AugCal: Improving Sim2Real Adaptation by Uncertainty Calibration on Augmented Synthetic Images

Prithvijit Chattopadhyay* **Bharat Goyal** **Boglarka Ecsedi** **Viraj Prabhu** **Judy Hoffman**
Georgia Tech
{prithvijit3,bharatgoyal,becsedi3,virajp,judy}@gatech.edu

Abstract

Synthetic data (Sim) drawn from simulators have emerged as a popular alternative for training models where acquiring annotated real-world images is difficult. However, transferring models trained on synthetic images to real-world applications can be challenging due to appearance disparities. A commonly employed solution to counter this Sim2Real gap is unsupervised domain adaptation, where models are trained using labeled Sim data and unlabeled Real data. Mispredictions made by such Sim2Real adapted models are often associated with miscalibration – stemming from overconfident predictions on real data. In this paper, we introduce AugCal, a simple training-time patch for unsupervised adaptation that improves Sim2Real adapted models by – (1) reducing overall miscalibration, (2) reducing overconfidence in incorrect predictions and (3) improving confidence score reliability by better guiding misclassification detection – all while retaining or improving Sim2Real performance. Given a base Sim2Real adaptation algorithm, at training time, AugCal involves replacing vanilla Sim images with strongly augmented views (Aug intervention) and additionally optimizing for a training time calibration loss on augmented Sim predictions (Cal intervention). We motivate AugCal using a brief analytical justification of how to reduce miscalibration on unlabeled Real data. Through our experiments, we empirically show the efficacy of AugCal across multiple adaptation methods, backbones, tasks and shifts.

1 Introduction

Most effective models for computer vision tasks (classification, segmentation, *etc*.) need to learn from a large amount of exemplar data (Dosovitskiy et al., 2020; Radford et al., 2021; Kirillov et al., 2023; Pinto et al., 2008) that captures real-world natural variations which may occur at deployment time. However, collecting and annotating such diverse real-world data can be prohibitively expensive – for instance, densely annotating a frame of Cityscapes (Cordts et al., 2016) can take upto ~ 1.5 hours! Machine-labeled synthetic images generated from off-the-shelf simulators can substantially reduce this need for manual annotation and physical data collection (Sankaranarayanan et al., 2018; Ros et al., 2016; Blaga & Nedevschi, 2019; Savva et al., 2019; Deitke et al., 2020; Chattopadhyay et al., 2021). Nonetheless, models trained on Sim data often exhibit subpar performance on Real data, primarily due to appearance discrepancies, commonly referred to as the Sim2Real gap. For instance, on GTAV (Sim) $\rightarrow$ Cityscapes (Real), an HRDA Sim-only model (Hoyer et al., 2022b) achieves an mIoU of only 53.01, compared to ~ 81 mIoU attained by an equivalent model trained exclusively on Real data.

While there is significant effort in improving the realism of simulators (Savva et al., 2019; Richter et al., 2022), there is an equally large effort seeking to narrow this Sim2Real performance gap by designing algorithms that facilitate Sim2Real transfer. These methods encompass both *generalization* (Chattopadhyay* et al., 2023; Huang et al., 2021; Zhao et al., 2022). – aiming to ensure strong out-of-the-box Real performance of Sim trained models – and *adaptation* (Hoyer et al., 2022b;c; Vu et al., 2019; Rangwani et al., 2022) – attempting to adapt models using labeled Sim data and unlabeled Real data. Such generalization and adaptation methods have demonstrated notable success in reducing the Sim2Real performance gap. For instance, Pasta (Chattopadhyay* et al., 2023) (a *generalization* method) improves Sim2Real performance of a Sim-only model from

*Correspondence to PC

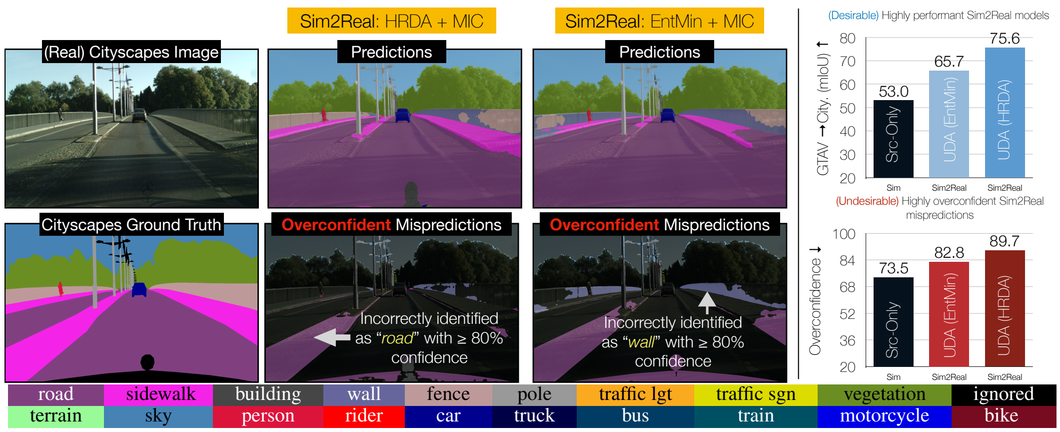

Figure 1: **Overconfident SIM2REAL mispredictions. [Left]** We show an example of what we mean by overconfident mispredictions. For SIM2REAL adaptation on GTAV→Cityscapes, we choose (DAFormer) HRDA + MIC (Hoyer et al., 2022c) and EntMin + MIC (Vu et al., 2019) (highly performant SIM2REAL methods) and show erroneous predictions on Cityscapes (bottom row). We can see that the model identifies *sidewalk* pixels as *road* (2nd column) and *fence* pixels as *wall* (3rd column) with very high confidence. **[Right]** We show how pervasive this "overconfidence" phenomena is. While better SIM2REAL adapted models – from (DAFormer) Source-Only (Hoyer et al., 2022b) to (DAFormer) EntMin + MIC (Vu et al., 2019) to (DAFormer) HRDA + MIC (Hoyer et al., 2022c) – exhibit improved transfer performance **[Top, Right]**, they also exhibit increased overconfidence in mispredictions **[Bottom, Right]**, affecting prediction reliability.

$53.01 \rightarrow 57.21$ mIoU. Furthermore, HRDA + MIC (Hoyer et al., 2022c) (an *adaptation* approach) pushes performance even higher to 75.56 mIoU.

While SIM2REAL performance may increase both from generalization or adaptation methods, for safety-critical deployment scenarios, task performance is often not the sole factor of interest. It is additionally important to ensure SIM2REAL adapted models make *calibrated* and *reliable* predictions on REAL data. Optimal calibration on real data ensures that the model's confidence in its predictions aligns with the true likelihood of correctness. Deploying poorly calibrated models can have severe consequences, especially in high-stakes applications (such as autonomous driving), where users can place trust in (potentially) unreliable predictions (Tesla Crash, 2016; Michelmore et al., 2018). We find that mistakes made by SIM2REAL adaptation methods are often associated with miscalibration caused by *overconfidence* – highly confident incorrect predictions (see Fig. 1 Left). More interestingly, we find that as adaptation methods improve in terms SIM2REAL performance, the propensity to make overconfident mispredictions also increases (see Fig. 1 Right). Our focus in this paper is to devise training time solutions to mitigate this issue.

