Single Image Test-Time Adaptation for Segmentation

Anonymous authors
Paper under double-blind review

Abstract

Test-Time Adaptation (TTA) methods improve the robustness of deep neural networks to domain shift on various tasks such as image classification, key-point estimation, or segmentation. However, the overwhelming majority of methods have been developed for image classification and new image segmentation methods are each evaluated under very different conditions and compared to a limited set of baselines, making understanding their performance difficult. This work explores adapting segmentation models to a single unlabelled image with no other data available at test-time. This allows for individual sample performance analysis while excluding orthogonal factors such as weight restart strategies. A diverse set of baselines, some modified from other domains or modalities, are first thoroughly validated on synthetic domain shifts and then tested on real datasets. The analysis highlights that simple optimization improvements such as proper choice of the loss function can greatly improve the performance of standard baselines such as pseudolabelling and that different methods and hyper-parameters are optimal for different kinds of domain shift, hindering the TTA performance where no prior knowledge about the domain shift is assumed.

1 Introduction

A common challenge in machine learning stems from the disparity between source, *i.e.* training, and target, *i.e.* deployment, data domains. Models optimized to minimize an error on a dataset from a specific domain are often expected to perform reliably in different domains. The discrepancy between training and deployment data, known as "domain shift", is very common; in fact, few things do not change in time, and training happens (well) before deployment. A domain shift may substantially degrade model performance at deployment time despite proper validation on training data, yet it is often not explicitly addressed and most machine learning effort has focused on the generalization problem.

In many practical scenarios, the characteristics of the target domain are not known beforehand, making the preparation of the model with traditional domain adaptation Rodriguez & Mikolajczyk (2019); Tzeng et al. (2017) techniques non trivial. Recent advances Wang et al. (2020b); Sun et al. (2020); Gandelsman et al. (2022) in the field suggest that under certain weak assumptions about the domain shift - such as stable label distribution across domains - it is possible to mitigate the performance degradation with methods based on the premise that input data, received during inference, carry information about its distribution that can be exploited for adaptation.

Test-Time Adaptation (TTA) is suitable for apriori unknown or difficult to predict domain shifts. Characterized as an unsupervised and source-free technique, TTA operates under the principle of adapting the model directly at the time of inference. The source-free nature, *i.e.* without access to the original training data, ensures compliance with data governance standards and enables adaptation in memory-constrained environments.

Single Image Test-Time Adaptation (SITTA) tailors a segmentation model at test time to each individual image. Since it operates on a single image, it does not introduce assumptions about the stability of data distribution over time. Each time starting from the weights fixed at training time, SITTA is safe to use when any form of memorization of the deployment data is prohibited. A major disadvantage is an increased computational time and no possibility to reuse the acquired knowledge. On the other hand, it could be

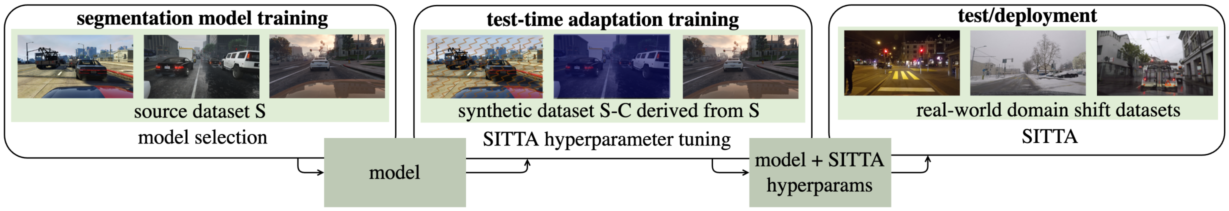

segmentation model training
source dataset S
model selection
model
test-time adaptation training
synthetic dataset S-C derived from S
SITTA hyperparameter tuning
model + SITTA hyperparams
test/deployment
real-world domain shift datasets
SITTA

Figure 1: The proposed experimental framework for Single Image Test-Time Adaptation (SITTA). The SITTA hyper-parameters are found on a synthetic dataset derived from the training set by applying a diverse set of corruption of different levels. The SITTA is then tested on real-world datasets with domain shift.

leveraged to reduce the computational cost of adapting to a sequence of similar images by only adapting to the most informative samples. Despite the advantages, only Khurana et al. (2021) primarily address SITTA. The mainstream of TTA research has centered on continual test-time adaptation in a changing environment Niu et al. (2023); Volpi et al. (2022); Wang et al. (2022). These methods typically gradually update the model parameters or accumulate image statistics for subsequent adaptation to individual images. While practical in many applications such as autonomous driving, it presents challenges for accurately assessing the strengths and weaknesses of different TTA strategies. This difficulty arises due to the evaluations being conducted over long sequences of images of varying levels and kinds of domain shift. Moreover, the sequence in which images are presented can significantly influence the overall performance metrics, adding a layer of complexity to understanding the true efficacy of these methods. acsitta not only streamlines the evaluation process but can also enhance our understanding and effectiveness in broader TTA applications where the focus is currently more on issues such as catastrophic forgetting, which are orthogonal to our work.

In this paper, we explore and improve the state of the art in Single Image Test-Time Adaptation (SITTA) for image segmentation. We center on SITTA with self-supervised loss functions as these methods are broadly applicable and easily adoptable across various segmentation models and domains, without being constrained by specific architectural or training requirements. Among the methods we employ, the strongest assumption is access to training data prior to deployment. Consequently, the many works relying on the existence of batch-normalization layers are not included in this work as they are incompatible with the modern transformer architectures which currently dominate the field. Likewise, methods such as image-reconstruction based TTA Wang et al.; Gandelsman et al. (2022) are not included due to their significant demands for training process modification and model architecture adjustments.

Five different TTA methods are evaluated in the SITTA setting. Most are evaluated on SITTA for the first time, some are introduced from other domains (medical) and some have never been evaluated on image segmentation. The methods include two established baselines: Entropy minimization and pseudolabelling (self-training). These are the only methods evaluated in the SITTA setup before. Other methods are a recently proposed segmentation TTA method that utilizes augmentation consistency to identify confident pseudolabel pixels and an additional method from image classification which robustifies the network to small domain shifts (represented by adversarially-attacked images), expecting increased robustness to other domain shifts as well. We also extend methods based on optimizing self-supervised loss functions with learned mask discriminators or mask refinement modules originally developed for medical image segmentation Karani et al. (2021); Valvano et al. (2021). While medical domain and its domain shifts of interest are different, these methods are general and easily applicable to non-medical images. They have not been applied to the traditional image segmentation TTA. Diverse self-supervised loss functions for image segmentation TTA generally remain underexplored. To date, the focus has predominantly been on prediction entropy and reconstruction loss. Our work aims to bridge this gap by investigating and applying these advanced segmentation-specific self-supervised techniques, originally conceived for medical imaging, to traditional image segmentation tasks.

We modify the learning of mask-refinement module to generate enhanced pseudolabels at test-time. The method is adapted to traditional images by accounting for more complex domain shifts. Novel adversarial training of the mask refinement module is introduced, replacing the originally proposed heuristic of swapping image patches. Swapping patches may results in unrealistic images and therefore masks and requires finding a suitable size of the patches and the ratio of swapped patches. Controlling the severity of mask corruption is non-trivial. Motivated by the intuition that the first pixels to be inverted in an untargeted adversarial attack are also those most likely to be impacted by domain shift, this work simulates domain shift impact on masks by projected gradient descent adversarial attack on the input image. This results in realistic corrupted masks in the first iterations of the attack. The adversarial optimization learning rate and the number of iterations then control the severity of the domain shift.

We address the problem that arises from the practice in the current landscape of segmentation TTA that the performance assessment is carried out with inconsistent adaptation settings. For instance, keeping batch normalization statistics constant or updating them during entropy minimization can yield substantially different outcomes. If only one of the two options is tested Khurana et al. (2021); Volpi et al. (2022), contradictory results are reported. Often, the evaluation compares only with the established baselines such as entropy-minimization Wang et al. (2020a) and batch-normalization Ioffe & Szegedy (2015a) statistics adaptation Schneider et al. (2020); Nado et al. (2020), ignoring recent progress. Many methods that improve over these baselines have been proposed in recent years for both segmentation and classification Nguyen et al. (a); Gao et al. (2023); Niu et al. (2023); Chen et al. (2022), and their relative merit is unknown, since a comprehensive comparison with a well-defined methodology is missing.

