MUM : Mix Image Tiles and UnMix Feature Tiles for Semi-Supervised Object Detection

JongMok Kim[1,2] JooYoung Jang[1,2] Seunghyeon Seo[2] Jisoo Jeong[2] Jongkeun Na[1] Nojun Kwak[2]
[1]SNUAILAB, South Korea [2]Seoul National University, South Korea
{win98man, jyjang1090, zzzlssh, soo3553}@snu.ac.kr, jake.na@snuailab.ai, nojunk@snu.ac.kr

Abstract

Many recent semi-supervised learning (SSL) studies build teacher-student architecture and train the student network by the generated supervisory signal from the teacher. Data augmentation strategy plays a significant role in the SSL framework since it is hard to create a weak-strong augmented input pair without losing label information. Especially when extending SSL to semi-supervised object detection (SSOD), many strong augmentation methodologies related to image geometry and interpolation-regularization are hard to utilize since they possibly hurt the location information of the bounding box in the object detection task. To address this, we introduce a simple yet effective data augmentation method, Mix/UnMix (MUM) , which unmixes feature tiles for the mixed image tiles for the SSOD framework. Our proposed method makes mixed input image tiles and reconstructs them in the feature space. Thus, MUM can enjoy the interpolation-regularization effect from non-interpolated pseudo-labels and successfully generate a meaningful weak-strong pair. Furthermore, MUM can be easily equipped on top of various SSOD methods. Extensive experiments on MS-COCO and PASCAL VOC datasets demonstrate the superiority of MUM by consistently improving the mAP performance over the baseline in all the tested SSOD benchmark protocols. The code is released at https://github.com/JongMokKim/mix-unmix.

1. Introduction

Deep neural networks have made a lot of progress on diverse computer vision tasks thanks to the availability of large-scale datasets. To achieve better and generalizable performance, a large amount of labeled data is indispensable, which however requires a vast amount of workforce and time for annotation [3, 12, 31]. Unlike image classification, which needs only a class label per image, object detection requires a pair of a class label and location information for multiple objects per single image. Therefore, it is more challenging to acquire enough amount of labeled data in object detection. To address the above problem, many recent works have focused on leveraging abundant unlabeled data when training the network with a small amount of labeled data, called *semi-supervised learning* (SSL) and *semi-supervised object detection* (SSOD).

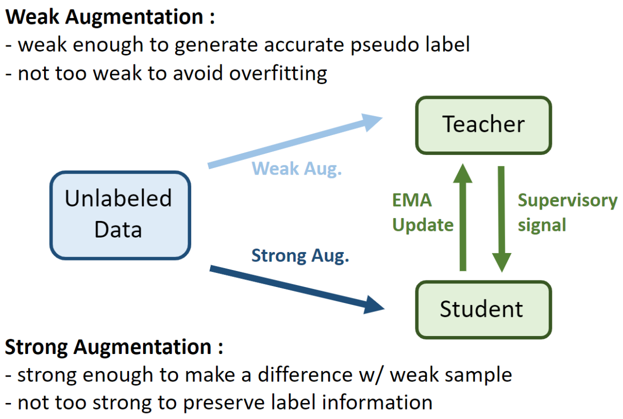
Weak Augmentation :
- weak enough to generate accurate pseudo label
- not too weak to avoid overfitting
Teacher
Unlabeled Data
Weak Aug.
EMA Update
Supervisory signal
Strong Aug.
Student
Strong Augmentation :
- strong enough to make a difference w/ weak sample
- not too strong to preserve label information

Figure 1. **Typical teacher-student (pseudo-labeling) framework for SSL.** To fully exploit the unlabeled data, building an intelligent teacher and employing an adequate data augmentation strategy for weak-strong pairs are very important in this framework.

In recent days, many SSL works rely on the teacher-student framework where a teacher network, typically a temporal ensemble model of the student, generates supervisory signals and trains the student network with them as shown in Fig. 1 [23, 35]. Data augmentation plays a significant role in this framework and most of the recent works apply strongly augmented inputs for the student model while weak augmentations are given to the teacher [33, 38]. Interpolation-regularization (IR), whose core idea is that the output of the interpolated input should be similar to the interpolated output of the original inputs, was originally developed as a data augmentation technique for supervised learning [46] and has been successfully applied to teacher-

This work was supported by the NRF grant (2021R1A2C3006659) and IITP grant (NO.2021-0-01343, Artificial Intelligence Graduate School Program) funded by the Korea government (MSIT).

