Supervision Interpolation via LossMix: Generalizing Mixup for Object Detection and Beyond

Thanh Vu[1,2*], Baochen Sun[2†], Bodi Yuan[2], Alex Ngai[2] Yueqi Li[2] , Jan-Michael Frahm[1]

[1]University of North Carolina at Chapel Hill
[2]Mineral
{thanhvu,baochens,bodiyuan}@mineral.ai, {alexander.s.ngai,yueqili.innovation}@gmail.com, jmf@cs.unc.edu

Abstract

The success of data mixing augmentations in image classification tasks has been well-received. However, these techniques cannot be readily applied to object detection due to challenges such as spatial misalignment, foreground/background distinction, and plurality of instances. To tackle these issues, we first introduce a novel conceptual framework called Supervision Interpolation (SI), which offers a fresh perspective on interpolation-based augmentations by relaxing and generalizing Mixup. Based on SI, we propose LossMix, a simple yet versatile and effective regularization that enhances the performance and robustness of object detectors and more. Our key insight is that we can effectively regularize the training on mixed data by interpolating their loss errors instead of ground truth labels. Empirical results on the PASCAL VOC and MS COCO datasets demonstrate that LossMix can consistently outperform state-of-the-art methods widely adopted for detection. Furthermore, by jointly leveraging LossMix with unsupervised domain adaptation, we successfully improve existing approaches and set a new state of the art for cross-domain object detection.

Introduction

Over the past decade, object detection has made remarkable progress, with impressive scores on challenging benchmarks such as MS COCO (Lin et al. 2014). However, state-of-the-art detectors still suffer from poor generalization abilities and struggle with data outside their training distribution, especially under domain shifts (Li et al. 2020; Oza et al. 2021). Recently, data mixing techniques, pioneered by Mixup (Zhang et al. 2018b), have emerged as an effective augmentation and regularization method for improving accuracy and robustness in deep neural networks. These techniques (Zhang et al. 2018b; Yun et al. 2019; Verma et al. 2019; Kim, Choo, and Song 2020; Dabouei et al. 2021; Hong, Choi, and Kim 2021) use a linear interpolation of both images and their labels to generate synthetic training data. The "mixing" process encourages the model to behave linearly between training examples, which can potentially reduce undesired oscillations for out-of-distribution predictions. Since its introduction in 2018, Mixup has garnered increasing attention and has been widely adopted for image classification problems (Zhang et al. 2018b; Yun et al. 2019; Verma et al. 2019; Xu et al. 2019; Kim, Choo, and Song 2020; Wu, Inkpen, and El-Roby 2020; Dabouei et al. 2021; Hong, Choi, and Kim 2021; Na et al. 2021; Liu et al. 2022a,b; Pinto et al. 2022; Liu et al. 2023). This motivates us to investigate Mixup-like augmentation for object detection.

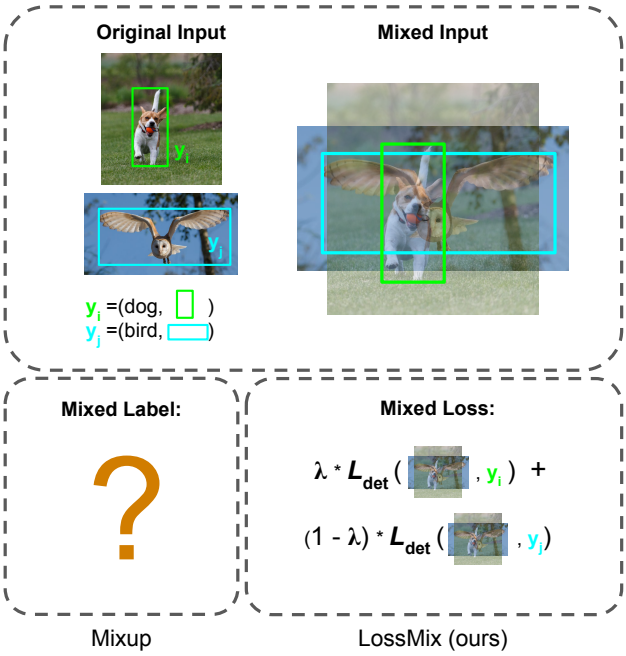

Figure 1: Compared to label-mixing (e.g., Mixup, CutMix), the proposed LossMix deploys interpolated losses instead of interpolated ground truths as the mixed supervision signals. This significantly simplifies the challenges involved in applying data mixing to higher-level tasks such as detection.

Unfortunately, Mixup cannot be applied to object detection task off the shelf. On one hand, the mixing of the category label of object instances is non-trivial (Fig. 2) due to issues such as spatial misalignment, foreground/background distinctions, and the plurality of instances. In contrast to classification, where images share the same shape and each

*Work done during a residency with Mineral.
†Project lead.

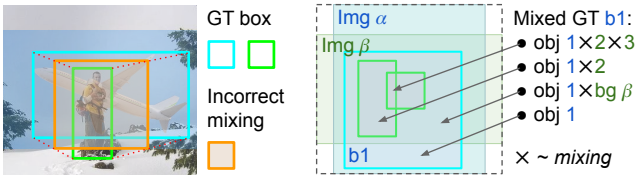

Figure 2: Challenges of applying Mixup to object detection. Left: Semantically incorrect interpolation of bounding box coordinates. Right: Complication of mixed category labels.

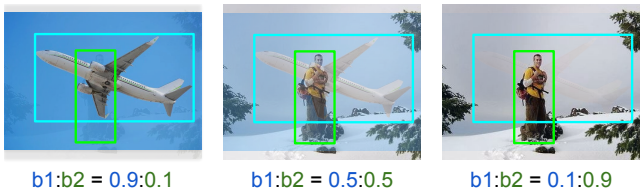

Figure 3: The state-of-the-art *Union* strategy is problematic because all ground truth bounding boxes are treated equally regardless of the mixing ratio.

only has one class label, object detectors need to handle images and objects of different aspect ratios and positions. This makes it impossible to guarantee the alignment of mixed objects and creates much more complexity for the interpolation of class labels. Fig. 2 (left) provides a visualization of these challenges. In addition, ground truth object detection annotations are composed of bounding box coordinates that cannot be naively interpolated without disturbing the localization ground truth (Fig. 2, right).

