independent components to represent data samples in multi-domain settings. We design new auxiliary cost functions for temperature optimization with consideration for style- and content-consistency to achieve better OOD calibration performance. We illustrate our CTS in Fig. 2 and explain the details of components in the following sections.

**Inter-domain style consistency loss.** The observations in Figs. 1a and 1c imply that consistency of predictions under style shift is closely related to OOD calibration. In particular, over-confident predictions tend to deteriorate calibration performance. Therefore, we aim to improve OOD calibration by rescaling predictions of the pre-trained model so that it can be consistent regardless of inter-domain style changes. Contrary to previous works (Motiian et al. 2017; Li et al. 2018; Kang et al. 2019) that focused on updating model parameters to learn style-invariant features during training, CTS optimizes a style-invariant temperature to recalibrate pre-trained models. In this context, using a validation set from source domains, we propose the style consistency loss $L_{style}$ as:

$$\begin{aligned} L_{style} &= D_{KL}[\mathbb{P}(y|f_i,T) \parallel \mathbb{P}(y|\overbrace{f^{(s_j,c_i)}}^{\text{style-shifted}})] \\ &= \sum_{k=1}^{K} \mathbb{P}(y=k|f_i,T) \log \frac{\mathbb{P}(y=k|f_i,T)}{\mathbb{P}(y=k|f^{(s_j,c_i)})}, \end{aligned} \tag{3}$$

where $D_{KL}[\cdot \parallel \cdot]$ is KL-divergence and $f^{(s_j,c_i)}$ is a style-shifted logit with style $s_i$ of sample $x_i$ substituted by $s_j$ of sample $x_j$. By strategically utilizing style information of validation samples from source domains, CTS can simulate a variety of style variations while optimizing the domain-invariant temperature to make consistent predictions. Concretely, as illustrated in Fig. 2, for the sample pair $(x_i, x_j)$, the corresponding intermediate features $(z_i, z_j)$ are separated into $(s_i, c_i)$ and $(s_j, c_j)$ through the operation in Eq. 2. Then, the style $s_i$ of sample $x_i$ is replaced with $s_j$ of sample $x_j$ to create a style-shifted feature, and it goes through the remaining layers to generate a style-shifted logit $f^{(s_j,c_i)}$ along with the original logit $f_i$. Lastly, the distance between temperature-scaled output $\mathbb{P}(y|f_i, T)$ and style-shifted output $\mathbb{P}(y|f^{(s_j,c_i)})$ is minimized such that the model is calibrated to generate style-invariant predictions. We note $L_{style}$ focuses on inter-domain style changes by minimizing the divergence between model outputs from features that differ only in style aspect.

**Intra-class content consistency loss.** Another important component that represents data samples in multi-domain settings is content: complex semantic features used to distinguish classes of samples. We note that even samples belonging to the same class can have different contents depending on some properties of the object (e.g., location, posture, shape, etc). As observed in Figs. 1b and 1d, content consistency is proportionally correlated with OOD calibration performance, similar to the trend observed with style. Also, models tend to make over-confident predictions on target samples that are relatively sensitive to the variations in content (as shown in Fig. 1d). Therefore, it is important to penalize such inconsistent predictions across intra-class content variation. To this end, CTS incorporates intra-class content consistency loss $L_{content}$ into the temperature optimization process as:

$$\begin{aligned} L_{content} &= D_{KL}[\mathbb{P}(y|f_i, T) \parallel \mathbb{P}(y|\overbrace{f^{(s_i,c_j)}}^{\text{content shifted}})] \\ &= \sum_{k=1}^{K} \mathbb{P}(y=k|f_i, T) \log \frac{\mathbb{P}(y=k|f_i, T)}{\mathbb{P}(y=k|f^{(s_i,c_j)})}, \end{aligned} \tag{4}$$

where $f^{(s_i,c_j)}$ is a content-shifted logit with content $c_i$ of sample $x_i$ substituted by $c_j$ of sample $x_j$. Note that $x_j$ belongs to the same category as $x_i$. As shown in Fig. 2, our CTS minimizes the divergence (as in Eq. 4) between the two predictions obtained from the original logit $f_i$ and the content-shifted logit $f^{(s_i,c_j)}$ while keeping style $s_i$ the same. After all, $L_{content}$ guides the model to make consistent predictions under the variation of contents.