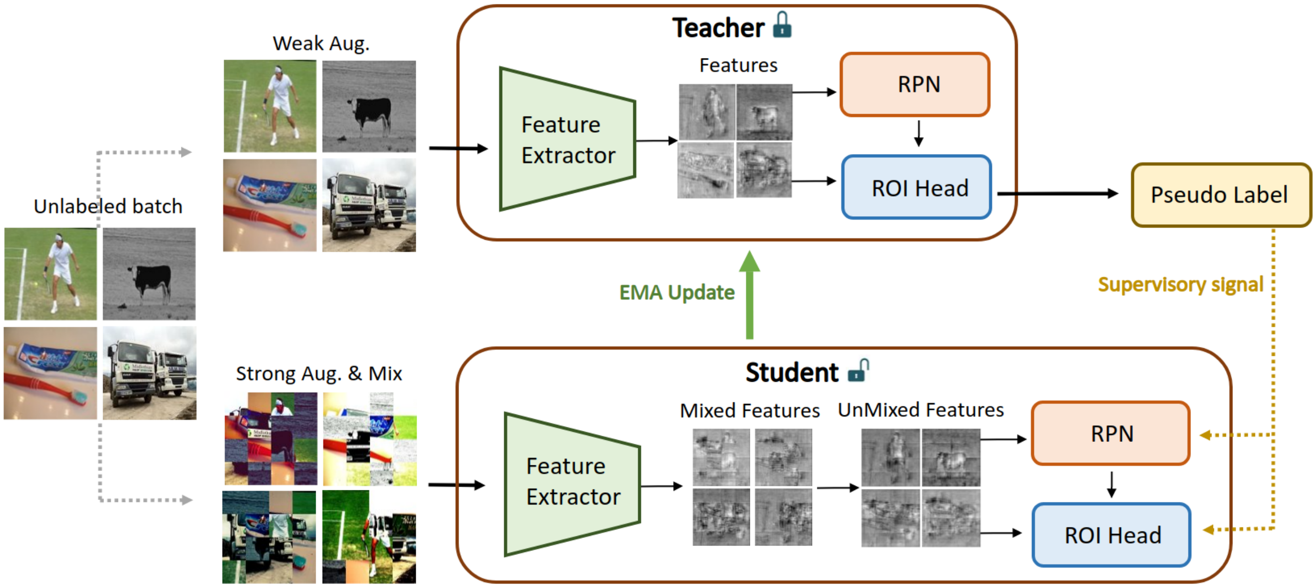

Figure 2. **Overview of Mix/UnMix (MUM) training system.** The teacher network generates a pseudo label to give a supervisory signal to the student, while weakly and strong & mixed augmented inputs are injected to the teacher and the student, respectively. In order to utilize the supervisory signal from the original shaped image, we unmix the mixed feature tiles and feed the unmixed features to the detection head in the student network. In each training step, the teacher network is slowly updated via EMA of the student's weights. For visual simplicity, we assume the batch size, N_T, and N_G are all identical to 4. For more details about the hyperparameters, N_T and N_G, see Sec.3.

student framework for SSL [6, 38]. It is a clever way to generate augmented input-output pairs without losing much contextual information and has also been extended to semantic segmentation by generating interpolated labels in a pixel-wise manner [14, 20].

However, it is challenging to make an interpolated label in the object detection task because it involves in multi-task learning, which consists of localization and classification. To tackle this problem, in this paper, we propose Mix/UnMix (MUM) method, which exploits IR in a much more efficient and more straightforward way for object detection (Fig. 2). MUM generates mixed images[1] by mixing image tiles in a batch and uses them as inputs to the student network. Then the feature maps extracted from the backbone are unmixed back to their original image geometry. The tiles maintain their positions in the original images through the mixing process so that it is possible for the feature maps to get back to their initial position through an unmixing phase. Therefore, the student network can learn from the mixed image without the interpolated (mixed) label. For the teacher network, input images are weakly augmented to generate highly confident pseudo labels as in other existing methods. As a result, the student can learn the robust features from the mixed and naturally occluded input image with the guide of a confident pseudo-label from the teacher network.

We benchmark Unbiased-Teacher [27] as a reliable baseline, which proposed a pseudo-labeling method for SSOD. Following the standard experimental setting of recent SSOD research, we adopted Faster-RCNN [30] as a default architecture. To verify the superiority of our algorithm, we test MUM on PASCAL VOC [13] and MS-COCO [26] dataset following the experimental protocols used in [27]. MUM achieves performance improvement against the baseline in every experimental protocol and could obtain the state-of-the-art performance in the SSOD benchmark experiment. Furthermore, thanks to the simplicity of MUM, the increase in the computational cost and complexity is negligible in the train phase, and it can be readily plugged in other SSOD frameworks as a data augmentation method. We also explore the versatility of MUM over different architectures through additional experiments with the Swin Transformer backbone. In addition, we tested performance of MUM for the supervised ImageNet classification task [10]. Our main contributions can be summarized as follows:

- We show the problem in applying the IR method to the pseudo-label-based semi-supervised object detection and propose a novel and simple data augmentation method, MUM, which benefits from IR.
- We experimentally prove our proposed method's superiority over a reliable baseline method through experiments and could obtain state-of-the-art performance on MS-COCO and PASCAL VOC dataset. Furthermore, we

[1] Mixed images can be considered as a type of interpolated images since they can be generated by patchwise interpolation with binary interpolation coefficients.

demonstrate the generalizability of our proposed method by still getting improved performance on a different backbone, Swin Transformer.

- Through thorough analysis of the feature maps, class activation maps and experimental results, we show the proposed MUM's compatibility with the SSOD problem.

2. Related Work

2.1. Semi-Supervised Learning

Since semi-supervised learning tackles practical problems regarding the cost of labeling and raw data acquisition, considerable progress has been made on improving the performance using only a few labeled data in combination with plenty of unlabeled data. Most SSL methods can be classified into two categories according how to generate supervisory signals from unlabeled raw data: consistency-based method [5, 6, 23, 29, 35, 38, 42] which induces consistent predictions for the same but differently augmented images and pseudo-labeling method [1, 2, 17, 24, 33, 43] which trains a student network with the highly confident label from a teacher network. As shown in Fig.1, to generate meaningful supervisory signals in the pseudo-labeling approach, it is necessary to equip with both a good teacher which can make better predictions than a student and an effective data augmentation method for generating data with different levels of difficulty under the same label. The most common and efficient method of building the teacher network is *Exponential Moving Average* (EMA) [35] which updates the teacher with a temporal ensemble of the student network. Regarding data augmentation, UDA [42], ReMixMatch [5] and FixMatch [33] apply RandAugment [9], CTAugment [5] and Cutout [11] as strong augmentations to generate data more difficult to learn than those from weak augmentations to make more meaningful supervisory signals. Interpolated-regularization is one of the efficient data augmentation methods in SSL and will be discussed further in Sec.2.3.