The current state-of-the-art approach (Zhang et al. 2019; Jocher 2020; Ge et al. 2021; Zhou et al. 2021; Wang et al. 2021; Zhang et al. 2022; Gao et al. 2022; Zheng et al. 2022; Yu et al. 2023; Zhang et al. 2023; Jocher, Chaurasia, and Qiu 2023) works around this by taking an unweighted, uniform *union* of all bounding boxes as new ground truth for the augmented image. Although this strategy has shown some success, there are several limitations. First, the approach does not follow the input-target dual interpolation principle that fuels the success of Mixup in classification, as it considers all component bounding boxes equally regardless of the actual mixing ratio λ. Second, when small mixing coefficients are used, e.g. $\lambda < 0.1$, the *Union* strategy can produce noisy mixed object labels (Fig. 3), potentially leading to sensitivity to noise and hallucinations in the model. These approaches expect the models to be able to predict all object instances with equal likelihood, regardless of their visibility. Finally, most of the previous studies have focused on using data mixing for semi-supervised (Zhou et al. 2021; Zheng et al. 2022) or few-shot learning (Gao et al. 2022), rather than general object detection, aside from (Zhang et al. 2019). More efforts exploring data mixing for general object detection are still needed to address these limitations.

To address these problems, we introduce two novel ideas: Supervision Interpolation (SI) and LossMix. SI generalizes Mixup's input-target formulation by relaxing the requirement to explicitly interpolate the labels. Instead, we hypothesize that it is possible to interpolate other forms of target supervision besides explicitly augmenting the ground truth. Based on this, we then propose LossMix, a simple but effective and versatile regularization that enables data mixing augmentation to strengthen object detection models and more. Our key insight is that we can effectively interpolate the losses, instead of the ground truth labels, according to the input's interpolation. Intuitively, from a data mixing perspective, LossMix interpolates the gradient signals that guide the models' learning, instead of explicitly augmenting the supervision labels like prior approaches (Zhang et al. 2018b; Yun et al. 2019). From an object detection perspective, LossMix weights the penalty for each prediction based on their augmented visibility. For example, LossMix would scale down the penalty for a failure to detect the plane with $\lambda = 0.1$ in Fig. 3 (right) since it has low visibility, while Union would treat both the plane and the person ($\lambda = 0.9$) equally. LossMix is flexible and can implicitly handle the mixing of sub-tasks (both classification and localization), while remaining true to the input-target dual interpolation principle that powered the success of the original Mixup (Zhang et al. 2018b). In short, our contributions are:

- We introduce Supervision Interpolation (SI), a conceptual reinterpretation and generalization of Mixup (Zhang et al. 2018b)-like input-label interpolation formulation.
- Based on SI, we propose LossMix, a simple but effective and versatile regularization that enables direct data mixing augmentation for object detection.
- We demonstrate that LossMix consistently outperforms state-of-the-art mixing methods for object detection on PASCAL VOC (2010) and MS COCO (2014) datasets.
- We leverage LossMix to enhance the recent Adaptive Teacher (Li et al. 2022b) framework and achieve a new state of the art for unsupervised domain adaptation.

Related Work

Data Mixing Augmentations The original Mixup (Zhang et al. 2018b) was designed for image classification tasks, proposing to use a convex combination of data and labels to expand the space of augmented training examples. Mixup, CutMix (Yun et al. 2019), and follow-up works (Verma et al. 2019; Xu et al. 2019; Kim, Choo, and Song 2020; Wu, Inkpen, and El-Roby 2020; Dabouei et al. 2021; Hong, Choi, and Kim 2021; Na et al. 2021; Chang, Tran, and Koishida 2021; Liu et al. 2022a,b; Pinto et al. 2022; Venkataramanan et al. 2022; Li et al. 2022a; Liu et al. 2023) have demonstrated the benefits of this interpolation-based augmentation for improving models' memorization and sensitivity to appearance changes. Since then, many works have utilized Mixup for image classification problems (Xu et al. 2019; Wu, Inkpen, and El-Roby 2020; Na et al. 2021). However, due to the challenges discussed above, only a few studies (Zhang et al. 2019; Jocher, Chaurasia, and Qiu 2023) have explored the use of Mixup for general object detection. Others tend to focus on settings with limited labeled data,

such as semi-supervised learning (Zhou et al. 2021; Zheng et al. 2022), few-shot learning (Gao et al. 2022), domain adaptation (Wang, Liao, and Shao 2021; Gao et al. 2022), or specific applications (Zhang et al. 2023; Yu et al. 2023).

Cross-Domain Object Detection There are two main approaches for object detection: one-stage object detectors that attempt to perform localization and classification simultaneously (Liu et al. 2016; Redmon et al. 2016), and two-stage object detectors that first generate object proposals and then perform classification and bounding box refinement in the second stage (Ren et al. 2015). Domain Adaptation based approaches aims to build a robust detectors that can generalize well to a target domain with limited or no labeled data. These methods can either explicitly align the feature distributions using a specific distance metric (Long et al. 2015, 2016; Sun, Feng, and Saenko 2016), or implicitly align the distributions using an adversarial loss (Ganin et al. 2016; Hoffman et al. 2018; Long et al. 2018) or GAN (Murez et al. 2018; Lee, Cho, and Im 2021). While most current works in domain adaptation focus on image classification (Abramov, Bayer, and Heller 2020; Lv et al. 2021; Ma et al. 2021; Meng et al. 2021; Berthelot et al. 2022; Harary et al. 2022; Hoyer, Dai, and Van Gool 2022; Liang et al. 2022; Liu, Durasov, and Fua 2022; Liu, Yang, and Hall 2022; Rangwani et al. 2022; Sun et al. 2022), a few have delved into object detection (Gu et al. 2019; Hsu et al. 2020; Deng et al. 2021; Munir et al. 2021; Ramamonjison et al. 2021; Li et al. 2022b). Data mixing is appealing in the context of UDA because of the opportunity to strategically blend cross-domain information during training. To our best knowledge, the topic of data mixing for cross-domain object detection remains largely understudied (Wang, Liao, and Shao 2021). In this work, we explore the application of LossMix to UDA.

Methodology

What does Mixup do? Mixup (Zhang et al. 2018b) trains models with virtual examples constructed by convex combinations of pairs of examples and their associated labels:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j \tag{1}$$

$$\tilde{y} = \lambda y_i + (1-\lambda)y_j \tag{2}$$

where (x_i, y_i) and (x_j, y_j) denote randomly sampled pairs of image and ground truth label and $\lambda \in [0, 1]$ denotes the interpolation coefficient. Training with Mixup-augmented data entails minimizing the empirical vicinal risk:

$$\mathcal{R}_v(f) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(f(\tilde{x}), \tilde{y}) \tag{3}$$

The idea is to regularize the learning using a prior knowledge that linear interpolations of the input features (input mixing) should yield linear interpolations of the corresponding output (label mixing). Such linear behavior in-between training examples can potentially mitigate undesired oscillations when predicting outside the training distribution.