Our experimental framework, depicted in Figure 1, involves testing with two different pretrained segmentation models. The first model was trained on the GTA5 Richter et al. (2016) datasets and tested on the Cityscapes Cordts et al. (2016) and ACDC Sakaridis et al. (2021) benchmarks. The other model is trained on COCO Lin et al. (2014) and evaluated on VOC Everingham et al. (2010). To fine-tune hyper-parameters and conduct a majority of our method performance analyses, we utilize an augmented version of the training datasets. This extension incorporates synthetic corruptions, encompassing a broad spectrum of corruption types and levels, inspired by Hendrycks & Dietterich (2019). The corruption types include different kinds of noise and blur, weather conditions such as fog or frost, as well as the jpeg compression and basic image intensity transformations. The advantage over commonly used synthetic datasets such as Synthia Ros et al. (2016) is that it can be derived from an arbitrary segmentation training dataset, providing precise control over the conditions and facilitating detailed analysis. The main limitation of this methodology is the focus on synthetic covariate shifts means that our evaluation does not encompass real-world domain shifts, which can be more complex and less predictable than their synthetic counterparts.

The main contributions of this paper are:

1. We conduct a comparative analysis of five different TTA techniques run in SITTA mode for image segmentation including established baselines and recently proposed promising methods. This work is the first to apply methods developed in the medical imaging domain to traditional image segmentation, filling a gap in the exploration of diverse self-supervised loss functions.
2. A novel adversarial refinement module training for Mask Refinement (Ref)-based TTA.
3. Improvements of baselines in single-image setup by replacing Cross-Entropy (CE) loss with Intersection over Union (IoU) loss.
4. First work emphasizing SITTA for segmentation, an underexplroed setup important in applications with strict data governance standards or high variability among individual images.
5. Analysis of SITTA performance emphasizing the unpredictability and variability of target data domains compared to source domains.

2 Background

Common approaches to domain adaptation change the style of labelled source images to resemble the training images Tzeng et al. (2017); Zhang et al. (2018) or train domain classifiers to guide the adaptation process. In practice, this is not always feasible since source data may not be available for example for privacy or memory limitation reasons, or we may only have a small number of target domain images available when data arrive individually/in small batches, rather than all at once. In continually evolving environments, the distribution may change by the time adaptation on a large target dataset is completed. Various modifications of the traditional domain adaptation scenario tackling the aforementioned limitations have recently emerged, for example by considering no access to source data or a continual domain shift Liu et al. (2021a); Volpi et al. (2022); Wang et al. (2022); Bartler et al. (2022).

In particular, test-time adaptation methods assume no source data is available and aim to exploit the information from as little as a single target domain image. Like other domain adaptation methods, TTA methods are often inspired by semi-supervised learning methods. For instance, the most common TTA baseline relies on minimization of the predictions entropy, a method inspired by Saito et al. (2019). Other methods rely on adapting the bacth-normalization statistics, inspired by methods like adaptive batch-normalization Li et al. (2018), or aggregating statistics to create so-called prototypes Tanwisuth et al. (2021) that can be used to build a classifier.

Some works also distinguish between TTA and Test-Time Training (TTT). The difference between TTA and TTT is that TTA methods such as Nguyen et al. (b); Karani et al. (2021) can be applied to arbitrary pre-trained models without any additional constraints while TTT methods like Gandelsman et al. (2022); Bartler et al. (2022); Liu et al. (2021b) require modifications to the training process. However, not all works make this distinction and the boundary is not always clear, as some methods like Karani et al. (2021) require to train an auxiliary deep net on the source data but do not modify the model pretrained weights. In this work, both will be jointly referred to as TTA for simplicity.

In Appendix A, other related domain adaptation scenarios and their relation to TTA are described.

Generally, TTA methods can be split into three groups: Adaptation in the input space, feature space and output space. **Input space adaptation** aims to translate the images from the source domain to the input domain. In practice, Gao et al. (2023) achieve this by feeding target images with added noise to a diffusion model trained on the source data, coupled with reconstruction guidance to preserve semantics. The model doesn't retain any knowledge from the adaptation, which can be both advantage as it is not susceptible to catastrophic forgetting, but it may limit the adaptation capabilities. **Adaptation in the feature space** is the most common approach and typically relies on optimizing the network parameters via a self-supervised loss function. This can be done directly, i.e. through prediction entropy minimization, or by training an auxiliary task such as image reconstruction. Another set of feature-adaptation approaches are parameter-free and rely on accumulating the image statistics, such as the mean and variance of image features, or by aggregating confident prediction features into so-called prototypes, which are then used for classification. **Output space adaptation** techniques aim to improve the network output without neither altering the network parameters and statistics nor the input image. This is done for instance in Karani et al. (2021) where an auxiliary network is trained to predict a refined mask. To the best of our knowledge, output space adaptation methods are typically only used to provide pseudo-masks, turning them into feature-space adaptation methods. This helps to iteratively improve the pseudo-masks and adapt to larger domain shifts. All the methods evaluated in this work can be considered as feature space adaptation methods, possibly via output space adaptation.

3 Related Work

Test-Time Adaptation methods for classification. Many recent methods propose improved strategies to update the batch normalization statistics Schneider et al. (2020); Nado et al. (2020). A limitation of these methods is the reliance on presence of batch nromalization, which is often not part of recent transformer-based architectures. In Wang et al. (2020b), the learnable parameters of the normalization layers are also updated via entropy minimization. While this method is often reported as unstable since single-image statistics may not be sufficient, the method can also only update the normalization layers learnable parameters, without the statistics update, making it generalizable to all currently used architectures.

On classification tasks, many methods outperforming the aforementioned baselines have been proposed. A combination of self-supervised contrastive learning to refine the features and online label refinement with a memory bank is proposed in Chen et al. (2022). Recently, a method based on updating the parameters of the normalization layers of the network by optimizing it for robustness against adversarial perturbation as a representative of domain shift was proposed in Nguyen et al. (b), outperforming similar test-time adaptation approaches. Rotation prediction is proposed in Sun et al. (2020) as self-supervised task to be learnt alongside the main one and then optimized at inference time. Lately, it was shown that reconstruction with masked auto-encoders is a very strong self-supervised task for test-time adaptation of classifiers by Gandelsman et al. (2022).

Test-Time Adaptation methods for segmentation. To the best of our knowledge, the only work also focused on adaptation to a single isolated image Khurana et al. (2021) is based on computing the statistics from augmented version of the input image, assuming batch-normalization layers are present in the network. Both Prabhu et al. (2021) and Wang et al. (2022) exploit augmented views of the input images to identify reliable predictions. The method of Prabhu et al. (2021) is based on the consistency of predictions between augmented views, which replaces prediction confidence for selecting reliable pixels. Cross entropy loss is then minimized on such reliable predictions, together with a regularization based on information entropy Li et al. (2020) to prevent trivial solutions. The method achieves impressive results, however, in contrast to our experiments, knowledge of the target domain shift is used for hyper-parameter tuning. The evaluation assumed a full test set available at once, focusing on source-free domain adaptation, rather than TTA, but the method is applicable to the TTA setup as well. In Volpi et al. (2022), the performance of entropy minimization in a continual setup is explored, proposing parameter restart to tackle weight drift, significantly improving performance. The focus is on driving datasets only. Similarly, Wang et al. (2022) also focus on continual adaptation. Again, augmentations of the images are generated to obtain more reliable predictions. Further, the network parameters are stochastically reset to their initial values to prevent forgetting of the source domain knowledge.