**Overall loss of our CTS.** CTS operates in a pair-wise manner; two validation samples in a batch from source domains provide mutual supervision to each other, to find a scaling temperature that improves consistency in terms of style and content. Without any target domain information, CTS effectively takes advantage of the style and content features in the source domains for temperature scaling. Consequently, CTS obtains consistency-guided temperature $T^*_{CTS}$ from a validation set as:

$$\begin{aligned} L_{total} &= L_{NLL} + \lambda_1 \cdot L_{style} + \lambda_2 \cdot L_{content}, \\ T^*_{CTS} &= \arg\min_{T} \left[ \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} L_{total} \right] \end{aligned} \tag{5}$$

Note that $L_{total}$ combines NLL loss in Eq. 1 with style and content consistency losses in Eq. 3 and Eq. 4. Here, $\lambda_1$ and $\lambda_2$ are coefficients for each loss, which are used to adjust the balance of two losses. These coefficients can be determined depending on each dataset's distributional characteristics by using its validation set, where more details are provided in supplementary material [1].

# Experiments

## Experimental Settings

**Datasets.** We evaluate our CTS on four datasets that consist of multiple domains: PACS (Li et al. 2017), Office-Home (Venkateswara et al. 2017), Digits-DG (Zhou et al. 2020) and VLCS (Fang, Xu, and Rockmore 2013). PACS consists of 7 classes from 4 different domains and Office-Home has 65 classes from 4 domains. Digit-DG contains 10 classes from 4 domains and VLCS has 4 domains with 5 classes.

**Performance metric.** As in previous works on calibration, we use expected calibration error (ECE) (Naeini, Cooper, and Hauskrecht 2015) as a performance metric in our experiments. ECE measures the calibration error based on the expected absolute difference between the model's averaged confidence and its accuracy after binning. By dividing the confidence score range $[0, 1]$ into $R$ bins at even intervals $(q_r, q_{r+1}]$, we define the $r$-th bin $B_r$ as the indices of samples contained in that bin, i.e. $B_r = \{i = 1...N \mid \mathbb{P}(y = \hat{y}_i|\cdot) \in (q_r, q_{r+1}]\}$. Then, ECE can be calculated by $\sum_{r=1}^{R} \frac{|B_r|}{N} |Acc(B_r) - Conf(B_r)|$, where $Acc(B_r) = |B_r|^{-1} \sum_{i \in B_r} \mathbb{1}(\hat{y}_i = y_i)$ represents the expected accuracy of samples in $B_r$ and $Conf(B_r) = |B_r|^{-1} \sum_{i \in B_r} \mathbb{P}(y = \hat{y}_i|\cdot)$ is the averaged confidence of $B_r$. Here, $\mathbb{1}(A)$ is an indicator function with $\mathbb{1}(A) = 1$ if $A$ is true and $\mathbb{1}(A) = 0$, otherwise. $y_i$ is the ground truth class and $\hat{y}_i = \arg\max_y \mathbb{P}(y|f_i, T)$ denotes the predicted class. We set $R = 10$ during experiments. All results are averaged over 5 different runs and we report the results with 95% confidence intervals.

**Baselines.** We consider the following latest TS baselines: 1) *Vanilla* (Hendrycks and Gimpel 2017): The baseline without any calibration techniques applied; 2) *TS with val/test* (Guo et al. 2017): TS with validation of the source domains/test set of target domain. The latter can be viewed as one of the ideal cases where the target domain information is given during calibration, which is impractical in many cases; 3) *PerturbTS* (Tomani et al. 2021): TS by injecting some perturbations into the validation set; 4) *CCDG* (Gong et al. 2021): Cluster-level TS with multiple calibration domains in the validation set. We consider three variants of *CCDG*. *CCDG-NN* assigns each target sample to its nearest neighbor (NN) cluster's