2.2. Semi-Supervised Object Detection

SSOD has gained significant attention for reducing burdensome cost of labeling in object detection task [18, 19, 27, 34, 39, 44, 47]. CSD [18] applied the consistency-regularization method, one of the mainstreams in SSL, into the object detection task. STAC [34] proposed the simple framework that trains a student network with pseudo-labels generated by a fixed teacher using unlabeled data. However, the fixed teacher network trained with only labeled data is insufficient to generate enough reliable pseudo-labels.

A line of recent works improves the teacher network and its pseudo-label by multi-phase training [39] or updates the teacher online by EMA [27, 44, 47], similar to MeanTeacher [35]. It leads to a reciprocal structure so that a teacher network generates supervisory signals helpful for improving the performance of a student network, and the teacher can also be strengthened by EMA update. Unbiased-Teacher [27] is composed of a simple SSOD framework that is robust to error propagation, using existing techniques such as EMA and Focal Loss [25]. It also made use of both strong and weak data augmentation, similar to FixMatch [33].

In contrast to SSL for classification tasks, the data augmentation methods in SSOD require the geometry of each augmented image to be identical for utilizing localization information from the teacher network's output as a supervisory signal. To overcome this constraint, we propose MUM that can mutate image geometry diversely and reduce the error propagation drastically.

2.3. Interpolation-based Regularization

IR is a method that derives high performance of a deep learning network by preprocessing inputs without noise injection and has been actively studied until recently [4, 7, 11, 15, 36, 37, 45, 46]. It generates new training samples by interpolating the original ones based on the inductive bias; the linear combination of two original samples' outputs should be similar to the output of the interpolated sample. Mixup [46], CutMix [45], Mosaic [15], and Cutout [11] are methods to synthesize and generate training samples and Manifold Mixup [37] deals with hidden representations in the feature level rather than with original images. Such methods can be regarded as strong data augmentation, and there have been several attempts to utilize them in SSL and SSOD.

ICT [38] trains a network by consistency loss between the interpolated prediction of two unlabeled samples and the prediction of an interpolated sample. MixMatch [6] and ReMixMatch [5] generate a guessed label from multi-view of a single unlabeled image, and then train it via Mixup [46] with labeled training sample. In addition, [14, 20] extends SSL to the semantic segmentation by generating mixed images via CutMix [45] and training with the same mechanism as ICT [38].

Unbiased-Teacher [27] also used Cutout [11] as a strong augmentation. However, Cutout results in information loss to the inputs because it drops pixel values of the random box-shape area in an image. Although ISD [19] applied IR adequately into the SSOD framework, it can be categorized more as a consistency-based method. Instant-Teaching [47] applied Mixup [46] and Mosaic [15] directly into pseudo-label-based SSOD framework, but the problem of class ambiguity of mixture between backgrounds and objects remains unsolved as mentioned in ISD [19]. To summarize, while Cutout [11] has a weak regularization effect, Mixup [46] has the class ambiguity issue in the interpolated label generation process. Motivated by these limitations, we propose MUM not only to avoid the problem caused by inter-