Test-Time Adaptation methods for medical imaging. In Karani et al. (2021), an autoencoder is proposed that translates predicted masks into refined mask. At test time, the segmenter is optimized to produce masks closer to the enhanced ones. However, this work assumes the whole test dataset is available at once, in contrast to our single-image setup. The work of Valvano et al. (2021) is similar to Karani et al. (2021) but instead of a masked-autoencoder, a GAN-like discriminator trained end-to-end together with the segmenter is used, as well as an auxiliary reconstruction loss.

These works assume domain shifts specific to the medical imaging domain such as the use of a different scanner and thus make the assumption that only low-level features are affected. Under this assumption, these works typically optimize a small adapter only, ie. the first few convolutional layers of the segmenter. Nonetheless, these methods are generalizable to image segmentation.

Enhancing existing TTA benchmarks. There are multiple concurrent works that identify similar issues and reporting results consistent with our experiments, mostly for image classification. The work of Yu et al. (2023) also highlights the issue of evaluating each method under very different conditions and provides a benchmark for image classification TTA encompassing different adaptation scenarions, as well as diverse backbones and domain shift datasets. Similarly to ours, a significant disparity between synthetic corruptions performance and natural shifts is observed. However, the hyper-parameters were selected based on a single kind of domain shift, which may bias the results. Another work adressing the issue of fair comparison of TTA methods is that of Mounsaveng et al. (2024) which provides an analysis of existing orthogonal classification TTA methods. Class rebalancing is one of the tricks proposed to improve the methods' performance. Also, sample filtration to remove noisy high-entropy images is employed. In contrast, we analyze performance of different methods based on prediction entropy, showcasing some methods can actually be highly effective on those noisy, high-entropy samples. Similarly to ours, the work shows that baselines can be greatly improved by very simple tricks. In Niu et al. (2023), label imbalance at test-time is again identified as an important factor harming the TTA performance. Again, the works focus is on image classification and epxlores different normalization layer kinds and stabilization techniques of entropy minimization while we focus on comparions of cross-entropy and a class-imbalance aware segmentation loss function, the IoU. Finally, Yi et al. (2023) study TTA for image segmentations and how well classification methods transfer to semantic segmentation TTA. They conclude that many of the classification TTA improvements do not transfer to segmentation and again highlight the class imbalance, which is typically greater for segmentation datasets.

4 Methods

In total, six different methods are implemented and evaluated, including both traditional TTA baselines and methods form other tasks or modalities. All of the methods consist in optimizing a self-supervised loss, the specifics of the loss being what differentiates the methods. It can be formalized as follows:

$$\theta_{i+1} = \arg\min_{\theta_i} \mathcal{L}(f_S^{\theta_i}, x)$$

where θ_i are the parameters of the segmentation network $f_S^{\theta_i}$ at the i-th iteration and $\mathcal{L}$ is the self-supervised loss function.

The methods considered are:

- **Entropy-Minimization (Ent)**, a method proposed by Wang et al. (2020b) inspired by semi-supervised learning where the self-sueprvised objective is the prediction entropy. It has been used as a baseline by the majority of the TTA work. In most work, only normalization layer parameters are updated to improve time efficiency. Whether batch-normalization statistics are updated as well varies.
- **Pseudo-Labelling (PL)**, also commonly referred to as self-training. The model is finetuned with pseudo-labels obtained from the pretrained segmentation model. There are many improvements and modifications, the standard approach is to threshold the predicted probabilities and only train the model on the most confident predictions.
- **Mask Refinement (Ref)** can be considered an enhanced pseudolabelling method where the pseudo-labels are obtained by a learnt refinement module that takes logit masks as inputs and outputs a refined segmentation mask. The idea has already been implemented in medical imaging Karani et al. (2021) but never tested on non-medical tasks.
- **Deep-Intersection-over-Union (dIoU)** is similar to Ref. However, a single-scalar quality estimate is predicted by a learnt module and minimized at test-time. It is similar to using a GAN-like discriminator.
- **Augmentation-Consistency (AugCo)**, proposed by Prabhu et al. (2021), is a method based on self-training enhanced by also optimizing for consistency between the original prediction and the prediction on augmented views, adapted to the single isolated image scenario.
- **Adversarial-Attack (Adv)** is the method proposed by Nguyen et al. (b) for image classification TTA, adapted to the single, isolated image segmentation.

Only the necessary modifications to make the methods applicable to the single, isolated image segmentation setup with no assumptions of specific network architecture were applied to the existing methods.

The only substantially different method from previous work is the learnt mask-refinement module, which will be described in the rest of this section. A description of other methods, as well as the details of their modifications in this work, is in Appendix B.

The rest of this section describes the TTA with mask refinement in more detail, including novel adversarial training of the refinement module proposed in this work.

TTA with mask refinement is based on the idea that since the output space changes much less than the input space, a mask translation module can be learnt to refine mask predictions on images from target distribution to resemble the masks obtained from source images. This is similar to domain adaptation methods based on learning a discriminator between source and target domain, with two key differences: Instead of a binary discriminator, a mask refinement module is learnt, so the output is not a scalar but a new segmentation mask. Also, the target domain is not known in advance and cannot be used to train the refinement network. At test time, the refinement network can be viewed as an enhanced pseudo-label

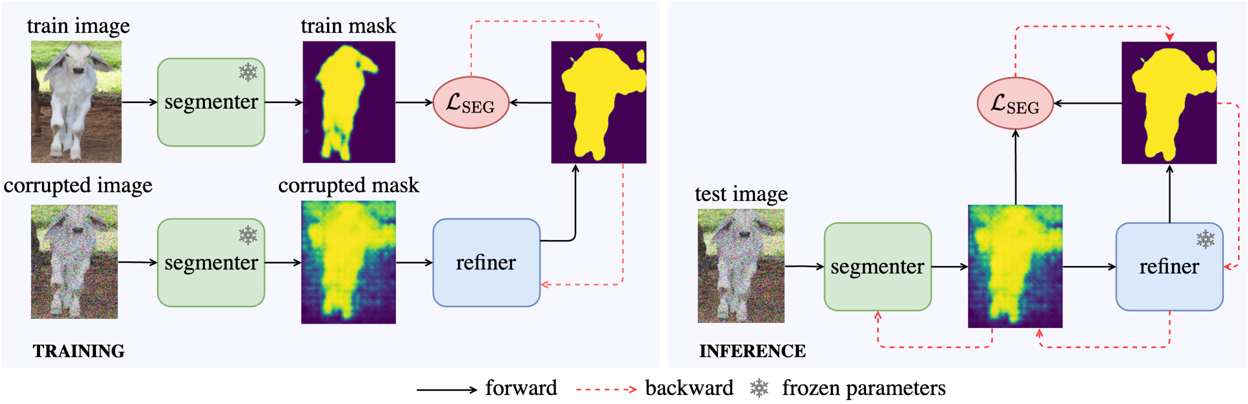


Figure 2: Mask refiner training (left) and Mask Refinement (Ref) TTA inference (right). During training, the segmenter outputs masks from a training image and a corrupted image simulating domain shift. The mask refiner is then trained to predict the clean image mask given the corrupted image mask as input only - no gradients flow to the segmenter. At inference time, the segmenter output is fed into the refiner model. The refiner model is then used as an enhanced pseudo-label to finetune the segmenter. A single gradient update is performed in each TTA iteration, then the masks are updated, since the segmenter output may have changed with the new weights, which in turn results in a new, possibly better, pseudo-label from the refiner. Visualized on single class prediction.

generation method. These pseudo-labels can then be used both as supervision for the segmenter, or directly replace the segmentation output without any parameter optimization. However, the second option is unlikely to tackle highly distorted masks since the refined mask cannot improve gradually.

To train the refinement network f_{R}^{ϕ} with learnable parameters ϕ, images from the source distribution and the pretrained segmentation network f_{S}^{θ} are required. Given an image x and x' generated from x by synthesizing a covariate domain shift (not changing the label), let us denote as $s = f_{\text{S}}^{\theta}(x)$ and $s' = f_{\text{S}}^{\theta}(x')$ the corresponding segmentation masks. Then, f_{R} is trained to predict s, given s' as input:

$$\arg\min_{\phi} \mathcal{L}_{\text{CE}}(f_{\text{R}}^{\phi}(s'), s) \tag{1}$$

Predicted masks s can also be replaced with ground truth g at training time:

$$\arg\min_{\phi} \mathcal{L}_{\text{CE}}(f_{\text{R}}^{\phi}(s'), g) \tag{2}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss.

