Milking CowMask for Semi-supervised Image Classification

Geoff French[1,2][a], Avital Oliver[1][b] and Tim Salimans[1][c]
[1]*Google Research, Brain Team, Amsterdam, The Netherlands*
[2]*School of Computing Sciences, University of East Anglia, Norwich, U.K.*

Keywords: Semi-supervised Learning, Image Classification, Deep Learning.

Abstract: Consistency regularization is a technique for semi-supervised learning that underlies a number of strong results for classification with few labeled data. It works by encouraging a learned model to be robust to perturbations on unlabeled data. Here, we present a novel mask-based augmentation method called *CowMask*. Using it to provide perturbations for semi-supervised consistency regularization, we achieve a competitive result on ImageNet with 10% labeled data, with a top-5 error of 8.76% and top-1 error of 26.06%. Moreover, we do so with a method that is much simpler than many alternatives. We further investigate the behavior of CowMask for semi-supervised learning by running many smaller scale experiments on the SVHN, CIFAR-10 and CIFAR-100 data sets, where we achieve results competitive with the state of the art, indicating that CowMask is widely applicable. We open source our code at https://github.com/google-research/google-research/tree/master/milking_cowmask.

1 INTRODUCTION

Training accurate deep neural network based image classifiers requires large quantities of training data. While images are often readily available in many problem domains, producing ground truth annotations is usually a laborious and expensive task that can act as a bottleneck. Semi-supervised learning offers the tantalising possibility of reducing the amount of annotated data required by learning from a dataset that is only partially annotated.

Semi-supervised learning algorithms based on consistency regularization (Sajjadi et al., 2016a; Laine and Aila, 2017; Oliver et al., 2018) have proved to be simple while effective, yielding a number of state of the art results over the last few years. Consistency regularization is driven by encouraging consistent predictions for unsupervised samples under stochastic augmentation. Using CutOut (DeVries and Taylor, 2017) – in which a rectangular region of an image is masked to zero – as the augmentation has proved to be highly effective, making significant contributions to the effectiveness of rich augmentation strategies (Xie et al., 2019; Sohn et al., 2020).

In this paper, we introduce a simple masking strategy that we call CowMask, whose shapes and appearance are more varied than the rectangular masks used by CutOut and RandErase (Zhong et al., 2020). When used to erase parts of an image in a similar fashion to RandErase, CowMask outperforms rectangular masks in the majority of semi-supervised image classification tasks that we tested.

We extend the Interpolation Consistency Training (ICT) algorithm (Verma et al., 2019) to use mask-based mixing, using both rectangular masks as in CutMix (Yun et al., 2019) and CowMask. Both CutMix and CowMask exhibit strong semi-supervised learning performance, with CowMask outperforming rectangular mask based mixing in the majority of cases. CowMask based mixing achieves semi-supervised image classification results that are comparable with the state-of-the-art on Imagenet and on multiple small image datasets, without the use of multi-stage training procedures or complex training objectives.

In Section 2 we discuss related work that forms the basis of our approach, alongside other semi-supervised learning algorithms for comparison. In Section 3 we present CowMask, the novel ingredient to our semi-supervised learning algorithm, that is described in Section 4. We present our experiments and results in Section 5. Finally we discuss our work and conclude in Section 7.

[a] https://orcid.org/0000-0003-2868-2237
[b] https://orcid.org/0000-0002-8301-0161
[c] https://orcid.org/0000-0002-7064-1474

French, G., Oliver, A. and Salimans, T.
Milking CowMask for Semi-supervised Image Classification.
DOI: 10.5220/0010773700003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 75-84
ISBN: 978-989-758-555-5; ISSN: 2184-4321

2 BACKGROUND

2.1 Semi-supervised Classification

A variety of semi-supervised deep neural network image classification approaches have been proposed over the last several years, including the use of auto-encoders (Wang et al., 2019; Rasmus et al., 2015), GANs (Salimans et al., 2016; Dai et al., 2017), curriculum learning (Cascante-Bonilla et al., 2020) and self-supervised learning (Zhai et al., 2019).