[1] Our supplementary material can be found in the github page at https://github.com/wonjeongchoi/CTS.git

| | PACS | | | | | VLCS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Methods | Art | Cartoon | Photo | Sketch | Avg. | Caltech | LabelMe | Pascal | Sun | Avg. |
| Vanilla | 11.07 | 12.32 | 1.30 | 14.62 | 9.83 ± 0.45 | 2.31 | 32.01 | 12.68 | 20.71 | 16.93 ± 0.38 |
| TS with val | 9.70 | 11.30 | 1.01 | 14.21 | 9.06 ± 0.47 | 5.38 | 29.03 | 7.01 | 14.57 | 14.00 ± 0.39 |
| PerturbTS | 8.24 | 6.33 | 15.70 | 6.49 | 9.19 ± 1.19 | 18.30 | 23.84 | 9.04 | 9.32 | 15.13 ± 0.91 |
| MDTS | 10.33 | 10.98 | 1.49 | 15.18 | 9.50 ± 0.88 | 2.78 | 28.24 | 7.37 | 13.94 | 13.08 ± 0.38 |
| CCDG-NN | 9.92 | 10.82 | 1.30 | 13.16 | 8.80 ± 0.44 | 5.29 | 28.44 | 7.24 | 13.93 | 13.73 ± 0.43 |
| CCDG-Reg | 9.14 | 11.18 | 1.45 | 13.46 | 8.81 ± 0.61 | 4.97 | 29.65 | 6.91 | 14.03 | 13.89 ± 0.51 |
| CCDG-Ens | 9.89 | 11.33 | 1.15 | 13.86 | 9.06 ± 0.41 | 4.69 | 29.21 | 7.16 | 14.38 | 13.86 ± 0.43 |
| **CTS (ours)** | 3.20 | 3.05 | 8.59 | 2.38 | **4.31 ± 0.93** | 11.30 | 24.09 | 6.01 | 10.18 | **12.90 ± 0.81** |
| TS with test | 3.82 | 5.34 | 1.15 | 4.07 | 3.60 ± 0.24 | 2.15 | 15.10 | 6.96 | 10.53 | 8.69 ±0.19 |

| | Office-Home | | | | | Digits-DG | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Methods | Art | Clipart | Product | Real | Avg. | MNIST | MNIST-M | SVHN | SYN | Avg. |
| Vanilla | 12.22 | 17.14 | 6.27 | 6.25 | 10.47 ± 0.21 | 1.65 | 23.64 | 18.10 | 6.79 | 12.55 ± 0.17 |
| TS with val | 11.41 | 16.47 | 5.54 | 5.75 | 9.79 ± 0.19 | 1.13 | 21.21 | 16.60 | 5.38 | 11.08 ± 0.15 |
| PerturbTS | 7.52 | 10.66 | 2.96 | 3.02 | 6.04 ± 0.62 | 7.88 | 4.98 | 4.68 | 11.73 | 7.32 ± 0.93 |
| MDTS | 10.95 | 17.81 | 6.46 | 5.99 | 10.30 ± 0.42 | 1.24 | 20.70 | 16.59 | 4.85 | 10.85 ± 0.19 |
| CCDG-NN | 11.51 | 16.46 | 5.47 | 5.79 | 9.81 ± 0.16 | 1.09 | 21.06 | 16.41 | 5.27 | 10.96 ± 0.12 |
| CCDG-Reg | 11.01 | 17.61 | 6.34 | 5.95 | 10.23 ± 0.41 | 1.10 | 21.86 | 16.59 | 5.24 | 11.19 ± 0.18 |
| CCDG-Ens | 11.47 | 17.07 | 6.03 | 5.92 | 10.12 ± 0.23 | 1.11 | 21.37 | 16.60 | 5.35 | 11.11 ± 0.15 |
| **CTS (ours)** | 4.56 | 8.33 | 3.24 | 2.88 | **4.75 ± 0.52** | 2.05 | 12.50 | 9.68 | 1.89 | **6.53 ± 0.31** |
| TS with test | 3.07 | 4.26 | 2.28 | 2.59 | 3.05 ± 0.31 | 1.50 | 6.04 | 6.77 | 4.64 | 4.74 ± 0.11 |

Table 1: Calibration errors on four different datasets. Each column represents the target domain in one-domain-leave-out setting.

temperature. *CCDG-Reg* employs regression-based mapping from sample-wise feature to the cluster-level temperature. *CCDG-Ens* is a method of ensembles of *TS with val*, *CCDG-NN* and *CCDG-Reg* in a logit space; 5) *MDTS* (Yu et al. 2022): Multi-domain TS with linear regression of temperatures.