At test-time, adapting to an image x, the model parameters are updated to minimize the IoU loss between mask prediction and a refined mask estimated by f_{R}:

$$\theta_{i+1} = \arg\min_{\theta_i} \mathcal{L}_{\text{IoU}}(f_{\text{R}}(\overline{f}_{\text{S}}^{\theta_i}(x)), f_{\text{S}}^{\theta_i}(x)) \tag{3}$$

where θ_i are the learnable parameters of f_{S}^{θ} at optimization iteration i and $\overline{f}_S$ denotes no gradient flow throughout the computations of f_{S}.

An overview of the training pipeline, as well as the TTA with mask-refinement, is in Figure 2.

Refinement module training requires generating masks resembling those that the model would output under domain shift. Since TTA assumes the domain shift is not known in advance, the goal is to generate a diverse set of domain shifts well representing realistic masks under domain shift. The advantage of the refinement module is that only the output space corrupted masks are needed, it doesn't matter how these were obtained since the refinement module is independent of the input images. In Karani et al. (2021), the corrutped masks are obtained through swapping input image patches, a heuristic method with many hyper-parameters that could lead to unrealistic artifacts in the masks. This work simulates the mask corruptions by using mask predictions on the images from the first few iterations of a Projected Gradient Descent (PGD) Madry et al. (2017); Kurakin et al. (2018) adversarial attack, using the inverted mask as target. The more iterations of the attack, the higher the mask corruption, but the less realistic it becomes. Examples of generated corrupted masks are shown in Appendix C. The intuition behind this adversarial approach is that in the first iterations, the most challenging pixels for the network are converted. Similarly, those image areas could be easily impacted by domain shift.

5 Experiments

The structure of this section is as follows:

1. **Evaluation metrics:** Given the focus of this study on SITTA and per-image performance analysis, we underscore the need for an image-level evaluation metric. The widely used mean Intersection over Union (mIoU) metric is typically applied at the dataset level and its adaptation for image-level assessment is not standardized.
2. **Experimental Setup:** Experiment settings shared across experiments such as network architectures or hyper-parameters. Creation of the synthetic SITTA training set derived from the segmentation training dataset is also explained.
3. **Experimental Results and Analysis:** Experiment results and analysis. The TTA methods are evaluated on two semantic segmentation models pretrained on the GTA5 Richter et al. (2016) and COCO Lin et al. (2014) datasets.

5.1 Evaluation metrics

The standard semantic segmentation evaluation metric is the mIoU, where the IoU score of each class is computed from predictions aggregated over the whole dataset

$$\mathrm{mIoU} = \frac{1}{\mathrm{C}} \sum_{c=1}^{\mathrm{C}} \mathrm{IoU}_c(m_c, g_c) \tag{4}$$

where m_c, g_c are the predictions and ground truth values for class c for all pixels across all images. Concatenating all the masks into a single one and then computing the metric would not change the results, each pixel has the same weight. This metric does not consider the size of objects or the difficulty of individual images. Per-image results cannot be compared, since not all classes are typically present in an image.

Two additional metrics are introduced to account for the limitations of the standard mIoU and make the evaluation more fine-grained. The first metric is designed to consider class imbalance and difficulty of individual images, focusing on per-class performance. It will be referred to as $\mathrm{m}\overline{\mathrm{IoU}}_c$ and is defined as

$$\mathrm{m}\overline{\mathrm{IoU}}_c = \frac{1}{\mathrm{C}} \sum_{c=1}^{\mathrm{C}} \frac{1}{|\mathrm{I}_c|} \sum_{i \in \mathrm{I}_c} \mathrm{IoU}(m_{ic}, g_{ic}) \tag{5}$$

where I_c is the set of images in which either the prediction or the ground truth mask contains class c and $|\mathrm{I}_c|$ is the total number of pixels.

The second metric is focused more on per-image performance and can be computed for a single image. It will be referred to as $\mathrm{m}\overline{\mathrm{IoU}}_i$ and is defined as

$$\mathrm{m}\overline{\mathrm{IoU}}_i = \frac{1}{|\mathrm{I}|}\sum_{i\in \mathrm{I}} \frac{1}{|\mathrm{C}_i|}\sum_{i\in\mathrm{C}_i} \mathrm{IoU}(m_{ic}, g_{ic}) \tag{6}$$

where C_i is the set of classes in the predicted masks and the ground truth. This is the metric reported in our experiments, unless stated otherwise. It allows for per-image performance comparison with the disadvantage of not accounting for class imbalance - less frequent classes (on the image-level) get smaller weight .

Similar metrics were recently considered by other works Volpi et al. (2022), typically only aggregating over images where the given class appears in the ground truth (as opposed to either ground truth or the prediction). This has the advantage that mistakes are only accounted for once, making the metric more optimistic than ours. On the other hand, information about the errors is lost, since the error is only computed for the ground truth class independently of what the incorrectly predicted class is.

5.2 Experiment setting

SITTA training set. The SITTA training set for each model is derived from a set of 40 images from the segmentation model's training dataset extended with a set of 9 synthetic corruptions at three severity levels from Hendrycks & Dietterich (2019) such as blur, noise or fog, simulating different domain shifts. The original images are also included, since the TTA methods should not harm the model on source domain images. Details about the corruptions can be found in Appendix D. These synthetic datasets based on the GTA5 and COCO datasets are referred to as GTA5-C and COCO-C, respectively. Since the original images without any corruption are also included, each SITTA training dataset consists of 1200 images in total (40 images, 9 + 1 corruptions, 3 corruption levels).

TTA hyper-parameters. For each TTA method, optimizing either all the network parameters or normalization parameters only is considered, resulting in at least two different setups for each method. Further, when applicable (the methods compute a segmentation loss based on masks, as opposed to another self-supervised loss such as the prediction entropy), the cross-entropy and IoU losses are compared. This results in four setups for the Ref, PL and AugCo methods. From learning hyper-parameters, the learning rate and number of TTA iterations are considered. The maximum possible number of iterations is 10 to limit the computational requirements. Reasonable learning rate values are first found via a grid search and then extended with other promising values based on the initial results.

Shared implementation details. The refinement network architecture is a U-Net Ronneberger et al. (2015) with an EfficentNet-B0 Tan & Le (2019) backbone pre-trained on ImageNet from the Timm library Wightman (2019). It is trained with the AdamW Loshchilov & Hutter (2017) optimizer with a learning rate of $1e^{-3}$ and the Cross-Entropy (CE) loss. The SGD optimizer is used for the TTA since early experiments with AdamW showed high divergence rate.

	Ent		PL				Ref				AugCo				Adv		dIoU	
params loss	full ent	norm ent	full ce	full iou	norm ce	norm iou	full ce	full iou	norm ce	norm iou	full ce	full iou	norm ce	norm iou	full kl	norm kl	full -	norm -
NA	35.18	35.18	35.18	35.18	35.18	35.18	35.18	35.18	35.18	35.18	35.18	35.18	35.18	35.18	35.20	35.20	35.18	35.18
TTA_{α^*}	35.18	35.58	35.54	37.21	35.60	37.09	35.18	**38.69**	36.88	36.50	35.27	35.66	35.35	35.39	35.20	35.20	35.18	35.18
Δ_{ABS}	$-\epsilon$	0.39	0.36	2.03	0.42	1.90	$-\epsilon$	3.51	1.70	1.32	0.09	0.48	0.17	0.21	$-\epsilon$	$-\epsilon$	$-\epsilon$	$-\epsilon$

Table 1: $\mathrm{m}\overline{\mathrm{IoU}}_i$ results aggregated across corruptions and levels in the GTA5-C dataset, compared to non-adapted (NA) performance. The TTA hyper-parameters α^* were selected for overall best performance of each method. The **overall** and per-method best results are highlighted. No positive hyper-parameters are denoted by $-\epsilon$ (the performance converges to 0 from below).