Many recent approaches are based on consistency regularization (Oliver et al., 2018), a simple approach exemplified by the π-model (Laine and Aila, 2017) and the Mean Teacher model (Tarvainen and Valpola, 2017). Two loss terms are minimized; standard cross-entropy loss and consistency loss for supervised and unsupervised samples respectively. Consistency loss measures the difference between predictions resulting from differently perturbed variants of an unsupervised sample. The π-model perturbs samples twice using stochastic augmentation and minimises the squared difference between class probability predictions. The Mean Teacher model builds on the π-model by using two networks; a teacher and a student. The student is trained using gradient descent as normal while the weights of the teacher are an exponential moving average of those of the student. The consistency loss term measures the difference in predictions between the student and the teacher under different stochastic augmentation.

A variety of types of perturbation have been explored. (Sajjadi et al., 2016b) employed richer data augmentation including affine transformations, while (Laine and Aila, 2017) and (Tarvainen and Valpola, 2017) used standard augmentation strategies such as random crop and noise for small image datasets. Virtual Adversarial Training (VAT) uses adversarial perturbations that maximise the consistency loss term.

2.1.1 More Recent Work

The following approaches were presented subsequent to the developement of the work that we describe in this paper.

Recent self-supervised methods – namely SimCLR (Chen et al., 2020) and TWIST (Wang et al., 2021) – have yielded strong semi-supervised classification results in a two step method consisting of self-supervised pre-training followed by supervised fine-tuning using the labelled subset of the training set.

CoMatch (Li et al., 2020) combines consistency regularization with self-supervised contrastive learning. Sample similarity between contrastive embeddings computed for unsupervised samples are used to compute weighted average pseudo-labels, thereby using similarity to other samples to improve the quality of the pseudo-label used as an unsupervised training target. Furthermore, agreement between a pseudo-label graph and a contrastive embedding similarity graph encourages clustering.

Meta Pseudo Labels (Pham et al., 2021) combines pseudo labelling – in which a teacher network predicts labels used to train a student – with meta-learning objectives that ensure that use the performance of the student on supervised samples to guide the training of the teacher.

2.2 Mixing Regularization

Recent works have demonstrated that blending pairs of images and corresponding ground truths can act as an effective regularizer. MixUp (Zhang et al., 2018) draws a blending factor from the Beta distribution that is used to interpolate images and ground truth labels. Interpolation Consistency Training (ICT) (Verma et al., 2019) extends this approach to work in a semi-supervised setting by combining it with the Mean Teacher model. The teacher network is used to predict class probabilities for a pair of images A and B and MixUp is used to blend the images and the teachers' predictions. The predictions of the student for the blended image are encouraged to be as close as possible to the blended teacher predictions.

MixMatch (Berthelot et al., 2019b) guesses labels for unsupervised samples by sharpening the averaged predictions from multiple rounds of standard augmentation and blends images and corresponding labels (ground truth for supervised samples, guesses for unsupervised) using MixUp (Zhang et al., 2018). The blended images and corresponding guessed labels are used to compute consistency loss.

2.3 Rich Augmentation

AutoAugment (Cubuk et al., 2019a) and RandAugment (Cubuk et al., 2019b) are rich augmentation schemes that combine a number of image operations provided by the Pillow library (Lundh et al.,). AutoAugment learns an augmentation policy for a specific dataset using re-inforcement learning, requiring a large amount of computation to do so. RandAugment on the other hand has two hyper-parameters that are chosen via grid search; the number of operations to apply and a magnitude.

Unsupervised data augmentation (UDA) (Xie et al., 2019) adds employs a combination of CutOut (DeVries and Taylor, 2017) and RandAugment (Cubuk et al., 2019b) in a semi-supervised setting achieving state-of-the-art results in small image benchmarks such as CIFAR-10. Their approach encourages consistency between the predictions for the original un-modified image and the same image with RandAugment applied.

ReMixMatch (Berthelot et al., 2019a) builds on MixMatch by adding distribution alignment and rich data augmentation using CTAugment or RandAugment (depending on the dataset). CTAugment is a variant of AutoAugment that learns an augmentation policy during training, and RandAugment is a pre-defined set of 15 forms of augmentations with concrete scales. It is worth noting that ReMixMatch uses predictions from standard 'weak' augmentation as guessed target probabilities for unsupervised samples and encourages predictions arising from multiple applications of the richer CTAugment to be close to the guessed target probabilities. The authors found that using rich augmentation for guessing target probabilities (a la MixMatch) resulted in unstable training.