Calibrating deep neural networks (for such SIM2REAL adaptation methods) is crucial, as they routinely make overconfident predictions (Guo et al., 2017; Gawlikowski et al., 2021; Minderer et al., 2021). While various techniques address miscalibration on "labeled data splits" for in-distribution scenarios, maintaining calibration in the face of dataset shifts, like SIM2REAL, proves challenging due to lack of labeled examples in the target (REAL) domain. To address this, we propose AUGCAL, a training-time patch to ensure existing SIM2REAL adaptation methods make *accurate*, *calibrated* and *reliable* predictions on real data. When applied to a SIM2REAL adaptation framework, AUGCAL aims to satisfy three key criteria: (1) retain performance of the base SIM2REAL method, (2) reduce miscalibration and overconfidence and (3) ensure calibrated confidence scores translate to improved reliability. Additionally, to ensure broad applicability, AUGCAL aims to do so by making two minimally invasive changes to a SIM2REAL adaptation training pipeline. First, by AUGmenting (Cubuk et al., 2020; Chattopadhyay* et al., 2023) input SIM images during training using an AUG transform that reduces distributional distance between SIM and REAL images. Second, by additionally optimizing for a CALibration loss (Hebbalaguppe et al., 2022; Liang et al., 2020a; Liu et al., 2022) at training time on AUGmented SIM predictions. We devise AUGCAL based on an analytical rationale (see Sec. 3.2.1 and 3.2.2) illustrating how it helps reduce an upper bound on desired target (REAL) calibration error. Through our experiments on GTAV→Cityscapes and VisDA SIM2REAL, we demonstrate how AUGCAL helps reduce miscalibration on REAL data. To summarize, we make the following contributions:

- We propose AUGCAL, a training time patch, compatible with existing SIM2REAL adaptation methods that ensures SIM2REAL adapted models make *accurate* (measured via adaptation performance), *calibrated* (measured via calibration error) and *reliable* (measured via confidence guided misclassification detection) predictions.

- We conduct SIM2REAL adaptation experiments for object recognition (VisDA (Peng et al., 2017)) and semantic segmentation (GTAV (Sankaranarayanan et al., 2018)→Cityscapes (Cordts et al., 2016)) with three representative UDA methods (pseudo-label based self-training, entropy minimization and domain adversarial training) and show that applying AUGCAL– (1) improves or preserves adaptation performance, (2) reduces miscalibration and overconfidence and (3) improves the reliability of confidence scores.
- We show how AUGCAL improvements are effective across multiple backbones, AUG and CAL options and highlight choices that are more consistently effective across experimental settings.

2 RELATED WORK

Unsupervised Domain Adaptation (UDA). We focus on UDA algorithms to address *covariate shifts* in the SIM2REAL context (Chattopadhyay* et al., 2023; Choi et al., 2021; Zhao et al., 2022; Huang et al., 2021; Rangwani et al., 2022; Hoyer et al., 2022c; Sankaranarayanan et al., 2018; Ros et al., 2016). This involves adapting a model to an unseen target (REAL) domain using labeled samples from a source (SIM) domain and unlabeled samples from the target domain. Here, the source and target datasets share the same label space and labeling functions, but differences exist in the distribution of inputs (Farahani et al., 2021; Zhang et al., 2019). SIM2REAL UDA methods (Ganin & Lempitsky, 2014; Hoffman et al., 2018; Saenko et al., 2010; Tzeng et al., 2014) range from *feature distribution matching* (Ganin & Lempitsky, 2014; Long et al., 2018; Saito et al., 2018; Tzeng et al., 2017; Zhang et al., 2019), explicitly addressing *domain discrepancy* (Kang et al., 2019; Long et al., 2015; Tzeng et al., 2014; Rangwani et al., 2022), *entropy minimization* (Vu et al., 2019) or *pseudo-label guided self-training* (Hoyer et al., 2022b;a;c). We observe that existing SIM2REAL UDA methods usually improve performance at the expense of increasingly overconfident mispredictions on (REAL) target data (Wang et al., 2020b). Our proposed method, AUGCAL, is designed to retain SIM2REAL adaptation performance while reducing miscalibration on real data for existing methods. We conduct experiments on three representative UDA methods – Entropy Minimization (Vu et al., 2019), Self-training (Hoyer et al., 2022b) and Domain Adversarial Training (Rangwani et al., 2022).

Confidence Calibration for Deep Networks. For discriminative models, confidence calibration indicates the degree to which confidence scores associated with predictions align with the true likelihood of correctness (usually measured via ECE (Naeini et al., 2015)). Deep networks tend to be be very poor at providing calibrated confidence estimates (are overconfident) for their predictions (Guo et al., 2017; Gawlikowski et al., 2021; Minderer et al., 2021), which in turn leads to less reliable predictions for decision-making in safety-critical settings. Recent work (Guo et al., 2017) has also shown that calibration worsens for larger models and can decrease with increasing performance. Several works (Guo et al., 2017; Lakshminarayanan et al., 2017; Malinin & Gales, 2018) have explored this problem for modern architectures, and several solutions have also been proposed –including temperature scaling (prediction logits being divided by a scalar learned on a held-out set (Platt et al., 1999; Kull et al., 2017; Bohdal et al., 2021; Islam et al., 2021)) and trainable calibration objectives (training time loss functions that factor in calibration (Liang et al., 2020a; Karandikar et al., 2021)). Improving network calibration is even more challenging in out-of-distribution settings due to the simultaneous lack of ground truth labels and overconfidence on unseen samples (Wang et al., 2020b). Specifically, instead of methods that rely on temperature-scaling (Wang et al., 2020a; 2022) or maybe require an additional calibration split, AUGCAL explores the use of training time calibration objectives (Munir et al., 2022) to reduce micalibration for SIM2REAL shifts.

3 METHOD

3.1 BACKGROUND

Notations. Let x denote input images and y denote corresponding labels (from the label space $\mathcal{Y} = \{1, 2, ..., K\}$) drawn from a joint distribution $P(x, y)$. We focus on the classification case, where the goal is to learn a discriminative model $\mathcal{M}_\theta$ (with parameters θ) that maps input images to the desired K output labels, $\mathcal{M}_\theta : \mathcal{X} \to \mathcal{Y}$, using a softmax layer on top. The predictive probabilities for the given input can be expressed as $p_\theta(y|x) = \text{softmax}(\mathcal{M}_\theta(x))$. We use $\hat{y} = \arg\max_{y\in\mathcal{Y}} p_\theta(y|x)$ to denote the predicted label for x and c to denote the confidence in prediction.