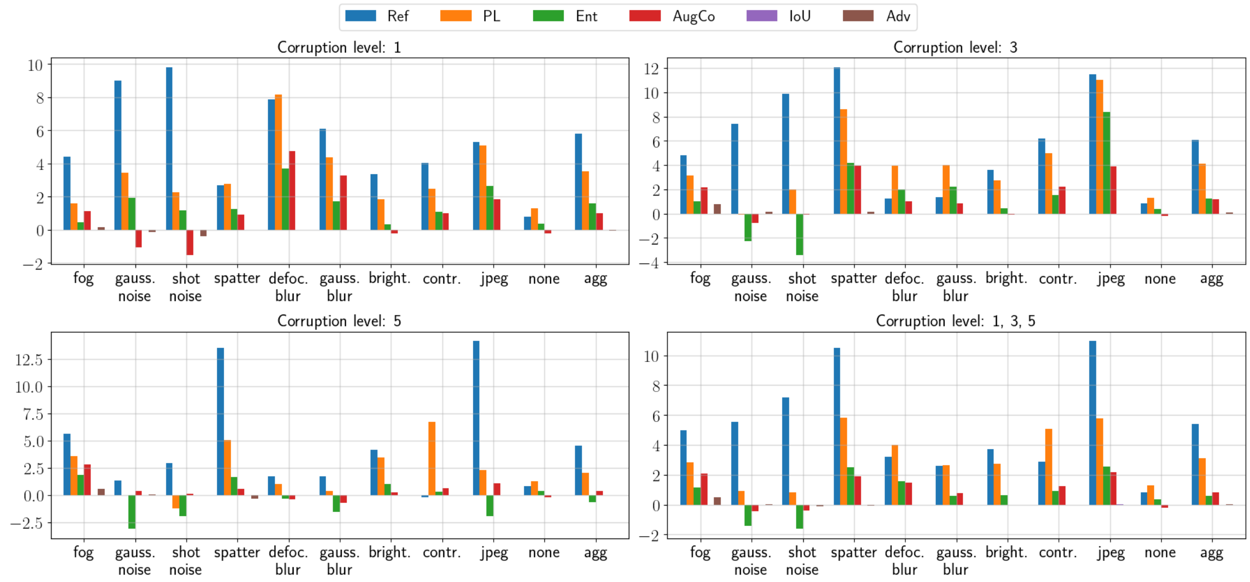
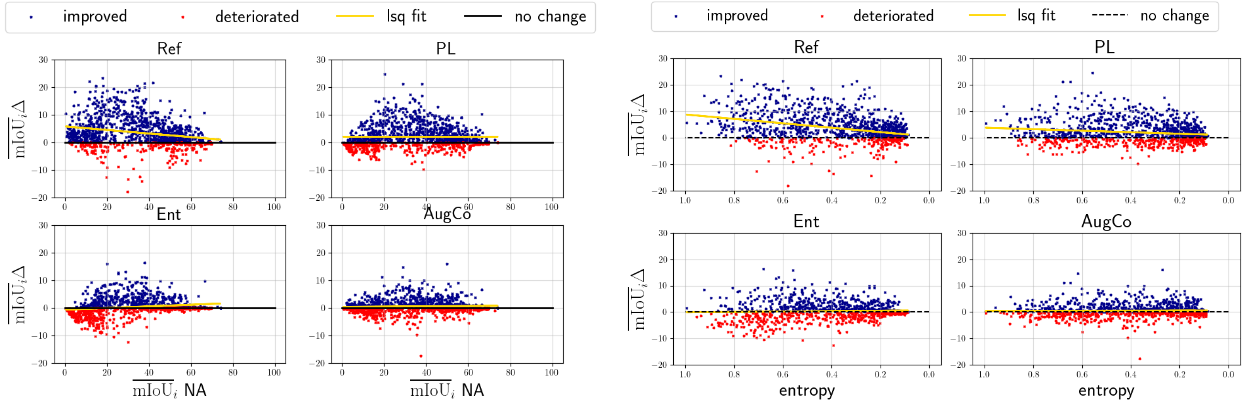


Figure 3: GTA5-C $\mathrm{m}\overline{\mathrm{IoU}}_i$ error reduction (%) depending on corruption levels. TTA with overall optimal hyper-parameters for GTA5-C.



(a) $\mathrm{m}\overline{\mathrm{IoU}}_i$ NA

(b) entropy NA

Figure 4: The relationship between per-image scores (a) or entropy (b) before and the score after adaptation on the GTA5-C dataset. The difference between non-adapted (NA) $\mathrm{m}\overline{\mathrm{IoU}}_i$ or entropy and the $\mathrm{m}\overline{\mathrm{IoU}}_i$ after TTA is shown ($\mathrm{m}\overline{\mathrm{IoU}}_i\Delta$). A least-squares line fitted to the points is shown in yellow.

5.3 Experiment results

GTA5 → Cityscapes, ACDC. This experiments explores the performance of the TTA methods on a model trained on a synthetic driving dataset, GTA5, evaluated on real-world driving dataset in clean weather conditions, as well as under adverse weather conditions. The GTA5-pretrained model is the best-performing model of Volpi et al. (2022) (DeepLabV2).

Since current methods do not consider different hyper-parameters for individual images, a single set of hyper-parameters with overall best performance across all corruptions and corruption levels is selected. The aggregated results with these overall optimal hyper-parameters on the SITTA training set can be found in Table 1. It can be observed that large improvements are achieved either by PL with IoU loss, optimizing normalization parameters only, or by Ref with IoU loss, optimizing all the parameters. The

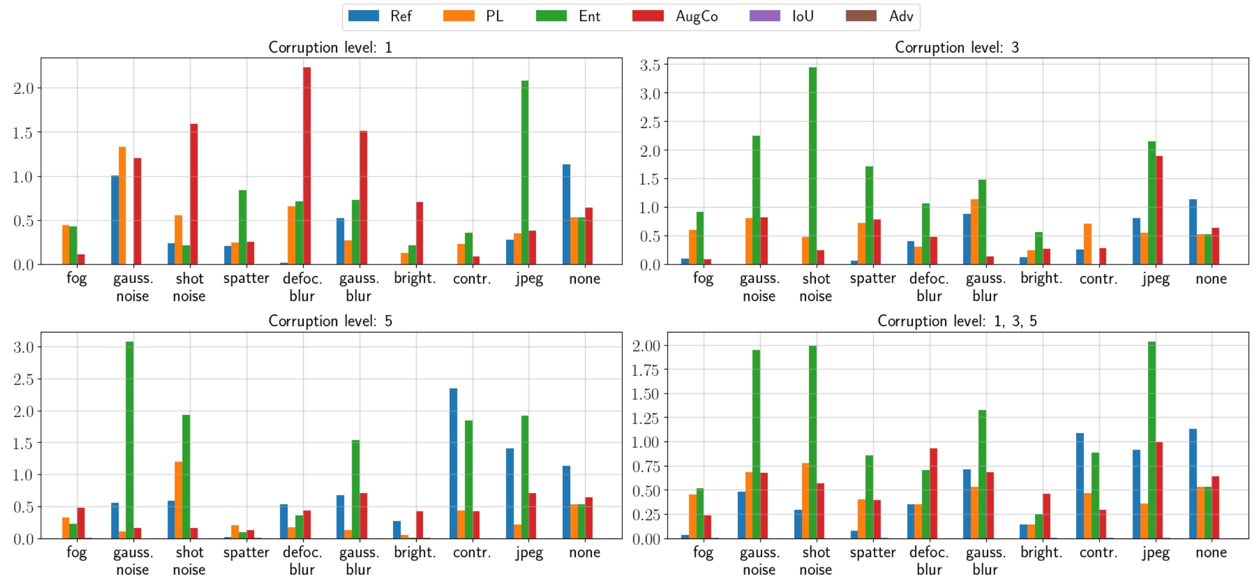


Figure 5: GTA5-C error reduction difference (%) between overall optimal hyperparameters and hyper-parameters selected for each corruption kind separately. The hyper-parameters were selected on GTA5-C.