FixMatch (Sohn et al., 2020) is a simple semi-supervised learning approach that uses standard 'weak' augmentation to predict pseudo-labels for unsupervised samples. The same samples are richly augmented using CTAugment and cross-entropy loss is computed using the pseudo-labels. Confidence thresholding (French et al., 2018) masks the unsupervised cross-entropy loss to zero for samples whose predicted confidence is below 95%.

2.4 Mask-based Regularization

Erasing a rectangular region of an image by replacing it with zeros – as in Cutout (DeVries and Taylor, 2017) – or noise – as in RandErase (Zhong et al., 2020) – has proved to be an effective augmentation strategy that yields improvements in supervised classification.

Cutout has proved to be highly effective in semi-supervised classification scenarios. The UDA authors (Xie et al., 2019) report impressive results, while the FixMatch authors (Sohn et al., 2020) report that CutOut alone is as effective as the combination of the other 14 image operations used in CTAugment.

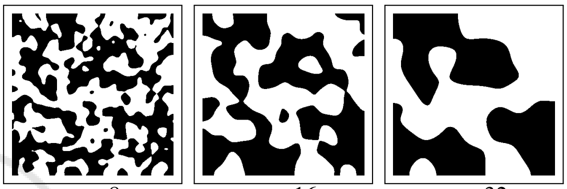
CutMix (Yun et al., 2019) replaces the blending factor in MixUp with a rectangular mask and uses it to mix pairs of images, effectively cutting and pasting a rectangle from one image onto another. This yielded significant supervised classification performance gains.

(French et al., 2020) analyzed semantic segmentation problems, finding that they exhibit a challenging data distribution where the cluster assumption – identified in prior work (Luo et al., 2018; Sajjadi et al., 2016a; Shu et al., 2018; Verma et al., 2019) as important to the success of consistency regularization – does not apply. They experiment with a variety of regularizers, obtaining strong results when using CutMix, suggesting mask-based mixing as a promising avenue for semi-supervised learning.

3 CowMask

Here, we propose CowMask; a simple approach to generating flexibly shaped masks, so called due to its' Friesian cow-like appearance. Example CowMasks are shown in Figure 1.

We note that the concurrent work FMix (Harris et al., 2020) uses an inverse Fourier transform to generate masks with a similar visual appearance.



$\sigma = 8$ $\sigma = 16$ $\sigma = 32$

Figure 1: Example CowMasks with $p = 0.5$ and varying σ.

Briefly, a CowMask is generated by applying Gaussian filtering of scale σ to normally distributed noise. A threshold τ is chosen such that a proportion p of the smooth noise pixels are below τ. Pixels with a value below τ are assigned a value of 1, or 0 otherwise. The scale of the mask features is controlled by σ – as seen in the examples in Figure 1 – and is drawn from a log-uniform distribution in the range $(\sigma_{min}, \sigma_{max})$. The proportion p of pixels with a value of 1 is drawn from a uniform distribution in the range (p_{min}, p_{max}). The procedure for generating a CowMask is provided in Algorithm 1.

Algorithm 1: CowMask generation algorithm. See Figure 1 for example output.

Require: mask size $H \times W$
Require: scale range $(\sigma_{min}, \sigma_{max})$
Require: proportion range (p_{min}, p_{max})
Require: inverse error function erf^{-1}
$\sigma \sim \log\mathcal{U}(\sigma_{min}, \sigma_{max})$ {Randomly choose sigma}
$p \sim \mathcal{U}(p_{min}, p_{max})$ {Randomly choose proportion}
$\mathbf{x} \sim \mathcal{N}^{H\times W}(0,1)$ {Per-pixel Gaussian noise}
$\mathbf{x}_s = \text{gaussian\_filter\_2d}(x, \sigma)$ {Filter noise}
$m = \text{mean}(\mathbf{x}_s)$ {Compute mean and std-dev}
$s = \text{std\_dev}(\mathbf{x}_s)$
$\tau = m + \sqrt{2} \cdot \text{erf}^{-1}(2p-1) \cdot s$ {Compute threshold}

