Proceedings Paper:
Chen, C., Debattista, K. and Han, J. (2024) Pseudo-labelling should be aware of disguising channel activations. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T. and Varol, G., (eds.) Computer Vision – ECCV 2024. The 18th European Conference on Computer Vision ECCV 2024, 29 Sep - 04 Oct 2024, Milan, Italy. Lecture Notes in Computer Science, 15121. , pp. 312-328. ISBN: 978-3-031-73035-1. ISSN: 0302-9743. EISSN: 1611-3349.

https://doi.org/10.1007/978-3-031-73036-8_18

Pseudo-Labelling Should Be Aware of Disguising Channel Activations

Changrui Chen[1], Kurt Debattista[1], and Jungong Han[1,2*]

[1] University of Warwick, WMG, UK
{Changrui.Chen,K.Debattista}@warwick.ac.uk
[2] University of Sheffield, Computer Science, UK
jungong.han@sheffield.ac.uk

Abstract. The pseudo-labelling algorithm is highly effective across various tasks, particularly in semi-supervised learning, yet its vulnerabilities are not always apparent on benchmark datasets, leading to suboptimal real-world performance. In this paper, we identified some channel activations in pseudo-labelling methods, termed **disguising channel activations** (abbreviated as **disguising activations** in the following sections), which exacerbate the confirmation bias issue when the training data distribution is inconsistent. Even state-of-the-art semi-supervised learning models exhibit significantly different levels of activation on some channels for data in different distributions, impeding the full potential of pseudo labelling. We take a novel perspective to address this issue by analysing the components of each channel's activation. Specifically, we model the activation of each channel as the mixture of two independent components. The mixture proportion enables us to identify the disguising activations, making it possible to employ our straightforward yet effective regularisation to attenuate the correlation between pseudo labels and disguising activations. This mitigation reduces the error risk of pseudo-label inference, leading to more robust optimization. The regularisation introduces no additional computing costs during the inference phase and can be seamlessly integrated as a plug-in into pseudo-labelling algorithms in various downstream tasks. Our experiments demonstrate that the proposed method achieves state-of-the-art results across 6 benchmark datasets in diverse vision tasks, including image classification, semantic segmentation, and object detection.

Keywords: Semi-supervised Learning · Pseudo-labelling

1 Introduction

Semi-supervised learning reduces the need for large amounts of labels, which is a pervasive challenge in data-driven algorithms for practical applications. Recently, methods employing pseudo-labelling [22] achieve state-of-the-art in numerous scenarios with incomplete labelling [17, 26, 47]. Nevertheless, we discover

* corresponding author

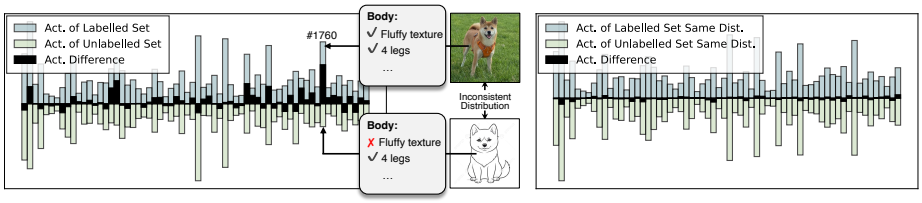


Fig. 1: The average activation (Act.) distribution of a layer in a model for the labelled and unlabelled set. We show the distribution of one category here. We sample once every 32 channels. The black bar is the difference between the two distributions. left) the distribution of the unlabelled set is different from the labelled set. right) the distribution of the unlabelled set is the same as the labelled set.

a noticeable channel activation discrepancy when the training data distribution consistency is not guaranteed. It misleads the pseudo-labeling algorithm to make decisions based on wrong features, causing the entire optimisation to fall into a *'confirmation bias'* [2] loop. In this paper, we investigate this phenomenon and propose a straightforward and effective method to address it.

Most semi-supervised learning algorithms are evaluated on benchmark datasets such as CIFAR [21], MSCOCO [25] *etc*. These datasets are inherited from fully supervised learning tasks and are organised by manually splitting into labelled/unlabelled sets with various label ratios. As a result, the labelled data and unlabelled data are drawn from one source with the same distribution. However, such a strong assumption is hard to uphold in practice. For example, labelled data is meticulously collected with high-quality cameras, while unlabelled data often comprises images from diverse sources, captured under different conditions or obtained from the internet. By analysing the activation of feature channels to the data of labelled and unlabelled sets in different distributions, we find that even the state-of-the-art methods [17] exhibit significantly different activation strengths on some channels for the same category. As shown in the left figure of Fig. 1, we collect images of a category and plot the channel activations of a well-trained model [17] dealing with the labelled and unlabelled set in different distributions. A longer black bar indicates a significant difference in the activation level. Some highly activated channels to labelled data, such as the channel #1760 in Fig. 1, are less sensitive to the unlabelled set. Such a phenomenon is invisible when the labelled and unlabelled data are in the same distribution (shown in the right of Fig. 1).

We attribute the phenomenon of channel activation discrepancy to the over-representation of information unique to either the labelled or unlabelled set, *i.e.*, the *set-private information*. For example, in Fig. 1, assume that the channel #1760 encodes the body-related information for recognising dogs. The high activation level of the channel #1760 for the labelled image indicates a very strong representation of the 'fluffy texture' and the '4 legs'. However, there is no texture information in the unlabelled image, thereby resulting in a significant lower activation level of the channel #1760. We refer to the intense activation caused by

the set-private information as **disguising channel activations** (abbreviated as **disguising activations** in the following sections). Once the decision-making process of the task-related module (*e.g.*, the linear classifier at the final layer in an image classification model) is overfitted to these disguising activations, it poses a high risk of poor performance for the pseudo-labelling algorithm on the unlabelled set as the channel activation values drop dramatically without the labelled-set-private information. Pseudo-labels with poor quality inevitably lead to the so-called *'confirmation bias'* issue.