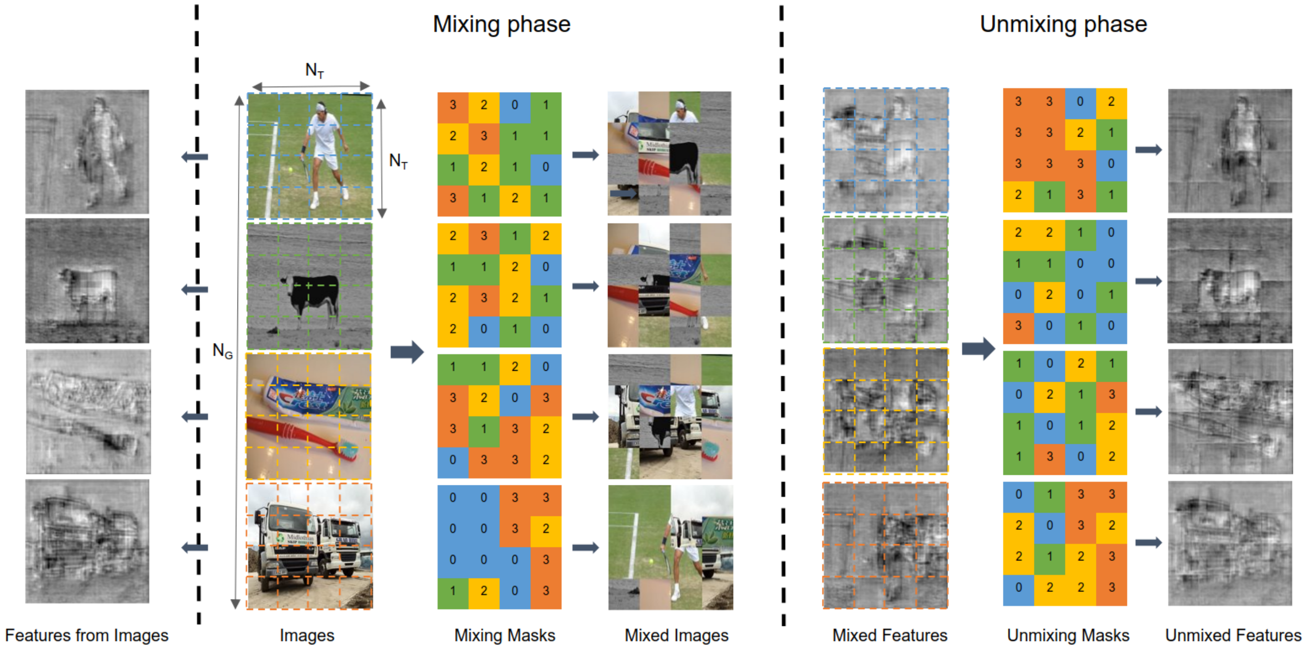

Figure 3. We provide the detailed operation of MUM with enlarged figures of images and features in Fig.2. With an assumption of $N_G = N_T = 4$, 4 images form a group, and each image is split into 4×4 tiles. Next, each input tile is mapped to the mixed image in the corresponding position of each mixing mask. Similar to the mixing phase, unmixed features are generated from mixed features with unmixing masks. Note that we generate mixing masks stochastically in each training step and unmixing masks are made from the mixing masks. Additionally, we provide the features from the original images to compare with unmixed features.

polated labels but also to still enjoy the IR effect.

3. Method

3.1. Preliminary

Problem definition. We deal with the semi-supervised object detection task, where a set of labeled data $\mathbf{D_s} = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and unlabeled data $\mathbf{D_u} = \{x_j^u\}_{j=1}^{N_u}$ is given for training. Here, x, y, N_s, N_u denote an image, the corresponding label, the number of labeled and unlabeled samples, respectively.

Baseline. Unbiased-teacher [27] is a well-designed architecture that employs the existing but competitive techniques like the Focal loss and EMA update method. They build a stable SSOD system with an unbiased teacher and its confident pseudo labels. To keep the above benefits, we choose it as our baseline. Following the baseline, we first build the teacher network via EMA:

$$\theta^{t+1} = \theta^t \cdot \delta + \theta \cdot (1 - \delta), \tag{1}$$

where θ^t, θ and δ denote the weights of the teacher at step t, the weights of the student, and EMA decay rate, respectively. Since the model performance is sensitive to the decay rate δ, it is essential to set the proper value to make the teacher better than the student. We will further discuss the effect of the decay rate δ on the system performance in Sec.5.

Next, we train the student network with the pseudo labels generated by the teacher network. The total training loss, $\mathcal{L}$, consisting of the supervised loss, $\mathcal{L}_s$, and the unsupervised loss, $\mathcal{L}_u$, can be described as follows:

$$\begin{aligned} \mathcal{L}_s &= \sum_i \mathcal{L}_{cls}(x_i^s, y_i^s) + \mathcal{L}_{reg}(x_i^s, y_i^s), \\ \mathcal{L}_u &= \sum_i \mathcal{L}_{cls}(x_i^u, \hat{y}_i^u) + \mathcal{L}_{reg}(x_i^u, \hat{y}_i^u), \\ \mathcal{L} &= \mathcal{L}_s + \lambda_u \cdot \mathcal{L}_u, \end{aligned} \tag{2}$$

where $\mathcal{L}_{cls}, \mathcal{L}_{reg}, \hat{y}^u$ and λ_u denote the loss for classification, the loss for bounding box regression, a pseudo label for an unlabeled image given by the teacher, and the balancing weight for the unsupervised loss, respectively.

3.2. Mixing Image/Unmixing Feature (MUM)

MUM. This section introduces the competitive data augmentation strategy, MUM (Mixing image tiles and UnMixing feature tiles), to leverage the unlabeled data effectively. Similar to the previous IR methods such as Mixup [46] and CutMix [45], we generate interpolated samples from each input mini-batch. We first split each image into $N_T \times N_T$ tiles. Simultaneously, we generate the same shaped $N_T \times$

Algorithm 1 Training procedure of the proposed MUM

```
Require: $(\mathcal{X}^s, \mathcal{Y}^s), \mathcal{X}^u$: pair of images and its labels, and unlabeled images
Require: $h(\cdot), \lambda_u$: loss function and balancing weight
Require: $f_{b,t}(\cdot)$, $f_{d,t}(\cdot)$: teacher object detection model (backbone and detector network)
Require: $f_{b,s}(\cdot)$, $f_{d,s}(\cdot)$: student object detection model (backbone and detector network)
Require: $m(\cdot)$, $u(\cdot)$: mixing and unmixing function
Require: $w(\cdot)$, $s(\cdot)$: weak and strong augmentation
for each $t \in [1, \text{max\_iterations}]$ do
    Prepare Data
        $\mathcal{A} \leftarrow w(\mathcal{X}^s) + s(\mathcal{X}^s), \mathcal{B} \leftarrow w(\mathcal{X}^u),\ \mathcal{C} \leftarrow s(\mathcal{X}^u)$
    Compute the supervised loss
        $\mathcal{P}^s \leftarrow f_{d,s}(f_{b,s}(\mathcal{A}))$
        $\mathcal{L}_{\mathcal{S}} \leftarrow h(\mathcal{P}^s, \mathcal{Y}^s)$
    Generate Pseudo Label
        $\hat{\mathcal{Y}}^u \leftarrow f_{d,t}(f_{b,t}(\mathcal{B}))$
    Mix Image Tiles & Unmix Feature Tiles
        $fm \leftarrow u(f_{b,s}(m(\mathcal{C})))$
    Compute the unsupervised loss
        $\mathcal{P}^u \leftarrow f_{d,s}(fm)$
        $\mathcal{L}_{\mathcal{U}} \leftarrow h(\mathcal{P}^u, \hat{\mathcal{Y}}^u)$
    Compute the total loss
        $\mathcal{L}_{Total} \leftarrow \mathcal{L}_{\mathcal{S}} + \lambda_u \cdot \mathcal{L}_{\mathcal{U}}$
    Update $f_s(\cdot)$ via $\mathcal{L}_{Total}$ and $f_t(\cdot)$ via EMA
end for
```