Unsupervised SIM2REAL Adaptation. In unsupervised domain adaptation (UDA) for SIM2REAL settings, we assume access to a labeled (SIM) source dataset $D_S = \{(x_i^S, y_i^S)\}_{i=1}^{|S|}$ and an unlabeled (REAL) target dataset $D_T = \{x_i^T\}_{i=1}^{|T|}$. We assume D_S and D_T splits are drawn from source and target distributions $P^S(x, y)$ and $P^T(x, y)$ respectively. At training, we have access to $D = D_S \cup D_T$. We operate in the setting where source and target share the same label space, and discrepancies exist only in input images. The model $\mathcal{M}_\theta$ is trained on labeled source images using cross entropy,

$$\sum_{i=1}^{|S|} \mathcal{L}_{CE}(x_i^S, y_i^S; \theta) = -\sum_{i=1}^{|S|} y_i^S \log p_\theta(\hat{y}_i^S | x_i^S) \text{ where } \hat{y}_i^S = \arg\max_{y \in \mathcal{Y}} p_\theta(y_i^S | x_i^S) \tag{1}$$

UDA methods additionally optimize for an adaptation objective on labeled source and unlabeled target data ($\mathcal{L}_{UDA}$). The overall learning objective can be expressed as,

$$\min_\theta \underbrace{\sum_{i=1}^{|S|} \mathcal{L}_{CE}(x_i^S, y_i^S; \theta)}_{\text{Source Loss}} + \underbrace{\sum_{i=1}^{|T|} \sum_{j=1}^{|S|} \lambda_{UDA} \mathcal{L}_{UDA}(x_i^T, x_j^S, y_j^S; \theta)}_{\text{Source Target Adaptation Loss}} \tag{2}$$

Different adaptation methods usually differ in terms of specific instantiations of this objective. While AugCal is applicable to any Sim2Real adaptation method in principle, we conduct experiments with three popular methods – Entropy Minimization (Vu et al., 2019), Pseudo-Label driven Self-training (Hoyer et al., 2022b) (for semantic segmentation) and Domain Adversarial Training (Rangwani et al., 2022) (for object recognition). We provide more details on these methods in Sec. G of appendix.

Uncertainty Calibration. For a perfectly calibrated classifier, the confidence in predictions should match the empirical frequency of correctness. Empirically, calibration can be measured using Expected Calibration Error (ECE) (Naeini et al., 2015). To measure ECE on a test set $D = \{(x_i, y_i)\}_{i=1}^{|D|}$, we first partition the test data into B bins, $D_b = \{(x, y) \mid r_{b-1} \leq c < r_b\}$, using the confidence values c such that $b \in \{1, ..., B\}$ and $0 = r_0 \leq r_1 \leq r_2 \leq \cdots \leq r_B = 1$. Then, ECE measures the absolute differences between accuracy and confidence across instances in every bin,

$$\text{ECE} = \sum_{j=1}^{B} \frac{B}{|D|} \left| \frac{1}{B} \sum_{i \in D_j} \mathbf{1}_{(y_i = \hat{y}_i)} - \sum_{i \in D_j} c_i \right| \tag{3}$$

We are interested in models that exhibit high-performance and low calibration error (ECE). Note that Eqn 3 alone does not indicate if a model is overconfident. We define overconfidence (OC) as the expected confidence on mispredictions. Prior work on improving calibration in out-of-distribution (OOD) settings (Wang et al., 2022) and domain adaptation scenarios (Wang et al., 2020b) typically rely on techniques like temperature scaling. These methods often necessitate additional steps, such as employing a separate calibration split or domains (Gong et al., 2021) or training extra models (e.g., logistic discriminators for source and target features (Wang et al., 2020b)). In contrast, we consider using training time calibration objectives (Liang et al., 2020b; Hebbalaguppe et al., 2022; Liu et al., 2022) that can be optimized in addition to task-specific objectives for improved calibration.

3.2 AugCal

3.2.1 Reducing Miscalibration on (Real) Target

Recall that $P^S(x, y)$ and $P^T(x, y)$ denote the source (Sim) and target (Real) data distributions. We assume $P(x, y)$ factorizes as $P(x, y) = P(x)P(y|x)$. We assume covariate shift conditions between P^S and P^T, *i.e.*, $P^T(x) \neq P^S(x)$ while $P^T(y|x) = P^S(y|x)$ – discrepancies across distributions exist only in input images. When training a model, we can only draw "labeled samples" (x, y) from $P^S(x, y)$. We do not have access to labels from $P^T(x, y)$. Our goal is to reduce miscalibration on (unlabeled) target data using training time calibration losses. Let $\mathcal{L}_{\text{CAL}}(x, y)$ denote such a calibration loss we can minimize (on labeled data). Using importance sampling (Cortes et al., 2010), we can get an estimate of the desired calibration loss on target data as,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x,y \sim P^T(x,y)} [\mathcal{L}_{\text{CAL}}(x, y)] &= \int_x \int_y \mathcal{L}_{\text{CAL}}(x, y) P^T(x, y) \, dx \, dy \\
&= \int_x \int_y \mathcal{L}_{\text{CAL}}(x, y) \frac{P^T(x) P^T(y|x)}{P^S(x) P^S(y|x)} P^S(x, y) \, dx \, dy \\
&= \mathop{\mathbb{E}}_{x,y \sim P^S(x,y)} \left[\underbrace{w_S(x)}_{\text{Importance Weight}} \underbrace{\mathcal{L}_{\text{CAL}}(x, y)}_{\text{Source Loss}} \right]
\end{aligned} \tag{4}$$

where $w_S(x) = \frac{P^T(x)}{P^S(x)}$ denotes the importance weight. Assuming $\mathcal{L}_{\text{CAL}}(x, y) \geq 0$[1], we can obtain an upper bound on step 4 (Pampari & Ermon, 2020; Wang et al., 2020b) as

$$\underset{x,y\sim P^T(x,y)}{\mathbb{E}}[\mathcal{L}_{\text{CAL}}(x, y)] = \underset{x,y\sim P^S(x,y)}{\mathbb{E}}[w_S(x)\mathcal{L}_{\text{CAL}}(x, y)] \tag{5}$$

$$\leq \sqrt{\underset{P^S(x)}{\mathbb{E}}\left[w_S(x)^2\right]\underset{P^S(x,y)}{\mathbb{E}}\left[\mathcal{L}_{\text{CAL}}(x, y)^2\right]} \tag{6}$$

$$\leq \frac{1}{2}\Bigg(\underbrace{\underset{P^S(x)}{\mathbb{E}}\left[w_S(x)^2\right]}_{\text{Shift Dependent}} + \underbrace{\underset{P^S(x,y)}{\mathbb{E}}\left[\mathcal{L}_{\text{CAL}}(x, y)^2\right]}_{\text{Source Dependent}}\Bigg) \tag{7}$$

where steps 6 and 7 use the Cauchy-Schwarz and AM-GM inequalities respectively. For a given model, the second RHS term in inequation 7 is computed purely on labeled samples from the source distribution and can therefore be optimized to convergence over the course of training. The gap in $\mathcal{L}_{\text{CAL}}$ across source and target is dominated by the importance weight (first term). Following (Cortes et al., 2010), the first term can also be expressed as,

$$\underset{P^S}{\mathbb{E}}\left[w_S(x)^2\right] = d_2\big(P^T(x)||P^S(x)\big) \tag{8}$$