To address this problem, we propose to identify the disguising activations and mitigate the correlation between task-related modules' decisions and them. To recognise disguising activations, we first consider what constitutes a disguising activation. The key principle is that *the disguising activations should contain sufficient one-set-private information while being relevant to the task-related layer's output.* Satisfying only one of the two conditions is insufficient. With a certain information capacity of a channel, if the activation is primarily contributed by the representation of the set-private information, enforcing the model to neglect this channel will not significantly improve the performance as this channel encodes limited task-related discriminative information. On the other hand, if the activation is dominated by crucial task-related information, neglecting it will confuse the model when learning a clear decision boundary as the model cannot converge without sufficient category discriminative features.

We propose a straightforward but effective strategy to highlight disguising activations with the above principle. Specifically, we model the activation of each channel as a mixture of two independent components — the set-private component and the task-related component. For example, in the training of a semi-supervised image classification algorithm, the first component encodes the information that exclusively exists in the labelled/unlabelled set, which should be useful for discriminating the set of the input image, while the second component is for the category attributes, which is crucial for classification. Such a modelling approach is more feasible than finding a one-to-one correspondence of channel activations and specific information. The mixture proportion indicates the contribution of these two components to the channel activation value in each channel and, most importantly, serves as the metric to discover disguising activations. Notably, comparing activations directly for each category to identify disguising activations statistically, as illustrated in Fig. 1, is not feasible since we lack access to the groundtruth labels of the unlabelled data. To lead the model to rely minimally on the disguising activations, we introduce channel masking regularisation guided by a mixture proportion. By doing so, the model will be encouraged to neglect disguising activations when making decisions, thereby yielding a better quality of pseudo-labels.

Extensive experiments are conducted on 36 settings of six benchmark datasets in three tasks — image classification, semantic segmentation, and object detection. The results show that the proposed method can improve the performance of the baseline pseudo-labelling based model by a significant margin. Moreover, our

method is not invasive and can be easily integrated into existing pseudo-labelling algorithms on various semi-supervised tasks without overhead.

Our contributions are summarised as follows:

- We discover and investigate the disguising activations problem in pseudo-labelling algorithms, as it exacerbates the confirmation bias issue in application scenarios.
- We identify disguising activations by modelling the activation of each channel as a mixture of two independent components. The mixture proportion guides us in regulating models to neglect such representations, thereby enhancing the quality of pseudo labels.
- We demonstrate how the baseline model equipped with the proposed method surpasses state-of-the-art by a significant margin on six benchmark datasets and three mainstream tasks, which reveals the effectiveness and generalisability of our method.

2 Related works

Semi-supervised Learning has been proposed to solve the problem of using a set of labelled data with a large amount of unlabelled data to optimise a model. Usually, algorithms in this topic are developed and evaluated on several benchmark datasets. The mainstream evaluation protocols are manually splitting the datasets inherited from the fully-supervised datasets with different label ratios into two sets. One set serves as the labelled set while the other one is the unlabelled set. Learning recognition patterns with limited labels and making the best use of the unlabelled data are crucial in solving this problem. There are three main categories of algorithms within this topic: a) generative models, b) graph-based methods, and c) pseudo-labelling models. Kingma *et al*. [20] proposed a stacked semi-supervised generative model, which appends a generative classifier to the latent representation produced by the encoder. Generative Adversarial Networks (GANs) [11] have also been explored as semi-supervised learning methods [30]. Apart from the generative models, graph-based models are introduced to model the data relationships to facilitate semi-supervised learning [29].

Pseudo Labelling permits predictions of an annotator model to be the pseudo-labels for optimising the model-self or a student model. Lee [22] first propose pseudo labelling with the prior of the low-density separation between classes. The form of the annotator model is implemented in a variety of ways. An exponential moving averaged (EMA) version [42] or even the student itself [41] was investigated to play the role of the teacher annotator. Several following works [3, 26, 41] achieved better performance by requiring models to produce consistent outputs when the inputs are perturbed. Image augmentations, such as flipping, cutout [9], or Gaussian Blurring, are usually applied to perturb input images. This technique was widely used by algorithms in the semi-supervised learning field, such as FixMatch [41] and achieved state-of-the-art performance [26, 41]. Within this framework, several works explored how to filter out low-quality pseudo labels

by using different policies. FlexMatch [47] dynamically adjust the threshold of the filtering threshold in a curriculum learning manner. SoftMatch [4] introduced a Gaussian distribution to assign weights for different unlabelled samples to solve the trade-off between the quality and quantity of the pseudo labels. Different from previous works, in this paper, we found that the channel activation discrepancy to data in different distributions hinders pseudo-labelling based methods from yielding high-quality pseudo labels. We propose a channel masking regularisation algorithm to solve this problem. Although domain adaptation has been explored for several years to solve the distribution shifting problems, most of domain adaptation methods make an application of pseudo-labelling without further studies [17,44]. A state-of-the-art pseudo-labelling algorithm — MIC [17] — which is used as the baseline in this paper, is inspired by several effective ideas in semi-supervised learning and domain adaptation. Even this state-of-the-art model still remains troubled by the disguising activations issue, which leads to potential risks for real-world application scenarios.

Channel Activation Analysis is not well-discussed in pseudo-labelling solutions. However, it is very popular in the network pruning area [10, 14]. Convolutional kernels are discarded if the activations are not relevant to the model's output to compress the model. Abbasi-Asl and Yu [1] adopted the most intuitive metric — accuracy reduction w/ and w/o the feature channels — to measure the channel importance. Some works [23] use the magnitude of the channel weight as the indicator of important filters. DomainDrop [12] analysed the channel activations in the domain generalisation scenario and proposed a channel dropout method to solve the domain private information issue. The proposed principle can only highlight the channels of great importance to the data set discrimination. In comparison, we propose a straightforward but effective method to find the activations which is of great importance to the model's decisions AND the data set discrimination.