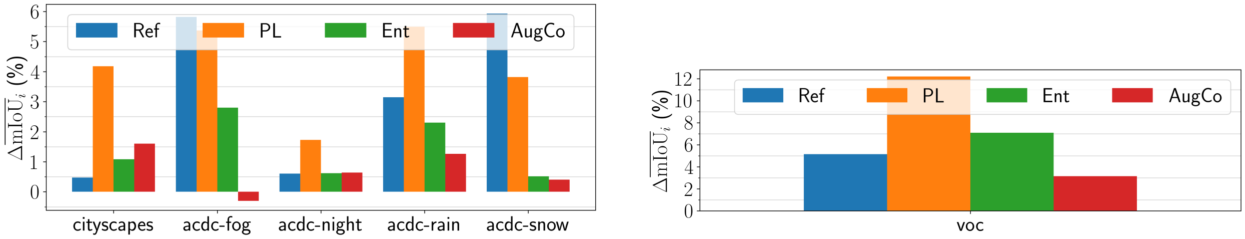
best-performing method is Ref, improving by 3.51 % over the non-adapted baseline (NA), reducing the total segmentation error by 5.41 %. Other methods only marginally improve over NA or show no improvement at all. Optimizing CE generally yields worse results than optimizing the IoU. While updating normalization parameters may only stabilize Ent, optimizing all the parameters for optimal performance of Ref is important. For other methods, the difference is small, yet optimizing normalization parameters only is faster and thus recommended.

In Figure 3, the total error reduction results with the same set of overall optimal hyper-parameters for each method are shown but for each corruption level and kind separately. It can be seen that it is not possible to find a single set of hyper-parameters that would perform well across all the corruption levels with these methods. While all methods improve performance on level 1 corruptions, from level 3, negative results can be observed for some of the corruptions, and all methods yield negative results on level 5. Ref outperforms the other methods on the majority of corruption kinds and corruption levels. The aggregated results across all corruptions showed that the negative results for level 5 and mixed results for level 3 are mostly outweighed by the gains on level 1, resulting in overall positive results.

In figure 5, it is shown that if one could select optimal hyper-parameters for each corruption kind and level, results would improve substantially. Moreover, Ref significantly outperforms the other methods on most corruptions, the blur corruption kinds being a notable exception. Significant improvements can be observed compared to other methods, especially on different kinds of noise and jpeg corruption at level 5. This analysis suggests that unless it is known in advance what kind of corruption or corruption levels will be present after deployment, strategies on method and hyper-parameter selection for each image should be considered.

Only the methods with overall positive TTA results are used for further analysis, namely Ref, PL, AugCo and Ent. The relationship between the non-adapted (NA) performance and the performance improvement on individual images for different methods is visualized in Subfigure 4a. The analysis shows Ref outperforms other methods, especially on images that had low initial $\mathrm{m}\overline{\mathrm{IoU}}_i$, while the performance of PL is more consistent across all initial scores but not as powerful for initial low scores. While Ent makes performance worse for low initial scores and improves more as the initial score increases, AugCo shows consistent improvements across all initial scores similarly to PL but to a smaller extent.

If the $\mathrm{m}\overline{\mathrm{IoU}}_i$ for each image were known, it could be used to either select a method performing best on those values or to select hyper-parameters. In Subfigure 4b, an analogous analysis is performed, replacing the



(a) GTA5 model

(b) COCO model

Figure 6: TTA $\text{m}\overline{\text{IoU}}_i$ error reduction (%) the driving test datasets with hyper-parameters selected on GTA5-C (left) and the VOC test dataset with hyper-parameters selected on COCO-C (right).

$\text{m}\overline{\text{IoU}}_i$ with segmentation prediction entropy. Similar results as with the $\text{m}\overline{\text{IoU}}_i$ can be observed, showing that the segmentation prediction entropy is a good proxy for initial segmentation $\text{m}\overline{\text{IoU}}_i$.

After selecting the best hyper-parameters for each method on the SITTA training set, the methods are evaluated on 5 test datasets: ACDC-Rain, ACDC-Fog, ACDC-Night, ACDC-Snow and Cityscapes. The Cityscapes represents a domain shift from synthetic to real images while ACDC datasets add adverse weather conditions, making the domain hsift even greater. The first four datasets are created by splitting the ACDC dataset by different conditions. Each of the test sets consists of 500 images. The test results are reported in Figure 6a. Similarly to SITTA training datasets, the Ref and PL methods perform best across all datasets. While not outperforming Ref on all the datasets, the performance of PL is consistently better than the other methods while Ref is outperformed or matched by the other methods on Cityscapes and ACDC-night.

The inconsistencies of results between SITTA training and test suggest that unless the domain shift conditions are known in advance, it is difficult to select hyper-parameters based on a general SITTA training set.

COCO → VOC. In this experiment, the performance of TTA methods is studied on a model trained on the COCO dataset and evaluated on the VOC dataset. The segmentation model is an official Torchvision DeepLabV3 model with a Resnet50 backbone trained on the COCO dataset with a subset of 20 VOC classes. In contrast to previous experiments, it is a real-to-real dataset domain shift. The results of different methods with parameters selected for the overall best performance across all corruptions and levels can be found in Table 2. Major improvements are obtained by the PL and Ref methods. PL outperforms Ref, in contrast to GTA5-C experiments. The best improvement is by 3.28 %, reducing the total segmentation error by 7.3 %. Again, best results for each method are always achieved by the IoU loss, outperforming CE in all cases. In contrast to GTA5-C, Ent achieves better results when optimizing all the network parameters. The same holds for PL. For Ref, optimizing all the parameters is again important. For other methods improvements over the non-adapted baseline are marginal.

	Ent		PL				Ref				AugCo				Adv		dIoU	
params	full	norm	full	full	norm	norm	full	full	norm	norm	full	full	norm	norm	full	norm	full	norm
loss	ent	ent	ce	iou	ce	iou	ce	iou	ce	iou	ce	iou	ce	iou	kl	kl	-	-
NA	55.01	55.01	55.01	55.01	55.01	55.01	55.01	55.01	55.01	55.01	55.01	55.01	55.01	55.01	55.16	55.16	55.01	55.01
TTA_{θ^*}	56.97	56.75	57.17	57.99	57.10	**58.30**	56.24	57.31	56.56	57.16	55.40	55.59	55.30	56.30	55.16	55.16	55.61	55.74
Δ_{ABS}	1.96	1.74	2.16	2.98	2.09	3.28	1.23	2.30	1.55	2.15	0.39	0.58	0.29	1.29	$-\epsilon$	$-\epsilon$	0.60	0.73

Table 2: $\text{m}\overline{\text{IoU}}_i$ results aggregated across corruptions and levels in the COCO-C dataset, compared to non-adapted (NA) performance. The TTA hyper-parameters α^* were selected for overall best performance of each method. The **overall** and per-method best results are highlighted. No positive hyper-parameters are denoted by $-\epsilon$ (the performance converges to 0 from below).