N_T mask to mix each image tile and get each feature tile back to its original position. Note that in the mixing phase all image tiles should be used once and keep their original geometric position in the image space for future reconstruction in the unmixing phase. In order to avoid the effect of mini-batch size on mixing, we predefine the number of images to form a group to mix as N_G. For example, assuming the mini-batch size of 32 and $N_G = 4$, then it forms 8 groups and the images are tiled and mixed within the corresponding group. The detailed example of MUM operation is provided in Fig.3.

Even though mixing tiles makes it hard to identify the edge or part of objects in images and feature maps, unmixing recovers the original position of features without loss of information. Unmixed features look degraded than the features from the original image since mixing tiles incurs severe occlusion so that each piece of feature tile can only make use of its local information. Therefore, MUM makes the student endeavor to predict like the teacher even with weak clues in features, and it is in line with the philosophy of previous studies [20, 33, 42] about weak-strong data augmentations.

Overall SSOD Framework. Employing MUM, we design the SSOD framework as shown in Fig.2. Similar to the baseline, we build the SSOD framework upon the pseudo-labeling method and the proposed MUM data augmentation. A mini-batch of unlabeled images is applied to the weak and strong augmentation as inputs to the teacher and student networks. The methods used to generate weak and strong augmentations are identical to those for the baseline [27]. Additionally for the student, we split and mix the input image tiles to generate mixed inputs and the feature maps of the mixed images are generated by the feature extractor. Then the mixed features are unmixed so that the original positions of all the tiles are restored. On the other hand, the teacher generates the supervisory signal for the inputs without the mixing process. Note that MUM can achieve the interpolation-regularization effect with a pseudo label of a single image because of the mixing-unmixing process in the student network. Including the above unsupervised learning process, the whole training process is described in Algorithm.1.

4. Experiments

Datasets. We evaluate our proposed method on two standard object detection datasets, PASCAL VOC [13] and MS-COCO [26], following the dominant benchmark of previous SSOD works [18, 27, 34, 47]. The benchmark has three protocols: (1) COCO-Standard: we randomly select 0.5, 1, 2, 5, and 10% of COCO2017-train dataset as labeled training data and treat the remaining data as unlabeled training data. (2) COCO-Additional: we utilize whole COCO2017-train dataset as labeled training data and the additional COCO2017-unlabeled dataset as the unlabeled training data. (3) VOC: we use VOC07-trainval set as the labeled training data and VOC12-trainval set as the unlabeled training data. To investigate the effect of the increased unlabeled data, we use COCO20cls [18] as additional unlabeled data. Model performance is tested on COCO2017-val and VOC07-test for evaluation following STAC [34] and Unbiased-Teacher [27].

Implementation Details. We use Faster-RCNN [30] with FPN [25] and ResNet-50 [16] initialized by ImageNet [10] feature extractor as base network architecture following Unbiased-Teacher [27]. We use training schedules of 180K, 360K, 45K and 90K iterations for COCO-Standard, COCO-Additional, VOC, and VOC with COCO20cls. Other training configuration is same as Detectron2 [41] and Unbiased-Teacher[2] for the sake of fair comparison. We use a low initial decay rate $\delta = 0.5$ and gradually increase it to 0.9996 at the same step of burn-in stage used in the baseline [27] instead of employing burn-in stage. MUM has its own two hyperparameters: N_G, N_T, which are the number of images to form a group and the number of tiles in each image axis,

[2] Code : https://github.com/facebookresearch/unbiased-teacher

Table 1. Experimental results ($AP_{50:95}$) on MS-COCO dataset with COCO-Standard and COCO-Additional protocols.

Methods	COCO-Standard					COCO-Additional
	0.5%	1%	2%	5%	10%	
Supservised	6.83±0.15	9.05±0.16	12.70±0.15	18.47±0.22	23.86±0.81	37.63
CSD [18]	7.41±0.21	10.51±0.06	13.93±0.12	18.63±0.07	22.46±0.08	38.82
STAC [34]	9.78±0.53	13.97±0.35	18.25±0.25	24.38±0.12	28.64±0.21	39.21
Instant Teaching [47]	-	18.05±0.15	22.45±0.15	26.75±0.05	30.40±0.05	39.6
ISMT [44]	-	18.88±0.74	22.43±0.56	26.27±0.24	30.53±0.52	39.6
Multi Phase [39]	-	-	-	-	-	40.1
Unbiased Teacher [27]	16.94±0.23	20.75±0.12	24.30±0.07	28.27±0.11	31.50±0.10	41.3
MUM(Ours)	**18.54±0.48**	**21.88±0.12**	**24.84±0.10**	**28.52±0.09**	**31.87±0.30**	**42.11**

Table 2. Experimental results on PASCAL VOC dataset compared with recent state-of-the-art results. Both protocols equally use VOC07 as labeled training dataset.