3 Methodology

In this section, we first define the research problem of this paper. We then introduce the overall framework of our method. The proposed activation components modelling and the estimation of the mixture proportion follow. Finally, we describe the channel masking regularisation.

3.1 Problem definition

In the problem of semi-supervised learning, two data subsets $\mathcal{D}^l$, and $\mathcal{D}^u$ are given for model optimisation, where $\mathcal{D}^l = \{(x_n^l, y_n^l)|_{n=0}^{N^l}\}$ is the subset with available ground truth label y^l, and $\mathcal{D}^u = \{x_n^u|_{n=0}^{N^u}\}$ is the unlabelled subset. N^l and N^u are the data numbers of these two sets respectively. The pseudo labelling algorithm is using a so-called teacher (*i.e.* the annotator) to predict the pseudo labels $\{y_n'^u\}$ to get $\mathcal{D}'^u = \{(x_n^u, y_n'^u)|_{n=0}^{N^u}\}$ for the student (*i.e.* the model)'s optimisation with $\mathcal{D}^l$. In this paper, we introduce an additional common

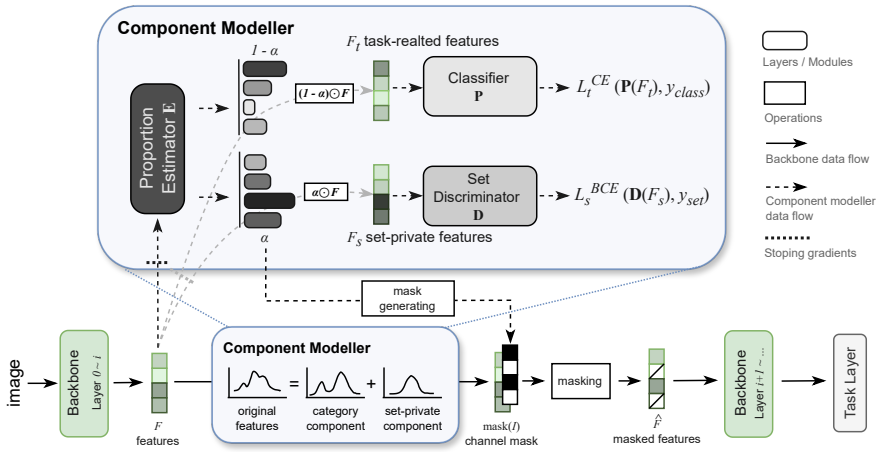


Fig. 2: bottom) The overall framework of our method. The original features, *i.e.*, the output of the convolutional layer i, are modelled as a mixture of two components. The mixture proportion is used to mask the original features to replace the original features in the forward-propagation of the following layers. top) The details of the component modeller. The gradient backpropagation is stopped at the beginning of the component modeller (represented by the dot bars). The proportion estimator consists of 4 linear layers with the LeakyReLU activation function. The output layer is a sigmoid function. We use two individual linear layers as the classifier **P** and the set discriminator **D**.

condition in application scenarios that there is a data distribution discrepancy between $\mathcal{D}^l$ and $\mathcal{D}^u$.

3.2 Overall

As we cannot access the groundtruth labels of the unlabelled data, it's impractical to highlight disguising activations statistically. Thus, we propose a learnable manner in this paper. Our method exclusively resides within the model (student)'s training with both labelled and unlabelled data, whereby the pseudo-labelling process is akin to other methods. We use a typical convolutional neural network for image classification, such as ResNet [13], as an example to demonstrate our overall framework.

As shown in the bottom of Fig. 2, for a convolution layer of index i with an activation function, the output feature tensor is processed by a *Component Modeller* before it is forwarded to the next layer. In the component modeller, we first model the original feature map as a mixture of two independent components with a mixture proportion. Then, we use the mixture proportion as guidance to mask the original feature map based on the principles proposed in Sec. 1. The masked feature map is then used to replace the original feature map in the following layers for the optimisation of the model.

3.3 Activation Components Modelling

Neural networks embed the information of input data, such as various attributes, into a high-dimensional feature space. In this paper, the feature tensor is represented by $F^{(i)} \in \mathbb{R}^{B\times C_i\times H_i\times W_i}$, where i is the layer index, B is the batch size, C_i is the channel number, H_i and W_i are the height and width of the feature map. For simplicity, we ignore the index term i, the size H, W, and the batch size B in the following, *i.e.*, $F \in \mathbb{R}^C$ is a feature map with a shape of $H \times W$ and C channels. Even though some research [46] claims that different channels of the feature map may represent different attributes, it is still hard to interpret the exact meaning of each channel. The reason is that channels and individual attributes do not have a one-to-one correspondence. Multiple different attributes may contribute to the activation of the same channel, with varying degrees of contribution. Therefore, in this paper, we model the mixture of activation components in each channel rather than enforce a one-to-one correspondence between channel activations and attributes. We consider two main components — the set-private one and the task-related one: $F = F_{\mathrm{s}} + F_{\mathrm{t}}$, where F_{s} and F_{t} are the set-private and task-related components of the feature map F, respectively. As the distribution of F_{s} and F_{t} are complicated, we cannot adopt a simple distribution, such as a Gaussian, to model them. Instead, we use a small neural network to learn how to model them as shown at the top part of Fig. 2. A learnable neural network $\mathbf{E}$ (called *Proportion Estimator* in Fig. 2) learns to produce a mixture proportion vector $\boldsymbol{\alpha} = [\alpha_0, ..., \alpha_{C-1}] = \sigma(\mathbf{E}(F))$, where C is the number of channels, σ is the sigmoid function, $\alpha \in [0, 1]$. Thus, F_{s} and F_{t} can be calculated as:

$$F_{\mathrm{s}} = \boldsymbol{\alpha} \odot F, \quad F_{\mathrm{t}} = (1 - \boldsymbol{\alpha}) \odot F, \tag{1}$$