Implementation of each method follows the corresponding paper and we leave the details to the supplementary material. We also compare our method with train-time calibration methods in the supplementary material and show that CTS achieves comparable performance.

**Implementation details.** As each dataset consists of four domains, we separate them into three source domains and one target domain following the one-domain-leave-out setup (Zhou et al. 2021c; Li et al. 2022). The three source domains are split again into training set and validation set following (Zhou et al. 2022, 2021b), where we use the training set for training the model and the validation set for post-hoc calibration. For a fair comparison, for all baselines except *TS with test*, the OOD calibration performance is measured with the following process: (i) The model is first trained using the training set of source domains. (ii) Post-hoc calibration is conducted following each scheme using the validation set of source domains. (iii) ECE of each scheme is obtained using the target domain. During the training process in step (i), we utilize ResNet-18 (He et al. 2016) pre-trained on ImageNet (Krizhevsky, Sutskever, and Hinton 2017) and adopt the base training setup (without any DG techniques) following (Zhou et al. 2021c) for all schemes, to mainly focus on the effect of post-hoc calibration. Later in Fig. 3, we also combine the baselines and our CTS with existing DG methods (Zhou et al. 2021c; Li et al. 2022). For our CTS, the intermediate features for style and content shifts are extracted at the output of the first residual block of ResNet-18 in all experiments. Other details and ablation studies on the effect of the different layer are also reported in the supplementary material.

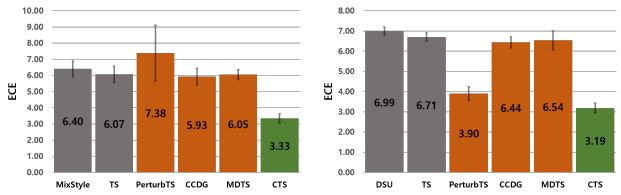
## Experimental Results

**Comparison with post-hoc OOD calibration methods.** Tab. 1 compares the ECEs of different schemes in each dataset. Since the post-hoc methods do not change the model parameters after calibration, all schemes have the same accuracy. In Tab. 1, we can see that our CTS achieves superior OOD calibration performance for all datasets compared to other post-hoc calibration methods. In particular, our method is much better for datasets with relatively large domain disparity, such as PACS and Digits-DG. This supports our claim that the temperature with style consistency can help achieve high calibration performance by regularizing the predictions for samples in the target domain, which has a large domain gap with the source domains. Also, for Office-Home dataset, when the coherent predictions within the same class are the matter due to the relatively large number of classes (65 classes in Office-Home), our CTS, which considers content consistency, can achieve the best performance by encouraging consistent predictions in the target domain. Since the model is pre-trained on ImageNet, we note that post-hoc calibration methods can negatively affect the calibration when the target domain is a photo in PACS dataset.

**Ablations for our CTS.** One key feature of our CTS is to balance between style and content consistencies by adjusting coefficients $\lambda_1$ and $\lambda_2$ in Eq. 5. This enables CTS to control the trade-off between two consistencies depending on the

| Methods | PACS | Office-Home | Digits-DG | VLCS |
|---|---|---|---|---|
| Vanilla | 9.83 | 10.48 | 12.55 | 16.93 |
| TS with val | 9.06 | 9.79 | 11.08 | 14.00 |
| Class-wise | 6.27 | 5.25 | 6.96 | 13.70 |
| CTS (only S) | 5.09 | 5.51 | 6.82 | 13.10 |
| CTS (only C) | 5.47 | 5.36 | 7.13 | 13.19 |
| **CTS (S&C)** | **4.31** | **4.75** | **6.53** | **12.90** |

Table 2: Ablation studies. We report the calibration errors for different variations of CTS: (i) class-wise, where the difference between confidence scores obtained from two samples within the same class is minimized; (ii) only S or only C, where either the style or content is solely considered instead of using both of them. The results demonstrate the effectiveness of each component in CTS with meaningful error gaps.