Limitation of Label Mixing As discussed in the Sec. and illustrated in Fig. 2, despite working well for classification, the problem quickly arise when considering Mixup for other higher-level tasks like object detection. This is precisely because Eq. 2 imposes a hard constraint for the augmented labels $\tilde{y}$ to be an explicit linear combination of the real labels y_i and y_j. In object detection, for every x_i, the label y_i contains a set of object annotations, each has their own class and bounding box coordinates: $y_i = \{(c_{i1}, b_{i1}), (c_{i2}, b_{i2}), \ldots\}$, with (c_{ik}, b_{ik}) denoting an object instance with class label c_{ik} and box coordinates b_{ik}. This results in an ill-defined $\tilde{y}$ and makes label mixing exponentially more complicated. Existing work (Zhang et al. 2019; Wang et al. 2021; Zhou et al. 2021; Zheng et al. 2022; Gao et al. 2022; Zhang et al. 2022) chose to take an unweighted union of y_i and y_j, yielding $\tilde{y} = \{(c_{i1}, b_{i1}), (c_{i2}, b_{i2}), \ldots, (c_{j1}, b_{j1}), (c_{j2}, b_{j2}), \ldots\}$. Although this heuristic may offer some improvement in practice, it does not faithfully interpolate labels since $\tilde{y}$ is independent of λ and may lead to sub-optimal results. For example, small λ could be problematic since some objects become barely visible (Fig. 3), creating noisy labels.

Supervision Interpolation

We identify that the root cause of the aforementioned issues for both Mixup and unweighted union strategy is the label mixing requirement defined in Eq. 2. Despite working well for simple classification tasks, this policy clearly creates much complications for higher-level tasks such as object detection. To address this, we propose Supervision Interpolation (SI), a conceptual reinterpretation and generalization of Mixup's input-label interpolation formulation. In SI, we train models using a dual of interpolated data $\tilde{x}$ and proportionally interpolated supervision signals. Formally, SI trains models using convex combinations of examples and correspondingly augmented supervision signals:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j \tag{4}$$

$$\tilde{S} = \lambda S_i + (1-\lambda)S_j \tag{5}$$

where x denotes the training input, e.g., images, S denotes the supervision signal, and $\lambda \in [0, 1]$ denotes the mixing coefficient. Intuitively, SI regulates the training by interpolating the supervision or gradient signals guiding the model, depending on the interpolated inputs. For example, Mixup-based augmentations (Zhang et al. 2018b; Yun et al. 2019; Dabouei et al. 2021; Hong, Choi, and Kim 2021; Liu et al. 2022b) and ICT (Jeong et al. 2021; Verma et al. 2022) in semi-supervised learning can be seen as a special case of SI where the supervision signals are the ground truth classification labels. Based on such a flexible SI framework, next we will introduce, LossMix, a versatile method that enables data mixing and helps strengthen object detectors.

LossMix

Given the conceptual framework of Supervision Interpolation (SI), we then introduce LossMix, an equally simple but more versatile, task-agnostic sibling of Mixup that interpolates the loss errors instead of target labels. Specifically, the LossMix-augmented data:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j \tag{6}$$

$$\tilde{y} = \{(y_i; \lambda), (y_j; (1-\lambda)\} \tag{7}$$

are coupled with an augmented loss function:

$$\tilde{\mathcal{L}}(f(\tilde{x}), \tilde{y}) = \lambda \mathcal{L}(f(\tilde{x}), y_i) + (1-\lambda)\mathcal{L}(f(\tilde{x}), y_j) \tag{8}$$

From a SI perspective, our supervision signal is the loss errors weighted relative to y_i and y_j, instead of an explicit interpolation of y_i and y_j. From the Mixup perspective, we have relaxed the constraint in Eq. 2 and only characterise the virtual target $\tilde{y}$ using weighted y_i and y_j without explicitly constraint the form of $\tilde{y}$. Note that as $\lambda \to 0.0$ or 1.0, LossMix optimization will approach the standard empirical risk minimization. We would like to highlight that LossMix is not only simple and effective, but also highly versatile:

Simple-yet-effective The idea of loss mixing is straightforward and intuitive, both conceptually and implementation-wise, allowing easy adaptation to existing frameworks. Nonetheless, by design, LossMix helps circumvent the semantic collapse of bounding box interpolation (Fig. 2) and approximation issues of unweighted union approach (Fig. 3). Moreover, our experiments demonstrate the effectiveness and robustness of LossMix despite the simplicity in the design, successfully yielding improvement across two standard object detection datasets: PASCAL VOC (Everingham et al. 2010) and MS COCO (Lin et al. 2014). Finally, by applying LossMix to domain adaptation, we can further enhance state-of-the-art methods in cross-domain object detection (Li et al. 2022b).

Generalizability LossMix leverages a versatile loss weighting formulation that is potentially applicable to different tasks (i.e., classification, detection, etc.) as well as different input mixing strategies (i.e., Mixup (Zhang et al. 2018b), CutMix (Zhang et al. 2018b), etc.). LossMix is loss-agnostic, which simplifies its application in object detection and can inspire more applications beyond cross- entropy loss based classification, e.g. regression tasks like localization or depth estimation. The core idea of LossMix lies in the mixing of target loss signal y and does not limit the input mixing. This opens the door for different strategies, including pixel-based (Zhang et al. 2018b), region-based (Yun et al. 2019), style-based (Hong, Choi, and Kim 2021) and more.

LossMix in Action

Classification & Segmentation For standard classification and segmentation with Cross-Entropy loss, we can show that LossMix optimization is equivalent to that of Mixup:

$$\begin{aligned}
&\tilde{\mathcal{L}}_{lossmix}(f(\tilde{x}), \tilde{y}_{lossmix}) && (9)\\
&= \lambda \mathcal{L}_{ce}(f(\tilde{x}), y_i) + (1-\lambda)\mathcal{L}_{ce}(f(\tilde{x}), y_j) && (10)\\
&= -\lambda y_i \log f(\tilde{x}) - (1-\lambda) y_j \log f(\tilde{x}) && (11)\\
&= -(\lambda y_i + (1-\lambda) y_j) \log f(\tilde{x}) && (12)\\
&= \mathcal{L}_{ce}(f(\tilde{x}), \lambda y_i + (1-\lambda) y_j) && (13)\\
&= \tilde{\mathcal{L}}_{mixup}(f(\tilde{x}), \tilde{y}_{mixup}) && (14)
\end{aligned}$$

This means LossMix enjoys the same advantages as Mixup when applied to CE-based classification and segmentation problems (Zhang et al. 2018b; Yun et al. 2019; Xu et al. 2019; Verma et al. 2019; Wu, Inkpen, and El-Roby 2020; Kim, Choo, and Song 2020; Hong, Choi, and Kim 2021; Dabouei et al. 2021; Na et al. 2021; Pinto et al. 2022; Liu et al. 2022a,b; Li et al. 2022a; Liu et al. 2023). It is worth noting that loss-mixing formulation for Mixup by itself is not new (Zhang et al. 2018a; Liu et al. 2022b). What sets LossMix apart is its use of the Supervision Interpolation concept to replace label mixing with loss mixing at a fundamental level. This significantly enhances its generalizability, making CE-based classification a special case rather than the only option. Since the benefits of LossMix/Mixup for classification are well studied, we focus on exploring LossMix for object detection and domain adaptation.