$\mathbf{c} = \mathbf{x}_s \leq \tau$ {Threshold filtered noise}
Return c

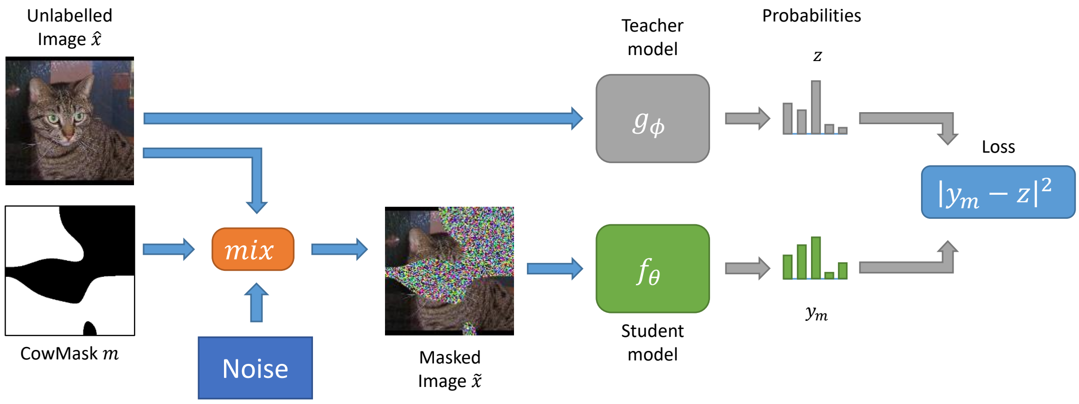


Figure 2: Illustration of the unsupervised mask based erasure consistency loss component of semi-supervised image classification. Blue arrows carry image or mask content and grey arrows carry probability vectors. Note that confidence thresholding is not illustrated here.

4 SEMI-SUPERVISED LEARNING METHOD

We adopt the Mean Teacher (Tarvainen and Valpola, 2017) framework as the basis of our approach. We use two networks; the student $f_\theta(\cdot)$ and the teacher $g_\phi(\cdot)$, both of which predict class probability vectors. The student is trained by gradient descent as normal. After every update to the student, the weights of the teacher are updated to be an exponential moving average of those of the student using $\phi' = \phi\alpha + \theta(1-\alpha)$. The momentum α controls the trade-off between the stability and the speed at which the teacher follows the student.

Our training set consists of a set of supervised samples S consisting of input images s and corresponding target labels t, and a set of unsupervised samples U consisting only of input images u. Given a labelled dataset we select the supervised subset randomly such that it maintains the class balance of the overall dataset[1] as is standard practice in the literature. All available samples are used as unsupervised samples. Our models f_θ are then trained to minimize a combined loss:

$$L = L_{\mathcal{S}}(f_\theta(s), t) + \omega L_{\mathcal{U}}(f_\theta(u), g_\phi(u))$$

where we use standard cross entropy loss for the supervised loss $L_{\mathcal{S}}(\cdot)$ and consistency loss for the unsupervised loss $L_{\mathcal{U}}(\cdot)$ that is modulated by the unsupervised loss weight ω.

We explore two different types of mask-based consistency regularization: *mask-based erasure* and *mask-based mixing*. In mask-based erasure we perturb our input data by erasing the part of the input image corresponding to a randomly sampled mask. In mask-based mixing we blend two input images together, with the blending weights given by the sampled mask. We follow the nomenclature of Cutout and CutMix, using the terms CowOut and CowMix to refer to CowMask based erasure and mixing respectively.

4.1 Mask-based Augmentation by Erasure

Mask-based erasure can function as an augmentation that can be added to the standard augmentation scheme used for the dataset at hand, with one caveat. Similar to prior work (Xie et al., 2019; Berthelot et al., 2019a; Sohn et al., 2020) we found it necessary to split our augmentation into a 'weak' standard augmentation scheme (e.g. crop and flip) and a 'strong' rich scheme; RandAugment in the case of the prior works mentioned or CowOut in our work. Weakly augmented samples are passed to the teacher network, generating predictions that are used as pseudo-targets that the student is encouraged to match for strongly augmented variants of the same samples. Using 'strong' erasure augmentation to generate pseudo-targets resulted in unstable training.