where $\odot$ is the element-wise multiplication. We use a binary set indicator y_{set} and the classification labels y_{class} to penalise the proportion estimator with these two components jointly:

$$\min_{\theta_{\mathbf{D}},\theta_{\mathbf{P}},\theta_{\mathbf{E}}} \mathcal{L}_{\mathrm{cm}}(F;\theta_{\mathbf{D}},\theta_{\mathbf{P}},\theta_{\mathbf{E}}) = \min(\mathcal{L}_s^{\mathrm{BCE}}(\mathbf{D}(F_{\mathrm{s}}), y_{\mathrm{set}}) + \mathcal{L}_t^{\mathrm{CE}}(\mathbf{P}(F_{\mathrm{t}}), y_{\mathrm{class}})), \tag{2}$$

where $\mathbf{D}$ is a set discriminator, $\mathbf{P}$ is a linear classifier. We use the ground truth y^l as y_{classs} for labelled data, and the pseudo labels y'^u for unlabelled data. $\theta_{\mathbf{D}}, \theta_{\mathbf{P}}, \theta_{\mathbf{E}}$ indicate the learnable parameters in $\mathbf{D}$, $\mathbf{P}$, and the proportion estimator $\mathbf{E}$. The loss functions are binary cross entropy loss $\mathcal{L}_d^{\mathrm{BCE}}$ and the cross entropy loss $\mathcal{L}_c^{\mathrm{CE}}$. To prevent the backbone model from being affected by $\mathcal{L}_{\mathrm{cm}}$, we stop the gradient backpropagation at the beginning of the components modeller.

The optimisation objective of the components' modelling is to ensure that the error risk of the set discrimination with set-private components F_{s} is low, while the task-related component F_{t} yields a high classification accuracy.

3.4 Mixture Portion Guided Masking

By performing the above-mentioned joint optimisation, F_s and F_t are playing an adversarial game as they are yielded by splitting the original feature F based on the mixture proportion ratio $\boldsymbol{\alpha}$. There are three main cases for $\boldsymbol{\alpha}$:

1. For a certain channel, if the activation is mainly contributed by the representation of the set-private information which should be vital in distinguishing the data's set, α should be large to get a large F_s according to Eq. (1). Otherwise, it is not guaranteed that the $\mathcal{L}_s^{\mathrm{BCE}}$ can converge.
2. Conversely, if α is small, the value of the channel in F_t should be large. It indicates that this activation is crucial for the classifier $\mathbf{P}$ to make a correct decision, ensuring a low $\mathcal{L}_t^{\mathrm{CE}}$.
3. α approaching a value close to 0.5 shows a neck-to-neck competition between F_s and F_t. In this case, the channel activation is contributed by both set-private and task-related information. This activation is referred to as the disguising channel activation in this paper. Upon altering the input data distribution, if the model expects a high activation value of this channel when making decisions, error risks arise as the activation undergoes a substantial decrease due to the absence of the set-private information.

Based on the above analysis, we propose a channel masking regularisation to decrease the importance of activations with α close to 0.5 for the model. We first calculate a ranking score S for each channel's activation and collect the indices of the top k with the highest ranking scores:

$$S = -|\alpha - 0.5|, \qquad I = \mathrm{argsort}(S)[:k], \tag{3}$$

where I is the channel index set of the activations with top k ranking scores. k is determined by a hyperparameter — the regularisation ratio. Then, we mask the feature map F with the index set I:

$$\hat{F} = F \odot \mathrm{mask}(I), \tag{4}$$

where $\mathrm{mask}(I)$ is a binary mask with 0 at the indices in I and 1 at the other indices. Finally, the masked feature map $\hat{F}$ is used to replace the original feature map F in the following layers for the optimisation of the model. The overall loss function of the model with our method is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cm}} + \mathcal{L}_{\mathrm{task}}(\mathbf{G}(\hat{F}); y_{\mathrm{task}}), \tag{5}$$

where $\mathcal{L}_{\mathrm{task}}$ is the loss function of the task module $\mathbf{G}$ (*i.e.*, the linear classifier in the classification model here) and y_{task} is the ground truth labels y^l or the pseudo-labels y'^u.

By doing so, the model will be forced to neglect the disguising activations with α close to 0.5 when making decisions, thereby resulting in a better quality of pseudo labels.

4 Experiments

In this section, we first employ image classification as the main task to evaluate our proposed method. Then, we conduct experiments on semantic segmentation and object detection — to show how our method generalises. Finally, we conduct ablation studies to analyse the effectiveness of our method.

4.1 Datasets and Metrics

In contrast to the mainstream semi-supervised learning evaluation protocol, which manually splits fully labelled datasets into labelled and unlabelled sets for training, we employ several datasets with diverse data distributions from the domain adaptation community to evaluate our method. Specifically, we conduct experiments on six benchmark datasets: Office-Home [43], VisDA-2017 [32], DomainNet [31], Cityscapes/Foggy [8, 39], Synthia [36], and GTA [35]. We choose data in different distributions as the labelled and unlabelled set respectively. The detailed information and how we use these datasets is discribed in the supplementary material. For image classification, we report the top-1 accuracy. For semantic segmentation, the metric is the mean Intersection over Union(mIoU). For object detection, we report the mean average precision (mAP, *i.e.*, the AP50 of the COCO evaluation style [25]).

4.2 Implementation

The source code can be found in github.com/GeoffreyChen777/plda. The training of each experiment was conducted on a signle Tesla V100.

MIC [17] is adopted as the baseline. To ensure fairness, we use the official source code with the recommended hyperparameters to train the baseline model on our hardware with 3 random seeds. By doing so, we can ensure that disabling our method in our source code can yield the same performance as the baseline reported in the following tables. The experimental results are the average of 3 runs with different random seeds (*i.e.*, 0, 1, 2).