where $d_\alpha(P||Q) = \left[\sum_x \frac{P^\alpha(x)}{Q^{\alpha-1}(x)}\right]^{\frac{1}{\alpha-1}}$ with $\alpha > 0$ is the exponential in base 2 of the Renyi-divergence (Rényi, 1960) between distributions P and Q. The calibration error gap between source and target distributions is therefore, dominated by the divergence between source and target distributions. Consequently, inequation 7 can be expressed as,

$$\underbrace{\underset{x,y\sim P^T(x,y)}{\mathbb{E}}[\mathcal{L}_{\text{CAL}}(x, y)]}_{\text{Target Calibration Loss}} \leq \underbrace{\frac{1}{2}d_2\big(P^T(x)||P^S(x)\big)}_{\text{Source and Target Divergence}} + \underbrace{\frac{1}{2}\underset{P^S(x,y)}{\mathbb{E}}\left[\mathcal{L}_{\text{CAL}}(x, y)^2\right]}_{\text{Source Calibration Loss}} = \underbrace{U(S, T)}_{\text{Upper Bound}} \tag{9}$$

where $U(S, T)$ denotes the upper bound on target calibration loss. Therefore, to effectively reduce miscalibration on target data, one needs to reduce the upper bound, $U(S, T)$, which translates to (1) reducing miscalibration on source data (second red term in 9) and (2) reducing the distributional distance between input distributions across source and target (first blue term in 9).

3.2.2 WHY AUGCAL?

Based on the previous discussion, to improve calibration on target data, one can always invoke a training time calibration intervention (CAL) on labeled source data to reduce $\mathbb{E}_{P^S(x,y)}[\mathcal{L}_{\text{CAL}}(x, y)]$. In practice, after training, we can safely assume that $\mathbb{E}_{P^S(x,y)}[\mathcal{L}_{\text{CAL}}(x, y)] = \epsilon \to 0$ (for some very small ϵ). We note that while this is useful and necessary, it is not sufficient. This is precisely where we make our contribution. To reduce both (red and blue) terms in 9, we introduce AUGCAL. To do this, in addition to a training time calibration loss, $\mathcal{L}_{\text{CAL}}$, AUGCAL assumes that access to an additional AUG transformation that satisfies the following properties:

1. $d_2\big(P^T(x)||P^S(\text{AUG}(x))\big) \leq d_2\big(P^T(x)||P^S(x)\big)$
2. After training, $\mathbb{E}_{P^S(x,y)}\Big[\mathcal{L}_{\text{CAL}}(x, y)\Big] \approx \mathbb{E}_{P^S(x,y)}\Big[\mathcal{L}_{\text{CAL}}(\text{AUG}(x), y)\Big] = \epsilon \to 0$

Property 1 states that the chosen AUG transformation brings transformed source data closer to target (or reduces SIM2REAL distributional distance). Property 2 states that over the course of training, irrespective of the data $\mathcal{L}_{\text{CAL}}(x, y)$ is optimized on (AUG transformed or clean source), $\mathbb{E}_{P^S(x,y)}\Big[\mathcal{L}_{\text{CAL}}(\cdot, \cdot)\Big]$ can achieve a sufficiently small value close to 0. Given an AUG transformation that satisfies the above stated properties, we can claim,

$$U^{\text{AUG}}(S, T) \leq U(S, T) \tag{10}$$

where

$$U^{\text{AUG}}(S, T) = \frac{1}{2}d_2\big(P^T(x)||P^S(\text{AUG}(x))\big) + \frac{1}{2}\underset{P^S(x,y)}{\mathbb{E}}\left[\mathcal{L}_{\text{CAL}}(\text{AUG}(x), y)^2\right] \tag{11}$$

[1]We make the reasonable assumption that the calibration loss function is always non-negative.

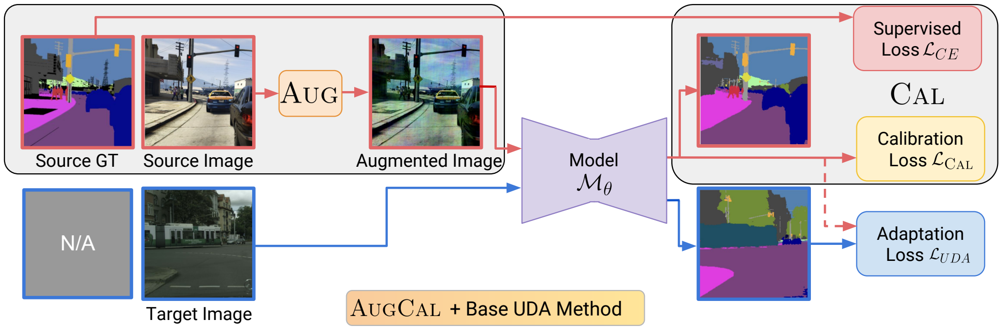

Figure 2: **AUGCAL pipeline**. AUGCAL consists of two key interventions on an existing SIM2REAL adaptation method. First source SIM images are augmented via an AUG transform. Supervised losses for SIM images are computed on the augmented image predictions. Additionally, AUGCAL optimizes for a calibration loss on AUGmented SIM predictions.

That is, an appropriate AUG transform, when coupled with $\mathcal{L}_{\text{CAL}}$, helps reduce a tighter upper bound on the target calibration error than CAL. We call this intervention – coupling AUG and CAL– AUGCAL. Naturally, the effectiveness of AUGCAL is directly dependent on the choice of AUG that satisfies the aforementioned properties.

Table 1: **Eligible AUG choices.** AUG transformations that reduce SIM2REAL distance and satisfy Property-1.

SIM2REAL	100x (RBF) MMD (SIM, REAL)	(PASTA-SIM, REAL)	(R.Aug-SIM, REAL)
VisDA	4.806	**4.171**	**3.464**
GTAV→City.	5.795	**5.214**	**5.213**

While most augmentations can satisfy property 2, to check if an augmentation is valid according to property 1, we compute RBF Kernel based MMD distances for (SIM, REAL) and (AUGmented SIM, REAL) feature pairs using a trained model.[2] In Table. 1, we show how PASTA (Chattopadhyay* et al., 2023) and RandAugment (Cubuk et al., 2020), two augmentations effective for SIM2REAL transfer, satisfy these considerations on multiple shifts (read Table. 1 left to right). PASTA and RandAug are additionally (1) inexpensive when combined with SIM2REAL UDA methods, and (2) generally beneficial for SIM2REAL shifts (PASTA via SIM2REAL specific design and RandAug via chained photometric operations).