Object Detection Since LossMix makes no assumption about the loss functions $\mathcal{L}$ in Eq. 8, we can easily apply it to object detection by interpolating both the classification loss and box regression loss. For example, for Faster RCNN (Ren et al. 2015), $\mathcal{L}$ takes the form of the standard supervised loss for two-stage detectors:

$$\begin{aligned}
\mathcal{L}_{det}(f(x), y) &= \mathcal{L}^{rpn}(f(x), y) + \mathcal{L}^{roi}(f(x), y)\\
&= \mathcal{L}^{rpn}_{cls}(f(x), y) + \mathcal{L}^{rpn}_{reg}(f(x), y)\\
&+ \mathcal{L}^{roi}_{cls}(f(x), y) + \mathcal{L}^{roi}_{reg}(f(x), y)
\end{aligned} \tag{15}$$

Here, $\mathcal{L}^{rpn}$ denotes the loss of Region Proposal Network (RPN) which generates candidate proposals, while $\mathcal{L}^{roi}$ denotes the loss for Region of Interest (ROI) branch. Both branches perform bounding box regression and classification tasks, specifically binary classification for RPN (object or not) and multi-class classification for ROI (Ren et al. 2015). Given a mixing coefficient λ, we can directly reweight all sub-task losses as follows (omitting f for brevity):

$$\tilde{\mathcal{L}}_{det}(\tilde{x}, \tilde{y}) = \lambda \mathcal{L}_{det}(\tilde{x}, y_i) + (1-\lambda)\mathcal{L}_{det}(\tilde{x}, y_j) \tag{16}$$

Domain Adaptation We apply LossMix alongside the recent Adaptive Teacher (Li et al. 2022b) (AT), a two-stage self-distillation method for cross-domain object detection. During the *Warmup* phase, we initialize both Teacher and Student models, who weights are shared, using standard object detection training with labeled source-domain data. We leverage LossMix to mix intra-source domain data to encourage better (non-directional) generalization and improve robustness on unseen data. However, initializing with pure source domain data risks biasing the Teacher model towards such a distribution (Deng et al. 2021), potentially yielding low-quality pseudo labels. Thus, we use unlabeled target images to mitigate this, specifically by mixing a small amount (e.g. $\lambda < 0.1$) of them into the labeled source images.

During the *Adaptation* phase, both models are jointly trained using the same cross-domain distillation (Li et al. 2022b; Deng et al. 2021). Thanks to the pseudo labels generated by the Teacher, we can perform intra-domain mixing with both labeled source (source-source) and pseudo-labeled target data (target-target). Moreover, we also deploy a balanced version of the inter-domain mixing used during the warmup phase. The reasons are twofold. First of all, with the presence of target pseudo labels, we now have the option to perform inter-domain mixing in the same manner as we

do for intra-domain labeled source mixing, which is something not feasible during the warmup phase. Secondly, the reason we do not want to continue using noise mixing is because it would become merely additional source domain data (the mixed-in unlabeled target image is only noise) and can potentially bias the model towards source distribution. Compare to this, the pseudo labels are much stronger signal that will push the model to learn target features.

Experiments: Object Detection

Experimental Settings

Datasets We conduct experiments on two standard benchmark datasets in object detection, namely PASCAL VOC (Everingham et al. 2010) and MS COCO (Lin et al. 2014). We follow (Zhang et al. 2019) and use the combination of PASCAL VOC 2007 *trainval* (5k images) and 2012 *trainval* (12k images) for training. Together they make up 16,551 images of 20 categories of common, real-world objects, each with fully annotated bounding boxes and class labels. The evaluation is done on PASCAL VOC 2007 *test* set (5K images). MS COCO (Lin et al. 2014) is composed of 80 object categories and is 10 times larger than PASCAL VOC. We train on *train2017* (118K images) and evaluated on *val2017* (5K images).

Baseline models We use three main baseline models to evaluate the performance of our proposed LossMix. The first one is a baseline, bare bone model without any Mixup-like data augmentation. Second, we compare LossMix against *Union* mixing, the current state-of-the-art approach widely used by prior studies (Zhang et al. 2019; Jocher 2020; Ge et al. 2021; Zhou et al. 2021; Wang et al. 2021; Zhang et al. 2022; Gao et al. 2022; Zheng et al. 2022; Yu et al. 2023; Zhang et al. 2023; Jocher, Chaurasia, and Qiu 2023) works around this by taking an unweighted, uniform *union* of all bounding boxes as new ground truth for the augmented image. Finally, we also compare with the "Noise" mixing strategy used by (Wang, Liao, and Shao 2021) for unsuperivsed domain adaptation. In a nutshell, it mixes input image A with a small amount of image B (e.g. $\lambda < 0.1$) acting only as color augmentation and discards any objects exists in B.

Implementation Details We leverage the open-source PyTorch-based Detectron2 (Wu et al. 2019) repository as our object detection codebase for experimentation. We use Faster RCNN (Ren et al. 2015) with ResNet (He et al. 2016)–FPN (Lin et al. 2017) backbone as our baseline model. By default, ImageNet1K (Deng et al. 2009) pretrained weights are used to initialize the networks. Unless otherwise specified, we use a batch size of 64 for faster convergence, an initial learning rate of 0.08, and the default step scheduler from Detectron2. We train PASCAL VOC for 18K iterations, which is about 70 epochs, and MS COCO for 270K iterations, or roughly 146.4 epochs. All experiments were trained with 8 NVIDIA GPUs, either V100 or A100.