4.3 Performance

Image Classification The backbone we used for image classification is ResNet-50 [13] for Office-Home, and ResNet-101 for VisDA-2017 and DomainNet. The regularisation ratio (top k ratio) is set to 0.2. The proposed mixture proportion-guided masking regularisation is injected after each residual block. All other hyperparameters are exactly the same as the baseline's.

Office-Home We use the data in different distributions from Office-Home to create 12 semi-supervised settings. The results are shown in Tab. 1. Our method outperforms the baseline model (MIC) by 1.3% on average. The improvement is significant in most of the settings. For example, our method improves the baseline model by 2.3% on the setting Real-Clp.

Table 1: The performance on Office-Home with ResNet-50 backbone. The best results in each algorithm group are highlighted in **bold**. The second-best results are underlined. We compare our method with pseudo-labelling based methods and other methods proposed in the domain adaptation topic.

Labelled Set / **Unlabelled Set**	**Art** / **Clp.**	**Art** / **Prd.**	**Art** / **Real**	**Clp.** / **Art**	**Clp.** / **Prd.**	**Clp.** / **Real**	**Prd.** / **Art**	**Prd.** / **Clp.**	**Prd.** / **Real**	**Real** / **Art**	**Real** / **Clp.**	**Real** / **Prd.**	**Avg**
R-50 [13]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
CDAN [27]	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
MDD [48]	54.9	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	60.2	82.3	68.1
MCC [19]	57.0	76.0	81.6	64.9	75.9	75.4	63.7	56.1	81.2	74.2	63.9	85.4	71.3
SDAT [33]	58.4	77.4	81.6	66.4	76.5	76.5	63.5	56.3	82.0	75.0	64.5	85.4	71.9
Ours(SDAT)	**59.6**	**77.7**	**82.0**	66.5	**77.0**	77.2	64.5	**57.7**	**82.3**	75.6	**65.5**	**85.7**	**72.6**
↓ **Pseudo-labelling based Methods** ↓													
FixMatch [41]	51.8	74.2	80.1	63.5	73.8	61.3	64.7	51.4	80.0	73.3	56.8	81.7	67.7
CKP [28]	54.2	74.1	77.5	64.6	72.2	71.0	64.5	53.4	78.7	72.6	58.4	82.8	68.7
IA [18]	56.2	77.9	79.2	64.4	73.1	74.4	64.2	54.2	79.9	71.2	58.1	83.1	69.5
CAPLS [44]	56.2	**78.3**	80.2	66.0	75.4	**78.4**	**66.4**	53.2	81.1	71.6	56.1	84.3	70.6
MIC [17]	60.8	76.2	80.4	65.6	73.7	74.6	63.6	57.5	80.8	74.3	65.1	84.8	71.5
Ours(MIC)	**62.2**	77.6	**81.0**	**67.6**	**75.5**	76.5	65.2	**59.1**	**81.8**	**75.0**	**67.4**	**85.2**	**72.8**

Table 2: The performance on VisDA2017 with ResNet-101 backbone. The best results are highlighted in **bold**. The second-best results are underlined.

Categories	**plane**	**bcycl**	**bus**	**car**	**horse**	**knife**	**mcyle**	**persn**	**plant**	**sktb**	**train**	**truck**	**Avg**
R-101 [13]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
MCD [38]	87.0	60.9	**83.7**	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
CDAN [27]	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.9
MCC [19]	88.1	80.3	80.5	**71.5**	90.1	93.2	85.0	71.6	89.4	73.8	85.0	36.9	78.8
SDAT [33]	95.8	85.5	76.9	69.0	93.5	97.4	88.5	78.2	**93.1**	91.6	86.3	55.3	84.3
↓ **Pseudo-labelling based Methods** ↓													
IA [18](R-50)	-	-	-	-	-	-	-	-	-	-	-	-	76.7
MIC [17]	96.7	81.9	82.4	66.8	94.2	**97.9**	89.5	81.4	92.3	92.9	89.5	51.8	86.5
Ours (MIC)	**97.0**	**90.1**	83.6	70.8	**95.7**	96.2	**90.4**	**83.1**	92.3	**95.3**	**89.5**	**54.3**	**87.4**

In addition, we integrate our method with SDAT [33], which is not a pseudo-labelling framework. Our method can also boost the performance of SDAT by 0.7% on average. The reason is that the proposed regularisation enforces the model to neglect channel activations caused by the set-private information of the labelled set, which misleads the model's prediction on unseen data. As pseudo-labelling methods such as MIC are suffering from the confirmation bias issue, more improvement is expected when using our method.

VisDA2017 Tab. 2 shows the results on VisDA2017 dataset. Even though the baseline model has already achieved high accuracy, our method still improves it by 0.9% on average. For some categories such as the *bicycle*, our method improves the baseline model by 8.2%.

DomainNet We report the results on DomainNet in Tab. 3. Similarly, our method achieves the best in most settings. The average improvement is 1.3%. For this dataset, we observed a training collapse of the baseline model on the Infograph-Sketch setting. Thus, we ignore the performance of this setting for all methods when calculating the average performance.

Mix-Office-Home These results on the 32 semi-supervised image classification settings demonstrate the effectiveness of our method. In addition, we create a

Table 3: The performance on DomainNet with ResNet-101 backbone. The best results are highlighted in **bold**. The second-best results are underlined. *As the baseline model cannot converge on the Infograph-Sketch labelled/unlabelled set pair, we ignore this setting in the calculation of the averaged accuracy for all methods.