3.2.3 AUGCAL INSTANTIATION

Given a SIM2REAL adaptation method, AUGCAL additionally optimizes for improved calibration on augmented SIM source images. Since AUGCAL is applicable to any existing SIM2REAL adaptation method, we abstract away the adaptation component associated with the pipeline and denote as $\mathcal{L}_{UDA}$ (see Eqn. 2). The steps involved in AUGCAL are illustrated in Fig. 2. Given a mini-batch, we first generate augmented views, $\text{AUG}(x^S)$, for SIM images x^S. Then, during training, we optimize $\mathcal{L}_{CE}$ on those augmented SIM views and $\mathcal{L}_{UDA}$ for adaptation. To improve calibration under augmentations, we optimize an additional $\mathcal{L}_{\text{CAL}}$ loss on augmented SIM images. The overall AUGCAL optimization problem can be expressed as,

$$\min_\theta \underbrace{\sum_{i=1}^{|S|} \mathcal{L}_{CE}(\text{AUG}(x_i^S), y_i^S; \theta)}_{\text{Source Task Loss}} + \underbrace{\sum_{i=1}^{|T|}\sum_{j=1}^{|S|} \lambda_{UDA}\mathcal{L}_{UDA}(x_i^T, \text{AUG}(x_j^S), y_j^S; \theta)}_{\text{Source Target Adaptation Loss}} + \underbrace{\sum_{i=1}^{|S|} \lambda_{\text{CAL}}\mathcal{L}_{\text{CAL}}(\text{AUG}(x_i^S), y_i^S; \theta)}_{\text{Source Calibration Loss}} \tag{12}$$

where λ_{UDA} and λ_{CAL} denote the respective loss coefficients and the changes to a vanilla SIM2REAL adaptation framework are denoted in teal.

Choice of "AUG". Strongly augmenting SIM images during training has proven useful for SIM2REAL transfer (Chattopadhyay* et al., 2023; Huang et al., 2021; Zhao et al., 2022; Cubuk et al., 2020). For AUGCAL, we are interested in augmentations that satisfy the properties outlined in Sec. 3.2.2. As stated earlier, we find that RandAugment (Cubuk et al., 2020) and PASTA (Chattopadhyay* et al., 2023) empirically satisfy this criteria. We use both of them for our experiments and provide details on the operations for these AUG transforms in Sec. B of appendix.

Choice of "CAL" ($\mathcal{L}_{\text{CAL}}$). AUGCAL relies on using training time calibration losses to reduce miscalibration. Prior work in uncertainty calibration has considered several auxiliary objectives to

[2]Under bounded importance weight assumptions, MMD can be interpreted as an upper bound on KL divergence (Wang & Tay, 2022).

calibrate a model being trained to reduce negative log-likelihood (NLL) (Hebbalaguppe et al., 2022; Kumar et al., 2018; Karandikar et al., 2021; Liang et al., 2020a). For "CAL" in AUGCAL, while we consider multiple calibration losses – DCA (Liang et al., 2020b), MbLS (Liu et al., 2022) and MDCA (Hebbalaguppe et al., 2022) – and find that DCA, simple "difference between confidence and accuracy (DCA)" loss proposed in (Liang et al., 2020a) is more consistently effective across experimental settings. DCA can be expressed as,

$$\mathcal{L}_{\text{CAL}} = \frac{1}{|S|}\left|\sum_{i=1}^{|S|} \mathbf{1}_{(y_i^S=\hat{y}_i^S)} - \sum_{i=1}^{|S|} p_\theta(\hat{y}_i^S|x_i^S)\right| \tag{13}$$

where $\mathbf{1}_{(y_i^S=\hat{y}_i^S)}$ and $p_\theta(\hat{y}_i^S|x_i^S)$ denote the correctness and confidence scores associated with predictions. The DCA loss forces the mean predicted confidence over training samples to match accuracy. In the following sections, we empirically validate AUGCAL across adaptation methods.

4 EXPERIMENTAL DETAILS

We conduct SIM2REAL adaptation experiments across two tasks – Semantic Segmentation (SemSeg) and Object Recognition (ObjRec). For our experiments, we train models using labeled SIM images and unlabeled REAL images. We test trained models on REAL images.

SIM2REAL Shifts. For SemSeg, we conduct experiments on the GTAV→Cityscapes shift. GTAV (Sankaranarayanan et al., 2018) consists of $\sim$ 25k densely annotated SIM ground-view images and Cityscapes (Cordts et al., 2016) consists of $\sim$ 5k REAL ground view images. We report all metrics on the Cityscapes validation split. For ObjRec, we conduct experiments on the VisDA SIM2REAL benchmark. VisDA (Peng et al., 2017) consists of $\sim$ 152k SIM images and $\sim$ 55k REAL images across 12 classes. We report all metrics on the validation split of (REAL) target images.

Models. We check AUGCAL compatibility with both CNN and Transformer based architectures. For SemSeg, we consider DeepLabv2 (Chen et al., 2017) (with a ResNet-101 (He et al., 2016) backbone) and DAFormer (Hoyer et al., 2022a) (with an MiT-B5 (Xie et al., 2021) backbone) architectures. For ObjRec, we consider ResNet-101 and ViT-B/16 (Dosovitskiy et al., 2020) backbones with bottleneck layers as classifiers. We start with backbones pre-trained on ImageNet (Deng et al., 2009).

Adaptation Methods. We consider three representative SIM2REAL adaptation methods for our experiments. For SemSeg, we consider entropy minimization (EntMin) (Vu et al., 2019) and high-resolution domain adaptive semantic segmentation (HRDA) (Hoyer et al., 2022b). For ObjRec, we consider smooth domain adversarial training (SDAT) (Rangwani et al., 2022). For both tasks, we further improve performance with masked image consistency (MIC) (Hoyer et al., 2022c) on target images during training. We use MIC (Hoyer et al., 2022c)'s implementations of the adaptation algorithms and provide more training details in Sec. C of appendix.

Calibration Metrics. We use ECE to report overall confidence calibration on REAL images. Since we are interested in reducing overconfident mispredictions, we also report calibration error on incorrect samples (IC-ECE) (Wang et al., 2022) and mean overconfidence for mispredictions (OC).

Reliability Metrics. While reducing overconfidence and improving calibration on real data is desirable, this is a proxy for the true goal of improving model reliability. To assess reliability, following prior work (de Jorge et al., 2023; Malinin et al., 2019), we measure whether calibrated confidence scores can better guide misclassification detection. To measure this, we use Prediction Rejection Ratio (PRR) (Malinin et al., 2019), which if high (positive and close to 100) indicates that confidence scores can be used as reliable indicators of performance (details in Sec. D of appendix).

Unless specified otherwise, we use PASTA as the choice of AUG and DCA (Liang et al., 2020b) as the choice of CAL in AUGCAL. We use $\lambda_{\text{CAL}} = 1$ for DCA.

5 FINDINGS

5.1 IMPROVING SIM2REAL ADAPTATION

Recall that when applied to a SIM2REAL adaptation method, we expect AUGCAL to – (1) retain SIM2REAL transfer performance, (2) reduce miscalibration and overconfidence and (3) ensure calibrated confidence scores translate to improved model reliability. We first verify these criteria.

▷ **AUGCAL improves or retains SIM2REAL adaptation performance.** Since AUGCAL intervenes on an existing SIM2REAL adaptation algorithm, we first verify that encouraging better calibration does not adversely impact SIM2REAL adapation performance. We find that performance is either retained or improved (*e.g.*, for EntMin + MIC in Table. 2 (a)) as miscalibration is reduced (Tables. 2(a) and (b), Perf. columns).