The total error reduction results with a single set of optimal hyper-parameters for each method are reported for each corruption level and kind in Figure 7. The results slightly differ from those for the GTA5-C, as

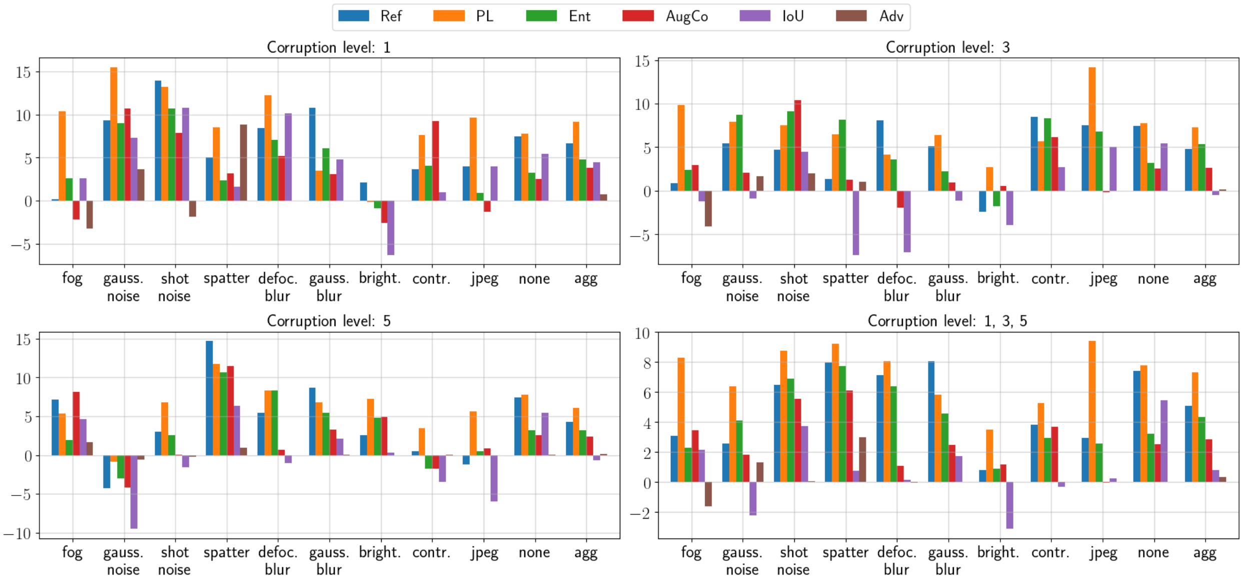
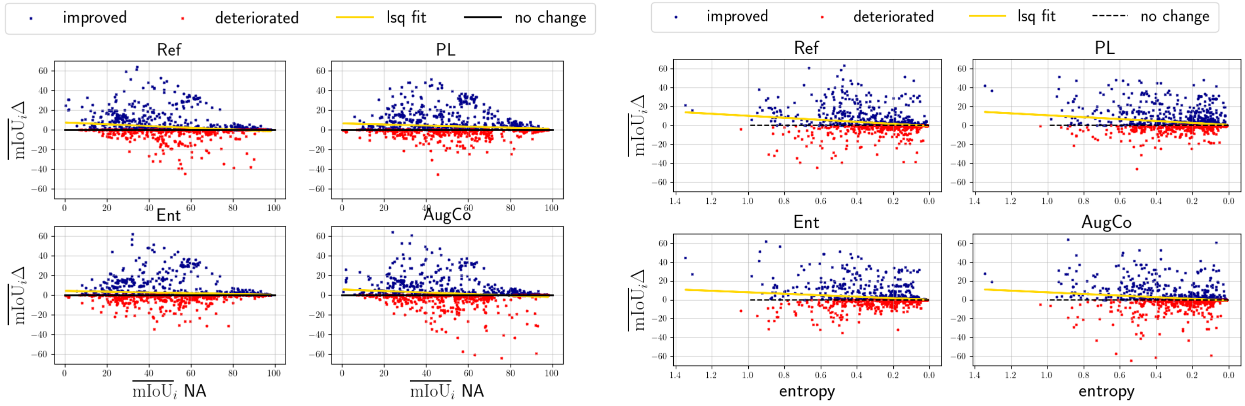


Figure 7: COCO-C $\mathrm{m}\overline{\mathrm{IoU}}_i$ error reduction (%) depending on corruption levels. TTA with overall optimal hyper-parameters for COCO-C.



(a) $\mathrm{m}\overline{\mathrm{IoU}}_i$ NA

(b) entropy NA

Figure 8: The relationship between per-image scores (a) or entropy (b) before and the score after adaptation on the COCO-C dataset. The difference between non-adapted (NA) $\mathrm{m}\overline{\mathrm{IoU}}_i$ or entropy and the $\mathrm{m}\overline{\mathrm{IoU}}_i$ after TTA is shown ($\mathrm{m}\overline{\mathrm{IoU}}_i\Delta$). A least-squares line fitted to the points is shown in yellow.

in this case, PL outperforms Ref. Both methods are again consistently better than the other methods, but positive results are reported for most corruptions on both level 1 and level 3.

In Figure 9, the results with optimal hyper-parameters for each method, corruption kind, and level are shown. This time, the results between different methods are much smaller. The PL consistently outperforms all other methods at all the corruption levels. Interestingly, the dIoU methods shows much stronger performance than in the GTA5-C experiments.

Figure 10b shows a comparison of the overall method performance on the SITTA training set. An oracle option is included where the method with best results is picked on per-image basis. There is a significant gap between the oracle and other methods, which further highlights that different methods are good in different cases and understanding the strengths of each methods can lead to greatly improved performance.

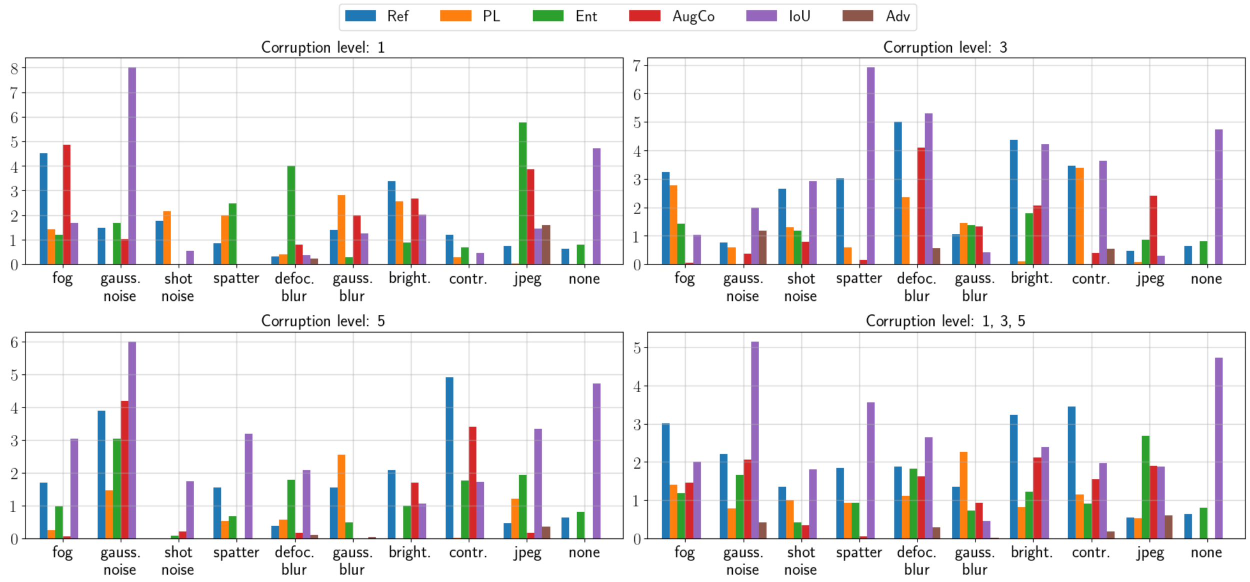
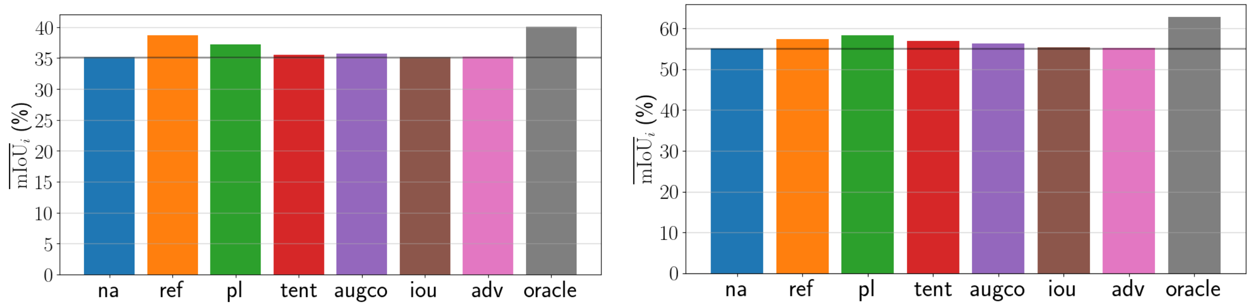


Figure 9: COCO-C error reduction difference (%) between overall optimal hyperparameters and hyper-parameters selected for each corruption kind separately. The hyper-parameters were selected on the COCO-C.