Labelled Set	Unlabelled Set	CDAN [27]	SDAT [33]		MIC [17]	Ours (MIC)
Clipart	Sketch	44.9	47.2	↓ Pseudo-labelling based Methods →	49.0	**49.7**
	Product	38.9	41.5		40.5	**44.1**
	Real	56.0	57.5		61.5	**61.6**
	Infograph	20.6	**22.0**		19.9	21.2
Sketch	Clipart	56.0	58.7		65.1	**65.3**
	Product	45.3	48.1		53.6	**54.4**
	Real	54.9	57.1		62.6	**63.1**
	Infograph	20.7	21.8		22.9	**23.2**
Product	Clipart	44.1	47.5		48.3	**49.6**
	Sketch	40.0	41.8		43.8	**44.2**
	Real	57.2	58.0		58.4	**58.8**
	Infograph	19.8	20.7		21.2	**21.2**
Real	Clipart	55.8	56.7		60.1	**61.1**
	Sketch	42.3	43.9		43.6	**44.6**
	Product	53.2	53.6		57.5	**57.8**
	Infograph	24.4	**25.1**		24.3	23.0
Infograph	Clipart	31.6	33.9		37.0	**39.2**
	Sketch*	26.4	27.9		-	-
	Product	29.3	30.3		35.5	**39.6**
	Real	43.6	48.1		49.0	**55.4**
Average		41.0	42.8		44.9	**46.2**

Table 4: The classification accuracy on the mix-distribution OfficeHome.

Methods	Avg Acc.
MIC [17]	77.4
Ours(MIC)	**78.5**

Table 5: The semantic segmentation mIoU with GTA/Cityscapes as the labelled/unlabelled set.

Methods	mIoU
DAFormer [15]	68.3
HRDA [16]	73.8
MIC [17]	74.8
Ours(MIC)	**75.9**

Table 6: The object detection mAP on Cityscapes (labelled) and Foggy Cityscapes (unlabelled).

Methods	mAP
DAFaster [6]	32.0
SWDA [37]	35.3
SIGMA [24]	44.2
SADA [7]	44.0
↓ PL based Methods ↓	
MTOR [49]	35.1
MIC [17]	47.6
Ours(MIC)	**48.4**

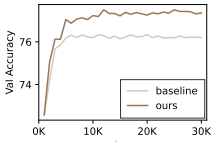
mix-distribution dataset based on Office-Home. As the distribution of unlabelled images in application scenarios is usually unknown, we use data from two distributions to create the unlabelled data. The accuracies of our method and the baseline mode are shown in Tab. 4. Our method boosts the baseline by 1.1%. More results are shown in the supplementary material.

Semantic Segmentation The backbone in semantic segmentation is a Vision Transformer (MiT) [45]. The experiments on semantic segmentation demonstrate the good generalisation ability of our method. The proposed mixture proportion-guided masking regularisation is injected in the ASPP module [5]. The regularisation ratio is set to 0.3 for Synthia and 0.1 for GTA.

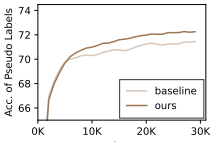
GTA We use GTA as the labelled set, and Cityscapes as the unlabelled set to conduct the semantic segmentation experiment. The results of the Cityscapes validation data are shown in Tab. 5. Our method improves the baseline model by 1.1% on mIoU.

Table 7: The semantic segmentation performance on the setting of Synthia as the labelled set while Cityscapes as the unlabelled set. All methods are based on pseudo-labelling algorithms.

Methods	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Sky	Person	Rider	Car	Bus	M.bike	Bike	mIoU
DAFormer [15]	84.5	40.7	88.4	41.5	6.5	50.0	55.0	54.6	86.0	89.8	73.2	48.2	87.2	53.2	53.9	61.7	60.9
HRDA [16]	85.2	47.7	88.8	**49.5**	4.8	57.2	65.7	60.9	85.3	92.9	79.4	52.8	89.0	64.7	63.9	64.9	65.8
MIC [17]	87.3	52.9	**89.8**	49.1	8.7	58.9	**66.8**	61.3	**86.4**	**94.5**	**81.2**	**58.6**	89.2	57.3	**67.3**	64.1	67.0
Ours(MIC)	**87.5**	**54.8**	89.2	48.1	**10.5**	**60.7**	65.9	**65.5**	84.4	93.5	80.2	57.3	**89.9**	**65.5**	66.4	**66.4**	**67.9**

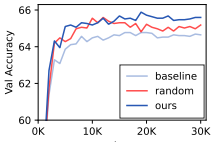


(a) Art(l) Product(u)

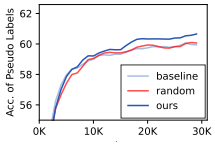


(b) Art(l) Product(u)

Fig. 3: a) The accuracy of the baseline model and the model with our method on the validation set. b) The accuracy of the pseudo-labels during training.



(a) Product(l) Art(u)



(b) Product(l) Art(u)

Fig. 4: The ablation study of the random masking strategy.

Synthia We conduct additional experiments on the Synthia dataset. The results are shown in Tab. 7. Our method improves the baseline model by 0.9% on mIoU.

Object Detection The backbone in object detection is ResNet-50 and Faster-RCNN [34]. The proposed mixture proportion-guided masking regularisation is injected before the RCNN head. The regularisation ratio is set to 0.3.

Foggy Cityscapes We show the experimental results on the object detection task in Tab. 6 to further demonstrate generalisation. Our method improves the mAP of the baseline model by 0.8%. The improvement is significant in various categories in this dataset, as shown in the supplementary material.

4.4 Ablation Studies and Discussion

By comparing the performance of 'Ours' in Tabs. 1 to 7 with the results of the corresonding baseline well demonstrates the effectiveness of our method. In addition, we ablate and discuss our method in image classification on the Office-Home dataset for further analysis by answering a few questions that may arise.

A. Does our method really improve the quality of pseudo-labels?

To answer this question, we record the accuracy of the pseudo-labels calculated by the ground truth during the training process and plot the curves in Fig. 3. The accuracy of the pseudo-labels of our method is significantly higher than the baseline model (MIC). More figures can be found in the supplementary.

B. Can we randomly mask the channels?

As shown in Fig. 4 (also in Tab. 8), the performance of the model with random masking (the red curve) is worse than our method (the dark blue curve). This

Table 8: The Ablation studies of different masking strategies (seed 0).