Table 2: **AugCal ensures Sim2Real adapted models make accurate, calibrated and reliable predictions.** We find that applying AugCal to multiple Sim2Real adaptation methods across tasks leads to better calibration (ECE, IC-ECE), reduced overconfidence (OC) and improved reliability (PRR) – all while retaining or improving transfer performance. Highlighted rows are AugCal variants of the base methods. For AugCal, we use Pasta as Aug and DCA as Cal. ± indicates standard error.

Method	**Perf. (↑)** mIoU	**Calibration Error (↓)** ECE	IC-ECE	OC	**Reliability (↑)** PRR
1 EntMin + MIC	65.71	$5.34_{\pm 0.35}$	$77.73_{\pm 0.26}$	$82.83_{\pm 0.55}$	$45.93_{\pm 0.54}$
2 + AugCal	**70.31**	$\mathbf{3.43}_{\pm 0.29}$	$\mathbf{72.97}_{\pm 0.26}$	$\mathbf{82.80}_{\pm 0.57}$	$\mathbf{62.66}_{\pm 0.55}$
3 HRDA + MIC	75.56	$2.86_{\pm 0.10}$	$81.92_{\pm 0.14}$	$89.72_{\pm 0.48}$	$68.91_{\pm 0.46}$
4 + AugCal	**75.90**	$\mathbf{2.45}_{\pm 0.09}$	$\mathbf{79.09}_{\pm 0.16}$	$\mathbf{88.26}_{\pm 0.49}$	$\mathbf{70.35}_{\pm 0.51}$

(a) GTAV→Cityscapes. (DAFormer).

Method	**Perf. (↑)** mIoU	**Calibration Error (↓)** ECE	IC-ECE	OC	**Reliability (↑)** PRR
1 SDAT + MIC	$92.53_{\pm 0.28}$	$7.67_{\pm 0.49}$	$91.45_{\pm 0.63}$	$89.13_{\pm 1.29}$	$63.78_{\pm 2.12}$
2 + AugCal	$\mathbf{92.87}_{\pm 0.06}$	$\mathbf{6.84}_{\pm 0.10}$	$\mathbf{89.25}_{\pm 0.36}$	$\mathbf{85.74}_{\pm 0.36}$	$\mathbf{67.80}_{\pm 0.78}$

(b) VisDA Sim2Real. (ViT-B).

Table 3: **AugCal is better than applying Aug or Cal alone.** On GTAV→Cityscapes and VisDA, we show that AugCal improves over just augmented Sim training (Aug) or just optimizing for calibration on Sim data (Cal). For AugCal, we use Pasta as Aug and DCA as Cal. ± indicates standard error.

Method	**Perf. (↑)** mIoU	**Calibration Error (↓)** ECE	IC-ECE	**Reliability (↑)** PRR
1 EntMin + MIC	65.71	$5.34_{\pm 0.35}$	$77.73_{\pm 0.26}$	$45.93_{\pm 0.54}$
2 + Aug	67.58	$4.30_{\pm 0.33}$	$77.59_{\pm 0.25}$	$48.05_{\pm 0.53}$
3 + Cal	68.70	$4.04_{\pm 0.26}$	$75.86_{\pm 0.26}$	$52.52_{\pm 0.54}$
4 + AugCal	**70.31**	$\mathbf{3.43}_{\pm 0.29}$	$\mathbf{72.97}_{\pm 0.26}$	$\mathbf{62.66}_{\pm 0.55}$

(a) GTAV→Cityscapes. (DAFormer).

Method	**Perf. (↑)** mAcc	**Calibration Error (↓)** ECE	IC-ECE	**Reliability (↑)** PRR
1 SDAT + MIC	$92.53_{\pm 0.28}$	$7.67_{\pm 0.49}$	$91.45_{\pm 0.63}$	$63.78_{\pm 2.12}$
2 + Aug	$92.69_{\pm 0.15}$	$9.48_{\pm 1.99}$	$90.48_{\pm 0.33}$	$65.68_{\pm 0.58}$
3 + Cal	$91.63_{\pm 0.71}$	$7.30_{\pm 0.09}$	$91.19_{\pm 0.13}$	$66.62_{\pm 1.70}$
4 + AugCal	$\mathbf{92.87}_{\pm 0.06}$	$\mathbf{6.84}_{\pm 0.10}$	$\mathbf{89.25}_{\pm 0.36}$	$\mathbf{67.80}_{\pm 0.78}$

(b) VisDA Sim2Real (ViT-B).

▷ **AugCal reduces miscalibration post Sim2Real adaptation.** On both GTAV→Cityscapes and VisDA, we find that AugCal consistently reduces miscalibration of the base method by reducing overconfidence on incorrect predictions. This is evident in how AugCal variants of the base adaptation methods have lower ECE, IC-ECE and OC values (AugCal rows, Calibration columns in Tables. 2 (a) and (b)). As an example, to illustrate the effect of improved calibration on real data, in Fig. 3, we show how applying AugCal can improve the proportion of per-pixel SemSeg predictions that are accurate and have high-confidence (> 0.95).

▷ **AugCal improvements in calibration improve reliability.** As noted earlier, we additionally investigate the extent to which calibration improvements for Sim2Real adaptation translate to reliable confidence scores – via misclassification detection on Real target data (see Sec. 4), as measured by PRR. We find that AugCal consistently improves PRR of the base Sim2Real adaptation method (PRR columns for AugCal rows in Tables. 2(a) and (b)) – ensuring that predictions made AugCal variants of a base model are more trustworthy.

5.2 Analyzing AugCal

We now analyze different aspects of AugCal.

▷ **Applying AugCal is better than applying just Aug or Cal.** In Sec. 3.2.1 and 3.2.2, we discuss how AugCal can be more effective in reducing target miscalibration than just optimizing for improved calibration on labeled Sim images. We verify this empirically in Tables. 3(a) and (b) for SemSeg and ObjRec. We show that while Aug and Cal, when applied individually, improve over a base Sim2Real method, they fall short of improvements offered by AugCal.

▷ **AugCal is applicable across multiple Aug choices.** In Sec. 3.2.2 and Table. 1, we show how both Pasta (Chattopadhyay* et al., 2023) and RandAugment (Cubuk et al., 2020) are eligible for AugCal. In Table. 4(a), we fix DCA as Cal and find that both Pasta and RandAugment are effective in retaining or improving performance, reducing miscalibration and improving reliability.

▷ **Ablating Cal choices for AugCal.** For completeness, we also conduct experiments by fixing Pasta as Aug and ablating the choice of Cal in AugCal. We consider recently proposed training time calibration objectives – Difference of Confidence and Accuracy (DCA) (Liang et al., 2020b), Multi-class Difference in Confidence and Accuracy (MDCA) (Hebbalaguppe et al., 2022) and Margin-based Label Smoothing (MbLS) (Liu et al., 2022) – as potential Cal choices (results for SemSeg outlined in Table. 4(b)). We find that while MDCA and MbLS can be helpful, DCA is more consistently helpful across tasks and settings.