Methods	Unlabeled	AP_{50}	$AP_{50:95}$
Supervised	None	72.63	42.13
CSD [18]	VOC12	74.70	-
STAC [34]		77.45	44.64
Instant Teaching [47]		78.3	48.7
ISMT [44]		77.2	46.2
Multi Phase [39]		77.4	-
Unbiased Teacher [27]		77.4	48.7
MUM(Ours)		**78.94**	**50.22**
CSD [18]	VOC12 + COCO20cls	75.1	-
STAC [34]		79.08	46.01
Instant Teaching [47]		79.0	49.7
ISMT [44]		77.75	49.59
Unbiased Teacher [27]		78.82	50.34
MUM(Ours)		**80.45**	**52.31**

respectively. We use $N_G = N_T = 4$, which were found in our ablation study.

4.1. Results

MS-COCO. We first evaluate our proposed method on MS-COCO dataset with two protocols, COCO-Standard and COCO-Additional. As shown in Table 1, our approach obtains ∼2%p mAP gain over the baseline [27] and surpasses all of the recent state-of-the-art results. Specifically, for the 0.5% protocol in Table 1, MUM achieves 18.54% mAP which improves 11.71%p over the supervised results, and its performance is comparable to Instant Teaching and ISMT in 1% protocol as well (18.05 and 18.88). MUM brings more improvements when the labeled data is scarce (COCO-Standard 0.5% and 1%) since it generates many training samples with natural occlusions and diverse appearances.

Pascal VOC. We next test the proposed MUM on the Pascal VOC dataset with two protocols in Table 2. As in MS-COCO, our method consistently outperforms the state-of-the-art methods and achieves 1∼2%p mAP improvement over the baseline in both AP_{50} and $\text{AP}_{50:95}$. Specifically, MUM has 7.82%p, 10.18%p improvement for AP_{50} and $\text{AP}_{50:95}$ respectively, over the supervised baseline. While Unbiased Teacher shows relatively weak competitiveness compared to the other researches in the VOC dataset unlike the above COCO results, our method still surpasses the other state-of-the-art results with a large margin. These results demonstrate that our proposed method, MUM, can improve the existing SSOD consistently in various datasets.

4.2. Ablation Study

Analysis of N_G and N_T. MUM needs two hyperparameters: N_G and N_T, which indicate the number of images to group and shuffle the tiles, and the number of tiles in each image axis. In order to investigate the effect of two hyperparameters, we examine the performance of MUM with $N_T \in \{2, 4, 8, 16\}$ and $N_G \in \{2, 4, 6, 12\}$ in Table 3. We find $N_G = N_T = 4$ is an appropriate choice to keep MUM with diverse appearances and semantic information without much loss in the geometric information. When N_T increases to 8 and 16, the performances drop drastically since the tile's size becomes too small to keep the semantic information of positive objects.

We also observe the performance drop with increased N_G. However, it was negligible compared to the N_T case. Especially when N_G further increases beyond 4, $\text{AP}_{50:95}$ decreases slightly, but AP_{50} increases a bit. We assume this phenomenon is because large N_G encourages the network to distinguish objectness and classify objects better by using more occluded images (AP_{50} increased), while it prohibits the network from getting more accurate bounding box position ($\text{AP}_{50:95}$ decreased). However, the performance differences are not significant.

Swin Transformer Backbone. To further investigate the generality of MUM, we replace ResNet with Swin Transformer [28] and examine the performance in COCO-Standard protocols (Table 4). We use Swin-T, which is comparable to ResNet-50 in terms of computational complexity, from open-source library timm [40]. We first examine Unbiased-Teacher [27] baseline with Swin backbone.

Table 3. Comparison of mAP with various values of N_G and N_T in COCO-Standard 1% protocol. For simplicity, we set the training step, batch size as 45K and 12, respectively. We use fixed random seeds to remove the randomness.

Methods	N_G	N_T	$AP_{50:95}$	AP_{50}	AP_{75}
Baseline	1	1	18.40	34.99	17.48
MUM	**4**	**4**	**18.99**	**36.09**	**18.31**
	4	2	18.52	35.25	17.61
		8	18.28	35.19	17.00
		16	16.46	31.93	15.22
	2	4	18.92	35.94	17.89
	6		18.85	36.27	17.66
	12		18.84	36.12	17.56

Table 4. Comparison of mAP with Unbiased Teacher and MUM with Swin Transformer backbone in COCO-Standard. For simplicity, we set the traning step, batch size as 60K and 16, respectively. We use fixed random seeds to remove the randomness. $^+$ denotes our experiments.

Methods	COCO-Standard				
	0.5%	1%	2%	5%	10%
Supervised	10.16	13.43	18.7	23.67	27.41
Unbiased-Teacher$^+$	15.95	19.8	24	27.88	30.48
MUM(Ours)	**16.52**	**20.5**	**24.5**	**28.35**	**30.58**

We set the EMA decay rate to an empirically-found value, $\delta = 0.999$, since the default value (0.9996) brings poor results, even worse than the supervised baseline. And then, we apply MUM to the baseline configuration. In every protocols, MUM achieves $\sim$ 1%p improvement over the baseline. The efficacy of MUM in Swin is relatively marginal compared to the CNN since MUM possibly hurts the characteristics of the long-range dependency of Transformer.

Supervised Classification. MUM can enjoy a regularization effect without any interpolated label, so we extend this idea to the supervised classification task. We conducted additional experiments for the ImageNet [10] classification task under a supervised-learning setting. We compared MUM with vanilla ResNet, Cutout, Mixup, and CutMix by following the experimental protocol and training framework of CutMix[3]. We unmix the mixed features after layer-1 of ResNet and set N_G and N_T as 4 found in SSOD experiments.