	Prd.(l) Art(u)	Art(l) Clp.(u)	Prd.(l) Real(u)
Baseline	64.6	61.0	81.2
Random	65.1	61.3	81.9
Task-related	64.7	61.8	82.0
Set-private	64.8	61.0	82.0
Ours	**65.6**	**62.1**	**82.5**

Table 9: Ablation study of different masking ratios (Prd.(l) Art(u), seed 0).

Top-K Ratio	0.0 (Baseline)	0.1	0.2	0.3	0.4	0.5
Acc.	64.6	65.3	**65.6**	64.5	63.3	62.9

demonstrates that the masking strategy is crucial to the performance of the model. Random masking exceeds the baseline model revealing that disguising activations exist in no doubt. In Fig. 4, at the beginning of the training, the performance of the random strategy is comparable with our mixture proportion guided masking strategy as our component modeller is still under-fitted. After some iterations, the performance of the random strategy drops significantly, while our method continues to improve. This demonstrates that the mixture proportion estimation network can guide us to regularise the model effectively.

C. What kind of masking strategy is better? Can we mask the activations of the most set-private information or task-related information?

As previously discussed, the experiment employing a random masking strategy has revealed that some activations should be masked while others should not. Through the estimation of the mixture proportion, we are able to ablate our model by masking channel activations dominated by different components to answer this question. As reported in Tab. 8, the performance is scarcely enhanced when masking activations contributed by a great proportion of the task-related component. This is because such masking confounds the task-related layer, hindering its ability to learn an accurate decision boundary in the absence of sufficient category discriminative information. Similarly, the accuracy is not significantly improved by masking activations with a large proportion of the set-private component. This can be attributed to the inefficiency of masking policy as these channels do not dominate the model's decision.

D. What will happen if we change the masking ratio?

We ablate our method with different masking ratios and report the results in Tab. 9. The one we adopted in this paper, *i.e.*, 0.2, achieves the best.

E. Can we confirm that the correlation between the disguising activations and the model's prediction is minimised?

Inspired by the Grad-CAM [40], we propose a metric — Grad-CAM Importance score (GCAM-I) — to measure and compare the importance of the disguising activations and other activations to the model's predictions. For a feature vector F, we perform a channel-wise Grad-CAM by calculating the gradient of the output *w.r.t.* the feature F to get the GCAM-I for activations of each channel. By employing our component modeller and the regularisation ratio (*e.g.*, 0.2 here), we choose 20% channel activations guided by the mixture proportion as the disguising activation to calculate the average GCAM-I as the disguising activations' GCAM-I(Disguising A. GCAM-I in Fig. 5). For other channels, we average the top 20% GCAM-I as the other crucial activations' GCAM-I (Other A. GCAM-I in Fig. 5). As shown in Fig. 5, the GCAM-I of the disguising activa-

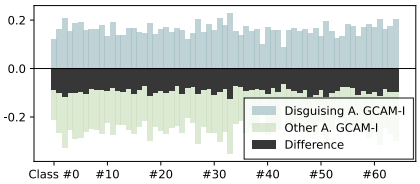


Fig. 5: The class-wise average Grad-CAM Importance (GCAM-I) of disguising activations and other crucial activations.

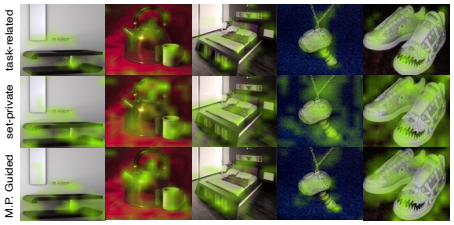


Fig. 6: The visualisation of activations masked by different strategies. 'M.P. guided' means the mixture proportion guided masking strategy.

tions is significantly lower than the one of other channels across all classes, which confirms that the predictions of the model are mainly based on other activations rather than the disguising activations. In other words, the correlation between the disguising activations and the model's prediction is mitigated.

F. Can we visualise the masked activations?

In Fig. 6, we visualise the feature maps of activations masked by the various strategies we discussed earlier via Grad-CAM. The initial row shows the activations containing a substantial proportion of the task-related component, while the second row is for the set-private component. The task-related component typically resides within objects of interest, exemplified by the ham in the first row. In contrast, the set-private component extends beyond the regions of specific objects and can be identified not only in object-related areas but also in numerous task-irrelevant regions, such as the background in the second row. Consequently, masking such activations is inefficient. The activations masked by the proposed mixture proportion guided strategy are shown in the third row. Our strategy concentrates on activations that are pertinent to the task while concurrently containing a considerable amount of set-private information.

5 Conclusions

This paper introduced and discussed the disguising activations in pseudo-labelling algorithms, which are a potential risk for a semi-supervised learning model as they exacerbate the confirmation bias issue in practice. By modelling the channel activation with two independent components, we successfully identified such disguising activations. Moreover, A straightforward but effective masking regularisation guided by the activation modelling results was proposed in this paper. It allows the model to learn how to make decisions without disguising activations to increase the robustness when dealing with data in different distributions. To the best of our knowledge, this paper is the first to identify, discuss and provide a solution to the disguising activations issue.

Acknowledgments We thank China Scholarship Council for the funding.