Results

PASCAL VOC dataset Tab. 1 shows the results for LossMix in comparison with the baseline model and prior methods on PASCAL VOC dataset. First, we can see that all data mixing methods offer some improvements over the base Faster RCNN model, even "Noise" despite the weak mixing augmentation. This validates our interest in studying data mixing regularization for object detection. Second, among the detectors that deploys different mixing strategies, those with LossMix clearly outperform others. Specifically, our method yields up to $+0.9$AP compared to Union, $+1.5$AP compared to Noise, and $+2.7$AP compared to no-mixing baseline. Overall, LossMix achieves the best performance across all three evaluation metrics, AP, AP_{50}, and AP_{75}, as well as both backbones, ResNet-50 and ResNet-101 FPN.

Backbone	Method	AP	AP_{50}	AP_{75}
ResNet-50 + FPN	Baseline	53.30	78.89	59.11
	Noise	54.36	80.46	59.92
	Union	55.05	82.02	61.72
	LossMix (ours)	**55.87**	**82.44**	**62.88**
ResNet-101 + FPN	Baseline	53.25	79.87	59.24
	Noise	54.90	81.52	60.78
	Union	55.01	82.56	61.50
	LossMix (ours)	**55.91**	**82.84**	**62.72**

Table 1: PASCAL VOC results with Faster RCNN detector and ResNet-50/101 FPN backbone. For each method, we report the best checkpoint based on AP50 metric following PASCAL VOC standard. Best results are in bold. Our proposed LossMix outperforms state-of-the-art approaches such as Union and Noise to achieve the best overall results.

MS COCO dataset Our results for MS COCO dataset is shown in Tab. 2 Here, we can see that the promising performance of LossMix on PASCAL VOC is also generalizable to a much bigger ($10\times$) dataset such as MS COCO as well. In particular, our method again achieves the best overall AP scores at 41.82 for ResNet-50 and 44.07 for ResNet-101. When considering all metrics, LossMix also outperforms the previous state-of-the-art mixing techniques in the majority of cases. We believe these results, coupled with the simplicity of loss mixing operation, make LossMix an appealing alternative to the current unweighted union practice for data mixing in object detection.

Ablation study Although at its core, LossMix simply proposes the mixing of loss signals, there can be different implementation variations and hyper-parameters. Tab. 3 provides an ablation study investigating how these options affect the performance of LossMix. Overall, LossMix is robust with these configurations; all offer improvement over the Baseline (no data mixing) and the popular Union (Gao et al. 2022; Wang et al. 2021; Zhang et al. 2022, 2019; Zheng et al. 2022; Zhou et al. 2021) strategy. Morevover, we can see that although mixing of classification losses ($\mathcal{L}_{cls}^{rpn}$ and $\mathcal{L}_{cls}^{roi}$) contributes the most, mixing box regression losses ($\mathcal{L}_{reg}^{rpn}$ and $\mathcal{L}_{reg}^{roi}$) can also help, yielding better localization results as shown by AP_{75} as well as better overall AP. It is important to highlight that even when incorporating only

Backbone	Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-50 + FPN	Baseline	40.41	60.95	43.95	24.59	43.82	51.82
	Noise	41.01	61.74	44.98	25.20	44.71	52.34
	Union	41.43	**62.75**	45.68	**25.59**	45.19	52.61
	LossMix (ours)	**41.82**	62.51	**45.81**	25.04	**45.48**	**54.03**
ResNet-101 + FPN	Baseline	42.28	62.65	46.03	25.32	45.77	54.42
	Noise	42.60	62.85	46.58	25.83	46.62	55.16
	Union	43.87	**65.00**	48.39	**26.88**	47.91	55.70
	LossMix (ours)	**44.07**	64.48	**48.40**	26.73	**48.11**	**56.80**

Table 2: MS COCO results with Faster RCNN detector and ResNet-50/101 FPN backbone. Best checkpoints are selected according to the AP metric following MS COCO evaluation format. Models are trained for 270K iterations. Best numbers are in bold. The proposed LossMix outperforms the baseline and state-of-the-art methods for the majority of metrics.

Method	Input mixing	Loss mixing				Evaluation		
		$\mathcal{L}_{cls}^{rpn}$	$\mathcal{L}_{reg}^{rpn}$	$\mathcal{L}_{cls}^{roi}$	$\mathcal{L}_{reg}^{roi}$	AP	AP_{50}	AP_{75}
Baseline (no mixing)	✗	✗	✗	✗	✗	53.88	79.31	59.99
Uniform Union (no loss mixing)	$\lambda \sim Beta(1.0, 1.0)$	✗	✗	✗	✗	55.28	81.97	61.78
LossMix: ROI-only	$\lambda \sim Beta(1.0, 1.0)$	✗	✗	✓	✓	56.15	82.07	62.90
LossMix: Localization-only	$\lambda \sim Beta(1.0, 1.0)$	✗	✓	✗	✓	55.31	82.15	62.61
LossMix: Classification-only	$\lambda \sim Beta(1.0, 1.0)$	✓	✗	✓	✗	56.55	**82.71**	63.38
LossMix: $\alpha = 0.2$	$\lambda \sim Beta(0.2, 0.2)$	✓	✓	✓	✓	56.21	81.95	63.33
LossMix: $\alpha = 5.0$	$\lambda \sim Beta(5.0, 5.0)$	✓	✓	✓	✓	56.18	82.21	62.72
LossMix: $\alpha = 20.0$	$\lambda \sim Beta(20.0, 20.0)$	✓	✓	✓	✓	56.30	82.10	63.11
LossMix + RegMixup (Pinto et al. 2022)	$\lambda \sim Beta(1.0, 1.0)$	✓	✓	✓	✓	56.31	81.92	62.40
LossMix + Early Stop (Liu et al. 2023))	$\lambda \sim Beta(1.0, 1.0)$	✓	✓	✓	✓	56.42	82.20	62.93
LossMix (default)	$\lambda \sim Beta(1.0, 1.0)$	✓	✓	✓	✓	**56.60**	82.17	**63.59**

Table 3: Ablation study on PASCAL VOC dataset. The base detector is Faster RCNN with ResNet-50 FPN backbone. Best AP checkpoints are reported. Best numbers are in bold. For early stopping, we train the model with LossMix for the first 16k iterations out of a total of 18k. The mixing coefficient λ is sampled from $Beta(\alpha, \alpha)$ distribution, following Mixup.

box classification losses, our proposed method goes beyond image-level Mixup. This is because, by re-weighting $\mathcal{L}_{cls}$, LossMix effectively addresses a range of challenges related to spatial misalignment, background information, and object plurality that we have discussed in previous sections. In contrast, Mixup is not specifically designed to tackle these and cannot be adopted directly for object detection. This underscores the distinct advantages of our approach.