(a) Compatibility with MixStyle (b) Compatibility with DSU

Figure 3: Compatibility with augmentation-based DG methods on PACS dataset. Each scheme is combined with either MixStyle (Zhou et al. 2021c) or DSU (Li et al. 2022).

degree of disparity in terms of styles and contents, which may vary depending on the dataset. In Tab. 2, we conduct ablation studies on CTS by considering different variations of our method. One variation we can think of is to directly minimize the difference between confidence scores obtained from two samples within the same category, in a class-wise manner (i.e., class-wise in Tab. 2). Interestingly, the class-wise method achieves a relatively good calibration performance compared to other baselines (e.g., TS with val), which substantiates our claim that consistency-guided temperature is a good solution for OOD calibration. We also consider adopting either the style consistency loss or content consistency loss solely (i.e., only S or only C in Tab. 2), instead of using both of them. One interesting observation is that style-consistency (i.e., only S) has more advantages in datasets such as PACS or Digits-DG where styles are relatively the main cause of disparity compared to contents, while content-consistency (i.e., only C) performs better when contents become more important as in Office-Home. Our CTS considering both styles and contents can strategically balance between them depending on the dataset, taking the best of both worlds.

**Compatibility with existing DG methods.** Since our CTS is a post-hoc calibration method based on TS, it has the advantage of being easily combined with existing domain generalization methods without compromising accuracy. To demonstrate this, we measure the OOD calibration performance by applying calibration methods to MixStyle (Zhou et al. 2021c) and DSU (Li et al. 2022), which are the recently proposed augmentation-based DG approaches. As shown in Fig. 3, our CTS significantly improves the calibration performance of DG methods compared to other baselines. These results pave the way to apply existing DG methods in trustworthy AI systems, further confirming the advantage of CTS.

| # of cali. domains | Methods | Art | Cartoon | Photo | Sketch | Avg. | Acc |
|---|---|---|---|---|---|---|---|
| 2 | CCDG-NN | 10.36 | 25.75 | 6.97 | 10.52 | 13.40 | 50.20 |
| | CCDG-Reg | 11.33 | 22.96 | 5.97 | 12.82 | 13.27 | 50.20 |
| | CCDG-Ens | 11.00 | 22.15 | 6.86 | 11.89 | 12.98 | 50.20 |
| | MDTS | 14.80 | 20.11 | 8.27 | 12.10 | 13.82 | 50.20 |
| | **CTS** | 9.91 | 14.59 | 11.01 | 12.84 | **12.09** | 50.20 |
| 3 | **CTS** | 3.20 | 3.05 | 8.59 | 2.38 | **4.31** | 77.95 |

Table 3: Effect of domain composition in multi-domain calibration schemes. Our CTS still achieves higher OOD calibration performance than others, confirming the effectiveness of our consistency-guided approach.

**Effect of domain compositions in train/validations sets.** In Tab. 3, we also report the OOD calibration results of multi-domain calibration schemes when the training and validation sets have separated domain compositions without overlapping as in (Gong et al. 2021; Yu et al. 2022): One domain is used for training, two are utilized for post-hoc calibration, and the remaining one is used to measure the ECE and accuracy. PACS dataset is utilized for experiments. In this setup, the accuracy of each scheme decreases from 77.95% to 50.2% since the model is trained with only one source domain. Interestingly, despite the decrease in the number of calibration domains, our CTS still achieves higher OOD calibration performance than other baselines. These results show that CTS can effectively utilize the intrinsic attributes of validation samples to encourage consistent predictions regardless of style and content variations, even when the number of domains for calibration is limited.