References

1. Abbasi-Asl, R., Yu, B.: Structural Compression of Convolutional Neural Networks. arXiv (2017)
2. Arazo, E., Ortego, D., Albert, P., Connor, N.E.O., McGuinness, K.: Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. In: IJCNN. pp. 1–8 (2020)
3. Chen, C., Han, J., Debattista, K.: Virtual Category Learning: A Semi-Supervised Learning Method for Dense Prediction with Extremely Limited Labels. PAMI pp. 1–17 (2024)
4. Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Raj, B., Savvides, M.: Softmatch: Addressing the Quantity-Quality Trade-off in Semi-supervised Learning. In: ICLR (2023)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. PAMI **40**(4), 834–848 (2018)
6. Chen, Y., Li, W., Sakaridis, C., Dai, D., Gool, L.V.: Domain Adaptive Faster R-CNN for Object Detection in the Wild. In: CVPR. pp. 3339–3348 (2018)
7. Chen, Y., Wang, H., Li, W., Sakaridis, C., Dai, D., Gool, L.V.: Scale-Aware Domain Adaptive Faster R-CNN. IJCV **129**(7), 2223–2243 (2021)
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic $ Urban Scene Understanding. In: CVPR. pp. 3213–3223. IEEE (2016)
9. Devries, T., Taylor, G.W.: Improved Regularization of Convolutional Neural Networks with Cutout. arXiv **abs/1708.04552** (2017)
10. Ding, X., Ding, G., Guo, Y., Han, J., Yan, C.: Approximated Oracle Filter Pruning for Destructive CNN Width Optimization. In: ICML. pp. 1607–1616 (2019)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
12. Guo, J., Qi, L., Shi, Y.: Domaindrop: Suppressing Domain-Sensitive Channels for Domain Generalization. In: ICCV (2023)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR. pp. 770–778 (2016)
14. He, Y., Liu, P., Wang, Z., Hu, Z., Yang, Y.: Filter Pruning via Geometric Median for Deep Convolutional Neural Networks Acceleration. In: CVPR. pp. 4340–4349 (2019)
15. Hoyer, L., Dai, D., Gool, L.V.: Daformer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation. In: CVPR. IEEE (2022)
16. Hoyer, L., Dai, D., Gool, L.V.: Hrda: Context-Aware High-Resolution Domain-Adaptive Semantic Segmentation. In: ECCV (2022)
17. Hoyer, L., Dai, D., Wang, H., Gool, L.V.: Mic: Masked Image Consistency for Context-Enhanced Domain Adaptation. In: CVPR (2023)
18. Jiang, X., Lao, Q., Matwin, S., Havaei, M.: Implicit Class-Conditioned Domain Alignment for Unsupervised Domain Adaptation. In: ICML (2020)
19. Jin, Y., Wang, X., Long, M., Wang, J.: Minimum Class Confusion for Versatile Domain Adaptation. In: ECCV. pp. 464–480 (2020)
20. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised Learning with Deep Generative Models. In: NeurIPS. pp. 3581–3589 (2014)

21. Krizhevsky, A.: Learning multiple layers of features from tiny images (2009)
22. Lee, D.H.: Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In: ICMLw. vol. 3, p. 896 (2013)
23. Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H.P.: Pruning Filters for Efficient ConvNets. In: ICLR (2017)
24. Li, W., Liu, X., Yuan, Y.: Sigma: Semantic-complete Graph Matching for Domain Adaptive Object Detection. In: CVPR. pp. 5291–5300 (2022)
25. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: ECCV. pp. 740–755 (2014)
26. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased Teacher for Semi-Supervised Object Detection. In: ICLR (2021)
27. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional Adversarial Domain Adaptation. In: NeurIPS. pp. 1647–1657 (2018)
28. Luo, Y.W., Ren, C.X.: Conditional Bures Metric for Domain Adaptation. In: CVPR. pp. 13984–13993. IEEE (2021)
29. Luo, Y., Zhu, J., Li, M., Ren, Y., Zhang, B.: Smooth Neighbors on Teacher Graphs for Semi-Supervised Learning. In: CVPR. pp. 8896–8905. IEEE (2018)
30. Odena, A.: Semi-Supervised Learning with Generative Adversarial Networks. ICLRw (2016)
31. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment Matching for Multi-Source Domain Adaptation. In: ICCV. pp. 1406–1415 (2019)
32. Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., Saenko, K.: Visda: The Visual Domain Adaptation Challenge. arXiv **abs/1710.06924** (2017)
33. Rangwani, H., Aithal, S.K., Mishra, M., Jain, A., Radhakrishnan, V.B.: A Closer Look at Smoothness in Domain Adversarial Training. In: ICML. pp. 18378–18399 (2022)
34. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. PAMI **39**(6), 1137–1149 (2017)
35. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for Data: Ground Truth from Computer Games. In: ECCV. pp. 102–118. Springer International Publishing (2016)
36. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In: CVPR. IEEE (2016)
37. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-Weak Distribution Alignment for Adaptive Object Detection. In: CVPR. pp. 6956–6965 (2019)
38. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. In: CVPR. pp. 3723–3732 (2018)
39. Sakaridis, C., Dai, D., Gool, L.V.: Semantic Foggy Scene Understanding with Synthetic Data. IJCVR **126**(9), 973–992 (2018)
40. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: IJCVR. vol. 128, pp. 336–359. Springer Science and Business Media LLC (2020)
41. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In: NeurIPS (2020)
42. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: ICLR (2017)
43. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep Hashing Network for Unsupervised Domain Adaptation. In: CVPR. pp. 5385–5394 (2017)

44. Wang, Q., Bu, P., Breckon, T.P.: Unifying Unsupervised Domain Adaptation and Zero-Shot Visual Recognition. In: IJCNN. IEEE (2019)
45. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and Efficient Design for Semantic Segmentation with Transformers. In: NeurIPS. pp. 12077–12090 (2021)
46. Yosinski, J., Clune, J., Nguyen, A.M., Fuchs, T.J., Lipson, H.: Understanding Neural Networks Through Deep Visualization. arXiv **abs/1506.06579** (2015)
47. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling. In: NeurIPS. pp. 18408–18419 (2021)
48. Zhang, Y., Liu, T., Long, M., Jordan, M.I.: Bridging Theory and Algorithm for Domain Adaptation. In: ICML. pp. 7404–7413. PMLR (2019)
49. Zhu, X., Pang, J., Yang, C., Shi, J., Lin, D.: Adapting Object Detectors via Selective Cross-Domain Alignment. In: CVPR. pp. 687–696. IEEE (2019)