Experiments: Domain Adaptation

Experimental Settings

Datasets We conduct our experiments for cross-domain object detection using two popular and challenging real-to-artistic adaptation setups (Chen et al. 2020; Deng et al. 2021; Kim et al. 2019; Li et al. 2022b; Saito et al. 2019; Shen et al. 2019; Xu et al. 2020): PASCAL VOC (Everingham et al. 2010) → Clipart1k (Inoue et al. 2018) and PASCAL VOC (Everingham et al. 2010) → Watercolor2k (Inoue et al. 2018). Compared to PASCAL VOC, Clipart1k and Watercolor2k (Inoue et al. 2018) represent large domain shifts from real-world photos to artistic images. Clipart1k dataset shares the same set of object categories as PASCAL VOC and contains a total of 1000 images. We split these into 500 training and 500 test examples. Watercolor2k dataset, which has 2000 images from 6 classes in common with the PASCAL VOC, are split into 1000 training and 1000 test images.

Implementation Detail We leverage the open source code of the state-of-the-art Adaptive Teacher (Li et al. 2022b) framework. The codebase is also built on top of Detectron2 (Wu et al. 2019). For fair comparison against previous works (Deng et al. 2021; Li et al. 2022b), we use Faster RCNN (Ren et al. 2015) with ResNet-101 (He et al. 2016) backbone. We follow the setup of (Li et al. 2022b) and scale all training images by resizing their shorter side to 600 while maintaining the image ratios. We keep all loss weight for labeled and pseudo-labeled examples to be 1.0 for simplicity and use the default weight of 0.1 for the discriminator branch. We also keep the confidence threshold as 0.8. We notice that the set of hyper-parameter reported in the (Li et al. 2022b) is not suitable for the open-sourced code. Thus, we tune AT to get the best performance for fair comparison and keep their original set of strong-weak augmentations.

Method	Source	SCL	SWDA	DM	CRDA	HTCN	UMT	AT*	Noise	Union	Ours
mAP	28.8	41.5	38.1	41.8	38.3	40.3	44.1	46.7	44.9	50.0	**51.1**

Table 4: PASCAL VOC → Clipart1k adaptation results. The Average Precision (in %) for all object classes from is reported, following Deng et al. (2021) and Li et al. (2022b). The methods presented are SCL (Shen et al. 2019), SWDA (Saito et al. 2019), DM (Kim et al. 2019), CRDA (Xu et al. 2020), HTCN (Chen et al. 2020), UMT (Deng et al. 2021), AT (Li et al. 2022b), and Source (Faster-RCNN (Ren et al. 2015)). Best results are in bold. *indicated reproduced results using the released code.

Method	bike	bird	car	cat	dog	person	mAP
Source	84.2	44.5	53.0	24.9	18.8	56.3	46.9
SCL	82.2	55.1	51.8	39.6	38.4	64.0	55.2
SWDA	82.3	**55.9**	46.5	32.7	35.5	66.7	53.3
DM	-	-	-	-	-	-	52.0
UMT	88.2	55.3	51.7	**39.8**	**43.6**	69.9	58.1
AT*	**95.8**	51.7	**57.8**	36.5	33.1	71.0	57.7
Ours	91.1	55.8	54.3	39.1	41.0	**74.3**	**59.3**

Table 5: PASCAL VOC → Watercolor2k adaptation results. The Average Precision (in %) is reported following (Li et al. 2022b). Best numbers are in bold. 2nd best are underlined. *indicated reproduced results using official code.

	α	warm	adapt	$\mathcal{L}_{cls}^{rpn}$	$\mathcal{L}_{reg}^{rpn}$	$\mathcal{L}_{cls}^{roi}$	$\mathcal{L}_{reg}^{roi}$	AP_{50}
AT Baseline								46.7
LossMix: ROI	1.0	✓	✓			✓	✓	49.6
LossMix: Loc	1.0	✓	✓		✓		✓	48.2
LossMix: Cls	1.0	✓	✓	✓		✓		48.1
LossMix: warm	1.0	✓		✓	✓	✓	✓	49.1
LossMix: adapt	1.0		✓	✓	✓	✓	✓	48.1
LossMix: α=0.2	0.2	✓	✓	✓	✓	✓	✓	47.8
LossMix: α=5.0	5.0	✓	✓	✓	✓	✓	✓	49.8
LossMix: α=20.	20.	✓	✓	✓	✓	✓	✓	49.4
LossMix (final)	1.0	✓	✓	✓	✓	✓	✓	**51.1**

Table 6: Ablation study for PASCAL VOC → Clipart1k.

Results

PASCAL VOC → Clipart1k We compare with state-of-the-art methods in cross-domain object detection using the popular PASCAL VOC → Clipart1k adaptation (Tab. 4). We report an mAP of 50.33% across all object categories, achieving the new state-of-the-art performance with +3.5% improvement on top of the prior state of the art set by the recent Adaptive Teacher (Li et al. 2022b). Despite AT's strong performance, our results suggest that large domain shifts are still challenging and reveal potential biases toward source domain, e.g. inherently in the warmup procedure of Mean Teacher. By strategically leveraging LossMix, we are able to mitigate these problems and further improve accuracy.

Comparing with SOTA mixing Tab. 4 also presents our comparison to different Mixup variations used by existing methods, namely Union (Gao et al. 2022; Wang et al. 2021; Zhang et al. 2022, 2019; Zheng et al. 2022; Zhou et al. 2021) and Noise (Wang, Liao, and Shao 2021). Specifically, AFAN (Wang, Liao, and Shao 2021) deploys a small λ value on target domain image without any pseudo labels. This strategy is similar to our noise mixing during the warmup, but is used throughout the training. Note that this approach performs worse than our AT basedline. This is because although noise mixing could be helpful in general, as shown by both AFAN (Wang, Liao, and Shao 2021) and our following ablation studies, heavily relying on it in the adaptation phase of Mean Teacher can lead to bias towards the source domain due to the fact that "mixed-in" target information is only limited to a tiny amount to act as a domain-aware augmentation. Indeed, we believe for cross-domain mean teacher, the pseudo labels are much stronger target signals and should be taken advantage of appropriately. We also see sub-optimal results for Union (Zhang et al. 2019) due to errors in the approximation of unweighted union, similar to detection experiments. Tab. 6 shows an ablation study for more insights.