**Additional experimental results.** Other experimental results including large-scale dataset, coefficient ablation ($\lambda_1, \lambda_2$), layer ablation, and comparison with train-time calibration methods are reported in the supplementary material.

## Conclusion

In this work, we proposed consistency-guided temperature scaling (CTS), a post-hoc calibration method that can be effectively used for robust OOD predictions. Our CTS strategically utilizes the style and content information of the source domains in a multi-domain setting, encouraging consistent predictions under domain shifts. By introducing auxiliary losses that can take consistencies into account in terms of two key aspects, styles and contents, our CTS can optimize the scaling temperature targeting OOD settings. Our approach achieves superior OOD calibration performance without requiring any target domain information and without compromising the model accuracy, making the scheme directly applicable to various trustworthy AI systems.

## Acknowledgments

This work was supported by IITP funds from MSIT of Korea (No. 2020-0-00626), NRF (No. 2019R1I1A2A02061135). Dong-Jun Han is the corresponding author.

## References

Fang, C.; Xu, Y.; and Rockmore, D. N. 2013. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, 1657–1664.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Gong, Y.; Lin, X.; Yao, Y.; Dietterich, T. G.; Divakaran, A.; and Gervasio, M. 2021. Confidence calibration for domain generalization under covariate shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8958–8967.

Gulrajani, I.; and Lopez-Paz, D. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hebbalaguppe, R.; Prakash, J.; Madan, N.; and Arora, C. 2022. A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16081–16090.

Hendrycks, D.; and Gimpel, K. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations*.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Kang, G.; Jiang, L.; Yang, Y.; and Hauptmann, A. G. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4893–4902.

Krishnan, R.; and Tickoo, O. 2020. Improving model calibration with accuracy versus uncertainty optimization. *Advances in Neural Information Processing Systems*, 33: 18237–18248.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84–90.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, 5542–5550.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5400–5409.

Li, X.; Dai, Y.; Ge, Y.; Liu, J.; Shan, Y.; and DUAN, L. 2022. Uncertainty Modeling for Out-of-Distribution Generalization. In *International Conference on Learning Representations*.

Liu, B.; Ben Ayed, I.; Galdran, A.; and Dolz, J. 2022. The devil is in the margin: Margin-based label smoothing for network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 80–88.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Motiian, S.; Piccirilli, M.; Adjeroh, D. A.; and Doretto, G. 2017. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, 5715–5725.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain generalization via invariant feature representation. In *International conference on machine learning*, 10–18. PMLR.

Mukhoti, J.; Kulharia, V.; Sanyal, A.; Golodetz, S.; Torr, P.; and Dokania, P. 2020. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33: 15288–15299.

Naeini, M. P.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Niculescu-Mizil, A.; and Caruana, R. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, 625–632.

Shankar, S.; Piratla, V.; Chakrabarti, S.; Chaudhuri, S.; Jyothi, P.; and Sarawagi, S. 2018. Generalizing Across Domains via Cross-Gradient Training. In *International Conference on Learning Representations*.

Thulasidasan, S.; Chennupati, G.; Bilmes, J. A.; Bhattacharya, T.; and Michalak, S. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Tomani, C.; and Buettner, F. 2021. Towards trustworthy predictions from deep neural networks with fast adversarial calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9886–9896.

Tomani, C.; Gruber, S.; Erdem, M. E.; Cremers, D.; and Buettner, F. 2021. Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10124–10132.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep hashing network for unsupervised

domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5018–5027.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Yu, Y.; Bates, S.; Ma, Y.; and Jordan, M. 2022. Robust calibration with multi-domain temperature scaling. *Advances in Neural Information Processing Systems*, 35: 27510–27523.

Zadrozny, B.; and Elkan, C. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, 609–616.

Zadrozny, B.; and Elkan, C. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 694–699.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2021a. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, K.; Yang, Y.; Hospedales, T.; and Xiang, T. 2020. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, 561–578. Springer.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2021b. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30: 8008–8018.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2021c. Domain Generalization with MixStyle. In *International Conference on Learning Representations*.