(a) Comparison of the overall performance of different methods on the GTA5-C validation set.

(b) Comparison of the overall performance of different methods on the COCO-C validation set.

Figure 10: The difference between non-adapted (NA) $\text{m}\overline{\text{IoU}}_i$.

Again, only Ref, PL, AugCo and Ent are used for further analysis. The relationship between the non-adapted (NA) performance and the performance improvement on individual images for different methods is visualized in Subgigure 8a. The distribution of initial non-adapted $\text{m}\overline{\text{IoU}}_i$ is different. The initial model is stronger than the GTA5 model. All methods show similar behavior - more improvement is achieved on images with a lower initial score, Ref and PL significantly outperform the other methods and again, Ref show better performance on images with low initial scores compared to other methods.

The relationship of segmentation prediction entropy and $\text{m}\overline{\text{IoU}}_i$ improvement by adaptation is shown in Subfigure 8b, supporting the notion that the entropy of prediction before adaptation is a good proxy for $\text{m}\overline{\text{IoU}}_i$.

The results on the VOC test set are shown in Figure 6b. PL slightly outperforms the other methods, but all the methods improved over the non-adapted baseline.

Additional results can be found in Appendix E.

6 Conclusions and limitations

This work investigates the performance of Single Image Test-Time Adaptation (SITTA) on segmentation problems. The first part explores the effect of previously neglected design choices. The results on a synthetic validation set reveal that while SITTA in the standard setting with CE loss does not improve performance much, substituting the CE with IoU improves performance substantially. The experiments on whether to update all or normalization parameters only are inconclusive; the results depends on the settings. Further, we find that entropy minimization, often reported as unstable for small batch sizes, performs well when the batch-normalization mean and variance are not updated at test-time.

Experiments were carried out on the GTA5 and COCO-pretrained models. In the GTA5-C synthetic datasets experiments, the refinement SITTA dominates, followed by pseudo-labelling. While the refinement is significantly better on some of the real-world test datasets, on other ones, pseudo-labelling performs best. In the COCO-C experiments, the top performers swap places: Pseudo-labelling is followed by refinement. On the test dataset, pseudo-labelling remains the best.

There are many common patterns in the GTA5 and COCO model experiments, but the TTA performance still depends on the model and the dataset. There is not a single method performing best in all cases. This motivated the oracle experiments which show that the results would improve substantially if the best method was chosen for each image. Evaluation with overall-optimal parameters and parameters found for each synthetic domain shift separately shows different kind of images benefit from different hyper-parameters. This diversity in performance underscores the necessity of a context-aware selection of adaptation techniques and hyper-parameters, based on the specific characteristics of the deployment domain.

We evaluate how performance depends on the difficulty of the segmentation task. First, we show that refinement performs well on images with low initial segmentation score in the GTA5 experiments. Next, we show that the initial score can be replaced by the prediction entropy which does not require labels. In the COCO experiments, the results are inconclusive, methods generally improve more on high-entropy images.

Limitations. First of all, to limit the scope of the study, we only focused on adaptation with self-supervised loss functions and no reliance on batch-normalization layers. While these methods tend to perform the best, their iterative optimization comes at an increased computational time. Methods alleviating this burden should be explored, such as only adapting to informative samples or methods inspired by efficient model finetuning. Secondly, the synthetic validation set created by applying artificial corruption to the training set images can only emulate covariate shift. The presence of label shift may contribute to some of the discrepancies between the validation and test results. Finally, only two models were considered and the effect of different model architectures on the individual methods is not known. While our work improves the understanding of TTA for semantic segmentation methods better, a benchmark for fair and thorough evaluation of the methods is still missing.

Broader Impact Statement

Maybe we could discuss a bit the 'Test-Time Poisoning Attacks Against Test-Time Adaptation Models' work and similar.

References

N Romero Aquino, Matheus Gutoski, Leandro T Hattori, and Heitor S Lopes. The effect of data augmentation on the performance of convolutional neural networks. *Braz. Soc. Comput. Intell*, 2017.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Alexander Bartler, Andre Bühler, Felix Wiewel, Mario Döbler, and Bin Yang. Mt3: Meta test-time training for self-supervised test-time adaption. In *International Conference on Artificial Intelligence and Statistics*, pp. 3080–3090. PMLR, 2022.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 295–305, 2022.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

Alain Fournier, Don Fussell, and Loren Carpenter. Computer rendering of stochastic models. *Communications of the ACM*, 25(6):371–384, 1982.

Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A Efros. Test-time training with masked autoencoders. *arXiv preprint arXiv:2209.07522*, 2022.

Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven adaptation to test-time corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11786–11796, 2023.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456, 2015a.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015b.

Neerav Karani, Ertunc Erdil, Krishna Chaitanya, and Ender Konukoglu. Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68:101907, 2021.

Ansh Khurana, Sujoy Paul, Piyush Rai, Soma Biswas, and Gaurav Aggarwal. SITA: Single Image Test-time Adaptation. 12 2021. URL https://arxiv.org/abs/2112.02355v3.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896, 2013.

Bo Li, Yezhen Wang, Tong Che, Shanghang Zhang, Sicheng Zhao, Pengfei Xu, Wei Zhou, Yoshua Bengio, and Kurt Keutzer. Rethinking distributional matching based domain adaptation. *arXiv preprint arXiv:2006.13352*, 2020.

Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1215–1224, 2021a.

Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34:21808–21820, 2021b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Saypraseuth Mounsaveng, Florent Chiaroni, Malik Boudiaf, Marco Pedersoli, and Ismail Ben Ayed. Bag of tricks for fully test-time adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1936–1945, 2024.

Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

A Tuan Nguyen, Thanh Nguyen-Tang, Ser-Nam Lim, and Philip H S Torr. TIPI: Test Time Adaptation with Transformation Invariance. a.

A Tuan Nguyen, Thanh Nguyen-Tang, Ser-Nam Lim, and Philip HS Torr. Tipi: Test time adaptation with transformation invariance. b.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023.

Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Augco: augmentation consistency-guided self-training for source-free domain adaptive semantic segmentation. *arXiv preprint arXiv:2107.10140*, 2021.

Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 102–118. Springer, 2016.

Adrian Lopez Rodriguez and Krystian Mikolajczyk. Domain adaptation for object detection via style consistency. *arXiv preprint arXiv:1911.10033*, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3234–3243, 2016.

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8050–8058, 2019.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10765–10775, 2021.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33:11539–11551, 2020.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Korawat Tanwisuth, Xinjie Fan, Huangjie Zheng, Shujian Zhang, Hao Zhang, Bo Chen, and Mingyuan Zhou. A prototype-oriented framework for unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 34:17194–17208, 2021.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

Gabriele Valvano, Andrea Leo, and Sotirios A Tsaftaris. Re-using adversarial mask discriminators for test-time training under distribution shifts. *arXiv preprint arXiv:2108.11926*, 2021.

Riccardo Volpi, Pau De Jorge, Diane Larlus, and Gabriela Csurka. On the road to online adaptation for semantic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19184–19195, 2022.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020a.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020b.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

Renhao Wang, Yu Sun, Yossi Gandelsman, Xinlei Chen, Alexei A Efros, and Xiaolong Wang. Test-Time Training on Video Streams. URL `https://video-ttt.github.io/`.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Chang'an Yi, Haotian Chen, and Yifan Zhang. A critical look at classic test-time adaptation methods in semantic segmentation. *arXiv preprint arXiv:2310.05341*, 2023.

Yongcan Yu, Lijun Sheng, Ran He, and Jian Liang. Benchmarking test-time adaptation against distribution shifts in image classification. *arXiv preprint arXiv:2307.03133*, 2023.

Yue Zhang, Shun Miao, Tommaso Mansi, and Rui Liao. Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, pp. 599–607. Springer, 2018.