PASCAL VOC → Watercolor2k Next, we are interested in answering the question whether or not the encouraging gains observed in PASCAL VOC → Clipart1k can be reproduced on a different dataset. To do this, we use Watercolor2k and evaluate the performance of PASCAL VOC → Watercolor2k adaptation. Note that after experimenting with Clipart1k, we narrowed down our set of hyper-parameters to ones that work best for both Adaptive Teacher and our method for fair competition. For Watercolor2k, to test our method's robustness, we directly perform grid search on this small set of hyper-parameters without any further tuning or manual supervision. Nonetheless, even without exhaustive tuning, our results in Table 5 show that we can still outperform AT (mAP=57.7) and archive mAP=59.3 (+1.5 gain).

Conclusion

We tackle the challenges of applying data mixing augmentations to object detection. Specifically, we introduce Supervision Interpolation (SI), a novel conceptual reinterpretation and generalization of Mixup. Given SI, we propose LossMix, a simple-yet-effective regularization that interpolates the losses instead of labels to enhance model learning. Our experiments show consistent accuracy improvements, outperforming popular object mixing strategies and achieving state-of-the-art domain adaptation results. We hope this inspires future data mixing research for detection and beyond.

References

Abramov, A.; Bayer, C.; and Heller, C. 2020. Keep it Simple: Image Statistics Matching for Domain Adaptation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*.

Berthelot, D.; Roelofs, R.; Sohn, K.; Carlini, N.; and Kurakin, A. 2022. AdaMatch: A Unified Approach to Semi-Supervised Learning and Domain Adaptation. In *International Conf. on Learning Representations*.

Chang, O.; Tran, D. N.; and Koishida, K. 2021. Single-Channel Speech Enhancement Using Learnable Loss Mixup. In *Proc. Interspeech 2021*, 2696–2700.

Chen, C.; Zheng, Z.; Ding, X.; Huang, Y.; and Dou, Q. 2020. Harmonizing transferability and discriminability for adapting object detectors. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Dabouei, A.; Soleymani, S.; Taherkhani, F.; and Nasrabadi, N. M. 2021. SuperMix: Supervising the Mixing Data Augmentation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Deng; Jia; Dong; Wei; Socher; Richard; Li; Li-Jia; Li; Kai; Fei-Fei; and Li. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Deng, J.; Li, W.; Chen, Y.; and Duan, L. 2021. Unbiased Mean Teacher for Cross-domain Object Detection. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Everingham, M.; Van Gool, L.; Williams, C.; Winn, J.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 88: 303–338.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Hugo Larochelle, F. L.; Marchand, M.; and Lempitsky, V. 2016. Domain-Adversarial Training of Neural Networks. In *JMLR*.

Gao, Y.; Yang, L.; Huang, Y.; Xie, S.; Li, S.; and Zheng, W.-s. 2022. AcroFOD: An Adaptive Method for Cross-domain Few-shot Object Detection. In *European Conf. on Computer Vision*.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. YOLOX: Exceeding YOLO Series in 2021. arXiv:2107.08430.

Gu, Q.; Zhou, Q.; Xu, M.; Feng, Z.; Cheng, G.; Lu, X.; Shi, J.; and Ma, L. 2019. PIT: Position-Invariant Transform for Cross-FoV Domain Adaptation. In *IEEE/CVF International Conf. on Computer Vision*.

Harary, S.; Schwartz, E.; Arbelle, A.; Staar, P.; Abu-Hussein, S.; Amrani, E.; Herzig, R.; Alfassy, A.; Giryes, R.; Kuehne, H.; Katabi, D.; Saenko, K.; Feris, R.; and Karlinsky, L. 2022. Unsupervised Domain Generalization by Learning a Bridge Across Domains. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *Int'l. Conf. on Machine Learning*.

Hong, M.; Choi, J.; and Kim, G. 2021. StyleMix: Separating Content and Style for Enhanced Data Augmentation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Hoyer, L.; Dai, D.; and Van Gool, L. 2022. DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Hsu, H.-K.; Yao, C.-H.; Tsai, Y.-H.; Hung, W.-C.; Tseng, H.-Y.; Singh, M.; and Yang, M.-H. 2020. Progressive domain adaptation for object detection. In *IEEE/CVF Winter Conf. on Applications of Computer Vision*.

Inoue, N.; Furuta, R.; Yamasaki, T.; and Aizawa, K. 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Jeong, J.; Verma, V.; Hyun, M.; Kannala, J.; and Kwak, N. 2021. Interpolation-based semi-supervised learning for object detection. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Jocher, G. 2020. YOLOv5 by Ultralytics.

Jocher, G.; Chaurasia, A.; and Qiu, J. 2023. YOLO by Ultralytics.

Kim, J.-H.; Choo, W.; and Song, H. O. 2020. Puzzle Mix: Exploiting Saliency and Local Statistics for Optimal Mixup. In *International Conf. on Machine Learning*.

Kim, T.; Jeong, M.; Kim, S.; Choi, S.; and Kim, C. 2019. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Lee, S.; Cho, S.; and Im, S. 2021. Dranet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Li, S.; Wang, Z.; Liu, Z.; Wu, D.; Tan, C.; Jin, W.; and Li, S. Z. 2022a. OpenMixup: A Comprehensive Mixup Benchmark for Visual Classification. arXiv:2209.04851.

Li, W.; Li, F.; Luo, Y.; Wang, P.; and sun, J. 2020. Deep Domain Adaptive Object Detection: a Survey. In *SSCI*.

Li, Y.-J.; Dai, X.; Ma, C.-Y.; Liu, Y.-C.; Chen, K.; Wu, B.; He, Z.; Kitani, K.; and Vajda, P. 2022b. Cross-Domain Adaptive Teacher for Object Detection. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Liang, J.; Hu, D.; Feng, J.; and He, R. 2022. DINE: Domain Adaptation from Single and Multiple Black-box Predictors. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conf. on Computer Vision*.

Liu, J.; Liu, B.; Zhou, H.; Li, H.; and Liu, Y. 2022a. TokenMix: Rethinking Image Mixing for Data Augmentation in Vision Transformers. In *European Conf. on Computer Vision*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. SSD: Single Shot MultiBox Detector. In *European Conf. on Computer Vision*.

Liu, W.; Durasov, N.; and Fua, P. 2022. Leveraging Self-Supervision for Cross-Domain Crowd Counting. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Liu, X.-C.; Yang, Y.-L.; and Hall, P. 2022. Geometric and Textural Augmentation for Domain Gap Reduction. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Liu, Z.; Li, S.; Wu, D.; Liu, Z.; Chen, Z.; Wu, L.; and Li, S. Z. 2022b. AutoMix: Unveiling the Power of Mixup for Stronger Classifiers. In *European Conf. on Computer Vision*.