▷ **AugCal is applicable across multiple task backbones.** Different architectures – CNNs and Transformers – are known to exhibit bias towards different properties in images (shape, texture, *etc.*) (Naseer et al., 2021). Since the choice of Aug transform (which can alter such properties) is central to the efficacy of AugCal, we verify if AugCal is effective across both CNN and Transformer backbones. To do this, we conduct our SemSeg, ObjRec experiments with both transformer (DAFormer, ViT-B) and CNN (DeepLabv2-R101, ResNet-101) architectures. We find that AugCal

Table 4: **Ablating AUG and CAL choices in AUGCAL.** For a DAFormer model on GTAV→Cityscapes, AUGCAL successfully reduces miscalibration and produces reliable confidence scores for SIM2REAL adaptation using both PASTA (P) and RandAug (R) as AUG choices. We also ablate the choice of CAL in AUGCAL across DCA, MDCA and MbLS and find that DCA is more consistently effective in reducing miscalibration across tasks and settings. $\lambda_{\text{CAL}} = 1$ for MDCA and $\lambda_{\text{CAL}} = 0.1, m = 10$ for MbLS. $\pm$ indicates standard error.

Method	AUG	Perf. (↑) mIoU	Calibration Error (↓) ECE	IC-ECE	Reliability (↑) PRR
3 EntMin		65.71	$5.34_{\pm 0.35}$	$77.73_{\pm 0.26}$	$45.93_{\pm 0.54}$
4 + AUGCAL	P	70.31	$3.43_{\pm 0.29}$	$72.97_{\pm 0.26}$	$62.66_{\pm 0.55}$
4 + AUGCAL	R	70.65	$2.34_{\pm 0.14}$	$73.77_{\pm 0.21}$	$66.65_{\pm 0.46}$
7 HRDA		75.56	$2.86_{\pm 0.10}$	$81.92_{\pm 0.14}$	$68.91_{\pm 0.46}$
8 + AUGCAL	P	75.90	$2.45_{\pm 0.09}$	$79.09_{\pm 0.16}$	$70.35_{\pm 0.51}$
8 + AUGCAL	R	74.10	$2.77_{\pm 0.17}$	$77.94_{\pm 0.18}$	$69.46_{\pm 0.46}$

(a) Ablating AUG in AUGCAL. (CAL = DCA).

Method	CAL	Perf. (↑) mIoU	Calibration Error (↓) ECE	IC-ECE	Reliability (↑) PRR
3 EntMin + MIC		65.71	$5.34_{\pm 0.35}$	$77.73_{\pm 0.26}$	$45.93_{\pm 0.54}$
4 + AUGCAL	DCA	70.31	$3.43_{\pm 0.29}$	$72.97_{\pm 0.26}$	$62.66_{\pm 0.55}$
4 + AUGCAL	MDCA	69.50	$3.22_{\pm 0.26}$	$72.65_{\pm 0.25}$	$59.96_{\pm 0.51}$
4 + AUGCAL	MbLS	68.77	$2.90_{\pm 0.24}$	$72.53_{\pm 0.23}$	$61.57_{\pm 0.48}$
7 HRDA + MIC		75.56	$2.86_{\pm 0.10}$	$81.92_{\pm 0.14}$	$68.91_{\pm 0.46}$
8 + AUGCAL	DCA	75.90	$2.45_{\pm 0.09}$	$79.09_{\pm 0.16}$	$70.35_{\pm 0.51}$
8 + AUGCAL	MDCA	75.50	$2.92_{\pm 0.15}$	$80.76_{\pm 0.16}$	$68.45_{\pm 0.47}$
8 + AUGCAL	MbLS	71.15	$2.68_{\pm 0.15}$	$70.21_{\pm 0.42}$	$69.33_{\pm 0.47}$

(b) Ablating CAL in AUGCAL. (AUG = PASTA)

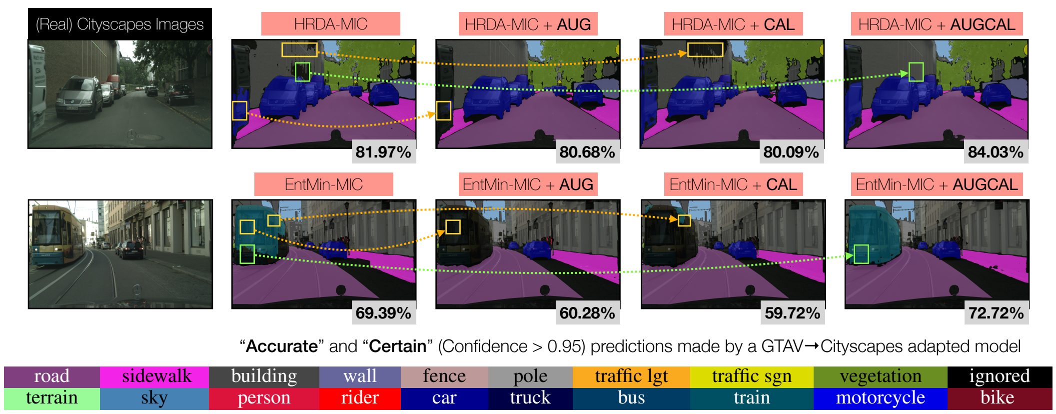

Figure 3: **AUGCAL increases the proportion of "accurate" and "certain" predictions**. For a (DAFormer) HRDA + MIC (row 1) and EntMin + MIC (row 2) on GTAV→Cityscapes, we show how different interventions affect the proportion of "accurate" and "certain" (confidence > 0.95) predictions (indicated in gray per column). Regions in black do not satisfy the "accurate" and "certain" filtering criteria. We see that compared to a base adaptation method, AUGCAL increases the proportion highly-confident correct predictions (green boxes). AUG and CAL applied alone can potentially reduce that proportion (yellow boxes). AUG is PASTA, CAL is DCA.

(with PASTA as AUG and DCA as CAL) is effective in reducing SIM2REAL miscalibration across all settings. We discuss these results in Sec. E of appendix.

▷ **How does AUGCAL compare with temperature scaling?** While we focus on "training-time" patches to improve SIM2REAL calibration, we also conduct an experiment to compare AUGCAL with "post-hoc" temperature scaling (TS) on VisDA. Specifically, we use 80% of VisDA SIM images for training models and rest (20%) for validation and temperature tuning. To ensure a fair comparison, we consider temperature tuning on both "clean" (C) and "PASTA augmented" (P) val splits. We find that irrespective of tuning on C or P, unlike AUGCAL, TS is ineffective and increases overconfidence and miscalibration. We present these results in Sec. E of appendix.

▷ **AUGCAL increases the proportion of accurate and certain predictions.** In Fig. 3, we show qualitatively for GTAV→Cityscapes SemSeg how AUGCAL increases the proportion of highly-confident correct predictions. In practice, we find that this improvement is much more subtle for stronger SIM2REAL adaptation methods, such as HRDA + MIC, compared to weaker ones, such as EntMin + MIC, which have considerable room for improvement.