As shown in Table 6, MUM outperforms the other methods except for CutMix with a top-1 error rate of 22.39%, which shows that MUM could also be used as a general data augmentation method for classification task. Compared to Cutout and Mixup, MUM generates much less information loss on the image, leading to a lower error rate. Furthermore, there is still room for improvement by finetuning the N_G, N_T, and layer location for unmixing.

Table 5. Ablation study on COCO-Standard 0.5% with Swin Transformer. T and T* denote default teacher ($\delta = 0.9996$), and refined teacher ($\delta = 0.999$) which is empirically found for Swin backbone. Note that the supervised only AP is 10.16 in Table 4.

	Cutout	MUM	T	T*	Teacher	Student
(1)	✓		✓		8.27	8.44
(2)	✓			✓	15.95	14.68
(3)	✓	✓	✓		14.55	14.22
(4)	✓	✓		✓	**16.52**	15.38

Table 6. Experiments results of MUM and existing IR methods, Cutout [11], Mixup [46], and CutMix [45] in supervised classification task.

Methods	Top-1 Err (%)	Top-5 Err (%)
Baseline	23.68	7.05
Cutout [11]	22.93	6.66
Mixup [46]	22.58	6.40
CutMix [45]	**21.40**	5.92
MUM(Ours)	22.39	6.44

5. Discussion

Teacher and Data augmentation. Building a good teacher and applying effective data augmentation is very important for pseudo-labeling-based SSOD systems, as mentioned in Fig. 1. In order to analyze how the two factors affect an SSOD system, we compare the worse and better approaches for building a teacher and augmenting data in Table 5. (1) With only Cutout (worse augmentation) and default EMA decay rate (worse teacher), the teacher's performance is even worse than the student's (8.44→8.27), and semi-supervised learning rather hurts mAP performance compared to the supervised learning (8.27 vs. 10.16). (2,3) If either one of MUM (better augmentation) and controlled EMA decay rate (better teacher) is used, semi-supervised learning turns to be helpful. A better teacher and better augmentation result in 5.79 and 4.39 mAP improvements (10.16 vs. 15.95, 14.55), respectively. It is remarkable that (3) still improves performance even with a worse teacher because MUM generates mixed input images which are difficult but worth learning and make SSOD helpful. Finally, using both better teacher and augmentation leads to the best performance (16.52) in (4). We confirm the importance of building a good teacher and data augmentation strategy in the SSOD framework from the experimental results.

Class-Activation-Map(CAM) Results. We further investigate the superiority of MUM over Unbiased Teacher by comparing the qualitative results of GradCAM [32] and box predictions in Fig. 4. We use Faster-RCNN with ResNet-50 and pre-trained weight in COCO-Standard 1% to get the results. We find that MUM concentrates more on the local region while baseline tries to look at global features, which allows the network with MUM to find small objects better

[3] Code : https://github.com/clovaai/CutMix-PyTorch

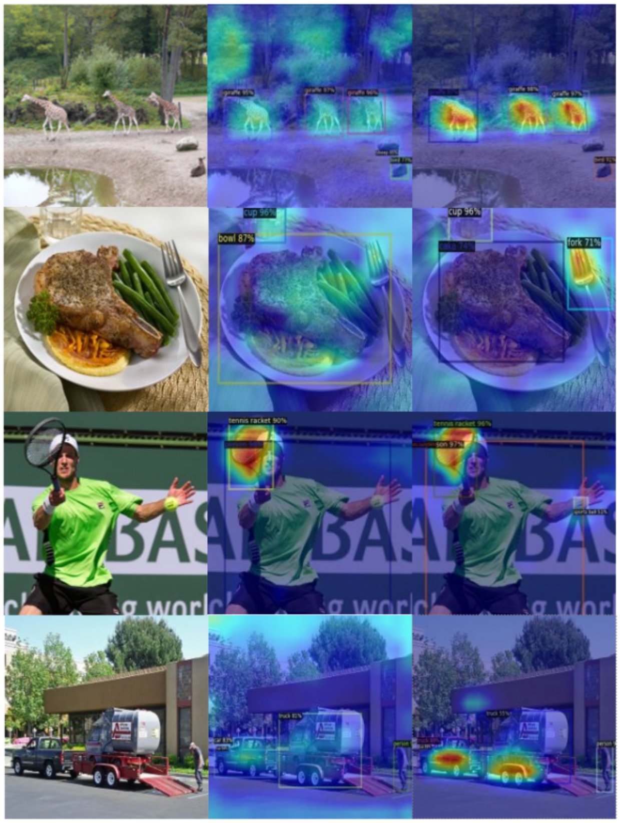

Figure 4. Class-Activation-Map (CAM) and box prediction results are provided. From left to right, each column shows the original images, outputs of Unbiased-Teacher and MUM, respectively. From top to bottom, the activated classes of each row are giraffe, fork, sports ball, and truck, respectively.

Table 7. Comparison of Unbiased Teacher and MUM by various APs in COCO-Standard 1% protocol.

Methods	$AP_{50:95}$	AP_S	AP_M	AP_L
Unbiased Teacher [27]	20.70	8.93	21.85	28.07
MUM(Ours)	21.81	**9.86**	**23.66**	27.91

such as sports ball and fork. Additionally, MUM classifies trucks and giraffe by highly focusing on each object. These results show that MUM encourages the network to extract meaningful features in the local region. Table 7 also provides quantitative results that MUM is more effective on small objects rather than large ones.