Liu, Z.; Wang, Z.; Guo, H.; and Mao, Y. 2023. Over-training with Mixup May Hurt Generalization. In *International Conf. on Learning Representations*.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning transferable features with deep adaptation networks. In *International Conf. on Machine Learning*.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2016. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*.

Lv, F.; Liang, J.; Gong, K.; Li, S.; Liu, C. H.; Li, H.; Liu, D.; and Wang, G. 2021. Pareto Domain Adaptation. In *Advances in Neural Information Processing Systems*.

Ma, X.; Wang, Z.; Zhan, Y.; Zheng, Y.; Wang, Z.; Dai, D.; and Lin, C.-W. 2021. Both Style and Fog Matter: Cumulative Domain Adaptation for Semantic Foggy Scene Understanding. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Meng, R.; Chen, W.; Yang, S.; Song, J.; Lin, L.; Xie, D.; Pu, S.; Wang, X.; Song, M.; and Zhuang, Y. 2021. Slimmable Domain Adaptation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Munir, M. A.; Khan, M. H.; Sarfraz, M. S.; and Ali, M. 2021. SSAL: Synergizing between Self-Training and Adversarial Learning for Domain Adaptive Object Detection. In *Advances in Neural Information Processing Systems*.

Murez, Z.; Kolouri, S.; Kriegman, D.; Ramamoorthi, R.; and Kim, K. 2018. Image to image translation for domain adaptation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Na, J.; Jung, H.; Chang, H. J.; and Hwang, W. 2021. FixBi: Bridging Domain Spaces for Unsupervised Domain Adaptation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Oza, P.; Sindagi, V. A.; VS, V.; and Patel, V. M. 2021. Unsupervised Domain Adaptation of Object Detectors: A Survey. arXiv:2105.13502.

Pinto, F.; Yang, H.; Lim, S.-N.; Torr, P. H. S.; and Dokania, P. K. 2022. RegMixup: Mixup as a Regularizer Can Surprisingly Improve Accuracy and Out Distribution Robustness. In *Advances in Neural Information Processing Systems*.

Ramamonjison, R.; Banitalebi-Dehkordi, A.; Kang, X.; Bai, X.; and Zhang, Y. 2021. SimROD: A Simple Adaptation Method for Robust Object Detection. In *IEEE/CVF International Conf. on Computer Vision*.

Rangwani, H.; Aithal, S. K.; Mishra, M.; Jain, A.; and Babu, R. V. 2022. A Closer Look at Smoothness in Domain Adversarial Training. In *International Conf. on Machine Learning*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*.

Saito, K.; Ushiku, Y.; Harada, T.; and Saenko, K. 2019. Strong-weak distribution alignment for adaptive object detection. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Shen, Z.; Maheshwari, H.; Yao, W.; and Savvides, M. 2019. SCL: Towards Accurate Domain Adaptive Object Detection via Gradient Detach Based Stacked Complementary Losses. arXiv:1911.02559.

Sun, B.; Feng, J.; and Saenko, K. 2016. Return of Frustratingly Easy Domain Adaptation. In *AAAI Conf. on Artificial Intelligence*.

Sun, T.; Lu, C.; Zhang, T.; and Ling, H. 2022. Safe Self-Refinement for Transformer-based Domain Adaptation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Venkataramanan, S.; Kijak, E.; Amsaleg, L.; and Avrithis, Y. 2022. AlignMixup: Improving Representations by Interpolating Aligned Features. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 19174–19183.

Verma, V.; Kawaguchi, K.; Lamb, A.; Kannala, J.; Solin, A.; Bengio, Y.; and Lopez-Paz, D. 2022. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145: 90–106.

Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold Mixup: Better Representations by Interpolating Hidden States. In Chaudhuri, K.; and Salakhutdinov, R., eds., *International Conf. on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 6438–6447. Long Beach, California, USA: PMLR.

Wang, H.; Liao, S.; and Shao, L. 2021. AFAN: Augmented Feature Alignment Network for Cross-Domain Object Detection. *IEEE Transactions on Image Processing*, 30: 4046–4056.

Wang, J.; Zhang, W.; Zang, Y.; Cao, Y.; Pang, J.; Gong, T.; Chen, K.; Liu, Z.; Loy, C. C.; and Lin, D. 2021. Seesaw Loss for Long-Tailed Instance Segmentation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Wu, Y.; Inkpen, D.; and El-Roby, A. 2020. Dual Mixup Regularized Learning for Adversarial Domain Adaptation. In *European Conf. on Computer Vision*.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2. Accessed: 2023-08-08.

Xu, C.-D.; Zhao, X.-R.; Jin, X.; and Wei, X.-S. 2020. Exploring categorical regularization for domain adaptive object detection. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Xu, M.; Zhang, J.; Ni, B.; Li, T.; Wang, C.; Tian, Q.; and Zhang, W. 2019. Adversarial Domain Adaptation with Domain Mixup. In *AAAI Conf. on Artificial Intelligence*.

Yu, H.; Yun, L.; Chen, Z.; Cheng, F.; Zhang, C.; and Lawrynczuk, M. 2023. A Small Object Detection Algorithm Based on Modulated Deformable Convolution and Large Kernel Convolution. *Intell. Neuroscience*, 2023.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *IEEE/CVF International Conf. on Computer Vision*.

Zhang, H.; Cisse, M.; Dauphin, Y.; and Lopez-Paz, D. 2018a. Mixup-CIFAR10. https://github.com/facebookresearch/mixup-cifar10/blob/main/train.py\#L138. Accessed: 2023-08-08.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018b. Mixup: Beyond empirical risk minimization. In *International Conf. on Learning Representations*.

Zhang, J.; Jin, J.; Ma, Y.; and Ren, P. 2023. Lightweight object detection algorithm based on YOLOv5 for unmanned surface vehicles. *Frontiers in Marine Science*, 9.

Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; and Wang, X. 2022. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. In *European Conf. on Computer Vision*.

Zhang, Z.; He, T.; Zhang, H.; Zhang, Z.; Xie, J.; and Li, M. 2019. Bag of Freebies for Training Object Detection Neural Networks. arXiv:1902.04103.

Zheng, S.; Chen, C.; Cai, X.; Ye, T.; and Tan, W. 2022. Dual Decoupling Training for Semi-supervised Object Detection with Noise-Bypass Head. In *AAAI Conf. on Artificial Intelligence*.

Zhou, Q.; Yu, C.; Wang, Z.; Qian, Q.; and Li, H. 2021. Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.