6 CONCLUSION

We propose AUGCAL, a method to reduce the miscalibration of SIM2REAL adapted models, often caused due to highly-confident incorrect predictions. AUGCAL modifies a SIM2REAL adaptation framework by making two minimally invasive changes – (1) augmenting SIM images via AUG transformations that reduce SIM2REAL distance and (2) optimizing for an additional calibration loss on AUGmented SIM predictions. Applying AUGCAL to existing adaptation methods for semantic segmentation and object recognition reduces miscalibration, overconfidence and improves reliability of confidence scores, all while retaining or improving performance on REAL data. AUGCAL is meant to be a task-agnostic, general purpose framework to reduce miscalibration for SIM2REAL adaptation methods and we hope such simple methods are taken into consideration for experimental settings beyond the ones considered in this paper.

Acknowledgements. This work has been partially sponsored by NASA University Leadership Initiative (ULI) #80NSSC20M0161, ARL and NSF #2144194.

7 Reproducibility Statement

We provide training, implementation and optimization details associated with our experiments in Sec. C of the appendix. Most of our experiments follow the implementations of Sim2Real adaptation methods from (Hoyer et al., 2022c), with our additional modifications on top. Additionally, in Sec. H of appendix, we provide details surrounding the assets (and corresponding licenses) used for our experiments. Assumptions surrounding the analytical justification behind AugCal (Sec. 3.2.1 and 3.2.2) have been presented in the same sections.

8 Ethics Statement

Our proposed patch, AugCal, is meant to improve the reliability of Sim2Real adapted models. We assess reliability in terms of confidence calibration (prediction scores aligning with true likelihood of correctness) and the extent to which calibrated confidence scores are useful for assessing prediction quality (measured via mis-classification detection). AugCal adapted models have promising consequences for downstream applications. A well-calibrated and reliable Sim2Real adapted model can increase transparency in Real predictions and facilitate robust decision making in safety-critical scenarios. That said, we would like to note that while AugCal is helpful for our specific measures of reliability, exploration along other domain specific notions of reliability remain.

References

Bianca-Cerasela-Zelia Blaga and Sergiu Nedevschi. Semantic segmentation learning for autonomous uavs using simulators and real data. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 303–310, 2019. doi: 10.1109/ICCP48234.2019.8959563.

Ondrej Bohdal, Yongxin Yang, and Timothy Hospedales. Meta-calibration: Learning of model calibration using differentiable expected calibration error. *arXiv preprint arXiv:2106.09613*, 2021.

Prithvijit Chattopadhyay, Judy Hoffman, Roozbeh Mottaghi, and Aniruddha Kembhavi. Robustnav: Towards benchmarking robustness in embodied navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15691–15700, 2021.

Prithvijit Chattopadhyay*, Kartik Sarangmath*, Vivek Vijaykumar, and Judy Hoffman. Pasta: Proportional amplitude spectrum training augmentation for syn-to-real domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19288–19300, October 2023.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11580–11590, 2021.

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.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *NIPS*, 2010.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020.

Pau de Jorge, Riccardo Volpi, Philip Torr, and Gregory Rogez. Reliability in semantic segmentation: Are we on the right track? *arXiv preprint arXiv:2303.11298*, 2023.

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *CVPR*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

Yunye Gong, Xiao Lin, Yi Yao, Thomas G Dietterich, Ajay Divakaran, and Melinda Gervasio. Confidence calibration for domain generalization under covariate shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8958–8967, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ramya Hebbalaguppe, Jatin Prakash, Neelabh Madan, and Chetan Arora. 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*, pp. 16081–16090, 2022.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pp. 1989–1998, 2018.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9924–9935, 2022a.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pp. 372–391. Springer, 2022b.

Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. *arXiv preprint arXiv:2212.01322*, 2022c.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6891–6902, 2021.

Mobarakol Islam, Lalithkumar Seenivasan, Hongliang Ren, and Ben Glocker. Class-distribution-aware calibration for long-tailed visual recognition. *arXiv preprint arXiv:2109.05263*, 2021.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4893–4902, 2019.

Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C Mozer, and Becca Roelofs. Soft calibration objectives for neural networks. *Advances in Neural Information Processing Systems*, 34:29768–29779, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, pp. 623–631. PMLR, 2017.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pp. 2805–2814. PMLR, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

Gongbo Liang, Yu Zhang, and Nathan Jacobs. Neural network calibration for medical imaging classification using dca regularization. In *International Conference on Machine Learning (ICML)*, 2020a.

Gongbo Liang, Yu Zhang, Xiaoqin Wang, and Nathan Jacobs. Improved trainable calibration method for neural networks on medical imaging classification. *arXiv preprint arXiv:2009.04057*, 2020b.

Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. 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*, pp. 80–88, 2022.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Andrey Malinin and Mark J. F. Gales. Predictive uncertainty estimation via prior networks. In *NeurIPS*, 2018.

Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019.

Rhiannon Michelmore, Marta Kwiatkowska, and Yarin Gal. Evaluating uncertainty quantification in end-to-end autonomous driving control. *arXiv preprint arXiv:1811.06817*, 2018.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021.

Muhammad Akhtar Munir, Muhammad Haris Khan, M Sarfraz, and Mohsen Ali. Towards improving calibration in object detection under domain shift. *Advances in Neural Information Processing Systems*, 35:38706–38718, 2022.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

Anusri Pampari and Stefano Ermon. Unsupervised calibration under covariate shift. In *arXiv:2006.16405*, 2020.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Nicolas Pinto, David D Cox, and James J DiCarlo. Why is real-world visual object recognition hard? *PLoS computational biology*, 4(1):e27, 2008.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, Arihant Jain, and Venkatesh Babu Radhakrishnan. A closer look at smoothness in domain adversarial training. In *International Conference on Machine Learning*, pp. 18378–18399. PMLR, 2022.

Stephan R Richter, Hassan Abu AlHaija, and Vladlen Koltun. Enhancing photorealism enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1700–1715, 2022.

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.

Alfréd Rényi. On measures of information and entropy. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, 1960.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010.

Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. In *International Conference on Learning Representations*, 2018.

Swami Sankaranarayanan, Yogesh Balaji, Carlos D. Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *ICCV*, 2019.

Guardian Tesla Crash. Tesla driver dies in first fatal crash while using autopilot mode — theguardian.com. https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk, 2016. [Accessed 16-May-2023].

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

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.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2517–2526, 2019.

Bailin Wang, Mirella Lapata, and Ivan Titov. Meta-learning for domain generalization in semantic parsing. *arXiv preprint arXiv:2010.11988*, 2020a.

Chong Xiao Wang and Wee Peng Tay. Practical bounds of kullback-leibler divergence using maximum mean discrepancy. *arXiv preprint arXiv:2204.02031*, 2022.

Dongdong Wang, Boqing Gong, and Liqiang Wang. On calibrating semantic segmentation models: Analysis and an algorithm. *arXiv preprint arXiv:2212.12053*, 2022.

Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable calibration with lower bias and variance in domain adaptation. *Advances in Neural Information Processing Systems*, 33:19212–19223, 2020b.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pp. 7404–7413, 2019.

Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.