Connection to Cutout. Cutout [11] can be used in semi- and supervised object detection task [8, 15, 27] as a strong augmentation method by replacing the pixel blocks with random noisy values and generating diverse appearances and occlusions in training images. However, the information loss in the image is inevitable since it blocks some areas with noise. In addition, semantic information of an image that is crucial to predict the correct label can be significantly lost in the worst cases. On the other hand, our method creates natural occlusion between positive objects similar to Cutout because MUM mixes different images. However, MUM is able to preserve the semantic information of inputs because it doesn't block the original image with random noise and has a reassembling process in the feature space. Fig. 5 provides examples of augmented images, and shows the difference between Cutout and MUM. Additionally, we conduct the supervised object detection experiments following the configuration of Detectron2 [41] with Cutout and MUM, and achieve 36.87 and 38.12 mAP, respectively. We guess the characteristics of preserving the information of MUM bring the results.

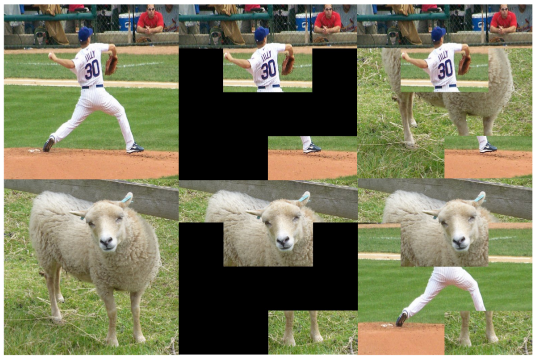

Figure 5. Comparison among original, Cutout, and MUM images. For simplicity and clear comparison, we assume the blocked region of Cutout is the same as mixed region of MUM and set N_G and N_T as 2 and 4, respectively.

6. Conclusion

In this paper, we investigate the pseudo-label-based SSOD system and propose the Mix/UnMix (MUM) data augmentation method, which mixes tiled input images and reassembles feature tiles to generate strongly-augmented images, while preserving the semantic information in the image space. On top of the pseudo-label-based SSOD framework, MUM obtains consistent performance improvement in SSOD benchmarks and achieves state-of-the-art results. We extend our experiments to a different backbone, Swin Transformer, and also applied MUM to a supervised ImageNet classification task. The experimental results show that our method is competitive with the existing IR methods and can also be used as a general regularization method for general architectures, and general tasks. We also provide Grad-CAM results to give further evidence why MUM works better. Additionally, we analyze the effect of teacher network and data augmentation to properly understand the MUM and SSOD framework. MUM has a weakness in accurately locating the prediction box since it splits the objects and blinds the edges. We believe that generating optimized mixing masks using saliency map of objects like [21, 22] could solve the above problem, and leave it as future work.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 3

[2] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27:3365–3373, 2014. 3

[3] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 1

[4] Christopher Beckham, Sina Honari, Vikas Verma, Alex M Lamb, Farnoosh Ghadiri, R Devon Hjelm, Yoshua Bengio, and Chris Pal. On adversarial mixup resynthesis. In *Advances in Neural Information Processing Systems*, pages 4348–4359, 2019. 3

[5] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019. 3

[6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5050–5060, 2019. 2, 3

[7] David Berthelot, Colin Raffel, Aurko Roy, and Ian Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer. In *International Conference on Learning Representations*, 2019. 3

[8] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 8

[9] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space, 2019. 3

[10] 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*, pages 248–255. Ieee, 2009. 2, 5, 7

[11] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017. 3, 7, 8

[12] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2012. 1

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 5

[14] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 2, 3

[15] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3, 8

[16] 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*, pages 770–778, 2016. 5

[17] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079, 2019. 3

[18] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32:10759–10768, 2019. 3, 5, 6

[19] Jisoo Jeong, Vikas Verma, Minsung Hyun, Juho Kannala, and Nojun Kwak. Interpolation-based semi-supervised learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11602–11611, June 2021. 3

[20] Jongmok Kim, Jooyoung Jang, and Hyunwoo Park. Structured consistency loss for semi-supervised semantic segmentation. *arXiv preprint arXiv:2001.04647*, 2020. 2, 3, 5

[21] Jang-Hyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. Co-mixup: Saliency guided joint mixup with supermodular diversity. *arXiv preprint arXiv:2102.03065*, 2021. 8

[22] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning*, pages 5275–5285. PMLR, 2020. 8

[23] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 1, 3

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

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3, 5

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 5

[27] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021. 2, 3, 4, 5, 6, 8

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 6

[29] Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 3

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 2, 5

[31] Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. Best of both worlds: human-machine collaboration for object annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2121–2131, 2015. 1

[32] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[33] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 1, 3, 5

[34] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 3, 5, 6

[35] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017. 1, 3

[36] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Between-class learning for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[37] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6438–6447, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 3

[38] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3635–3641. International Joint Conferences on Artificial Intelligence Organization, 7 2019. 1, 2, 3

[39] Zhenyu Wang, Yali Li, Ye Guo, Lu Fang, and Shengjin Wang. Data-uncertainty guided multi-phase learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4568–4577, 2021. 3, 6

[40] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 6

[41] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5, 8

[42] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019. 3, 5

[43] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 3

[44] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5941–5950, 2021. 3, 6

[45] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *arXiv preprint arXiv:1905.04899*, 2019. 3, 4, 7

[46] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 1, 3, 4, 7

[47] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4081–4090, 2021. 3, 5, 6