The π-model (Laine and Aila, 2017) and the Mean Teacher model (Tarvainen and Valpola, 2017) both use a Gaussian ramp-up function to modulate the effect of consistency loss during the early stages of training. Reinforcing the random predictions of an untrained network was found to harm performance. In place of a ramp-up we opt to use confidence thresholding (French et al., 2018). Consistency loss is masked to zero for samples for which the teacher networks' predictions are below a specified threshold. FixMatch (Sohn et al., 2020) uses confidence thresholding for similar reasons.

Our procedure for computing unsupervised consis-

[1] We use `StratifiedShuffleSplit` from Scikit-Learn (Buitinck et al., 2013)

tency loss based on erasure is provided in Algorithm 2 and is illustrated in Figure 2. For our small image experiments we found that the best value for the unsupervised weight factor ω is 1.

Algorithm 2: CowOut: erasure-based unsupervised loss.

Require: unlabeled image $\mathbf{x}$, CowMask $\mathbf{m}$
Require: teacher model g_ϕ
Require: student model f_θ
Require: confidence threshold ψ
$\hat{\mathbf{x}} = \text{std\_aug}(\mathbf{x})$ {standard augmentation}
$\mathbf{z} = \text{stop\_gradient}(g_\phi(\hat{\mathbf{x}}))$ {teacher pred.}
$q = \max_i \mathbf{z}[i] \geq \psi$ {confidence mask}
$\varepsilon \sim N(0, I)$ {generate noise image}
$\hat{\mathbf{x}}_m = \hat{\mathbf{x}} * \mathbf{m} + \varepsilon * (1 - \mathbf{m})$ {apply mask}
$\mathbf{y}_m = f_\theta(\hat{\mathbf{x}}_m)$ {student prediction}
$d = q * ||\mathbf{y}_m - \mathbf{z}||_2^2$ {cons. loss}
Return d

4.2 Mask-based Mixing

Alternatively, we can construct an unsupervised consistency loss by mask-based *mixing* of images in place of erasure. Our approach for mixing image pairs using masks is essentially that of Interpolation Consistency Training (ICT) (Verma et al., 2019). ICT works by passing the original image pair to the teacher network, the blended image to the student and encourages the students' prediction to match the blended teacher predictions. Where ICT draws per-pair blending factors a beta distribution, we mix images using a mask, and mix probability predictions with the mean of the mask (the proportion of pixels with a value of 1).

Confidence thresholding required adaptation for use with mix-based regularization. Rather than applying confidence thresholding to the blended teacher probability predictions we opted to blend the confidence values before thresholding as this gave slightly better results. Further improvements resulted from modulating the consistency loss by the proportion of samples in the batch whose predictions cross the confidence threshold, rather masking the loss for each sample individually.

The procedure for computing unsupervised mix consistency loss is provided in Algorithm 3 and illustrated in Figure 3. We found that a higher weight ω was appropriate for mix consistency loss; we used a value of 30 for our small image experiments.

Algorithm 3: CowMix: mixing-based unsupervised loss.

Require: unlabeled images $\mathbf{x}_a$, $\mathbf{x}_b$
Require: CowMask $\mathbf{m}$
Require: teacher model g_ϕ
Require: student model f_θ
Require: confidence threshold ψ
$\hat{\mathbf{x}}_a = \text{std\_aug}(\mathbf{x}_a)$ {standard augmentation}
$\hat{\mathbf{x}}_b = \text{std\_aug}(\mathbf{x}_b)$
$\mathbf{z}_a = \text{stop\_gradient}(g_\phi(\hat{\mathbf{x}}_a))$ {teacher pred.}
$\mathbf{z}_b = \text{stop\_gradient}(g_\phi(\hat{\mathbf{x}}_b))$
$c_a = \max_i \mathbf{z}_a[i]$ {confidence of prediction}
$c_b = \max_i \mathbf{z}_b[i]$
$\hat{\mathbf{x}}_m = \hat{\mathbf{x}}_a * \mathbf{m} + \hat{\mathbf{x}}_b * (1 - \mathbf{m})$ {mix images}
$p = \text{mean}(\mathbf{m})$ {scalar mean of mask}
$\mathbf{z}_m = \mathbf{z}_a * p + \mathbf{z}_b * (1 - p)$ {mix tea. preds.}
$c_m = c_a * p + c_b * (1 - p)$ {mix confidences}
$q = \text{mean}(c_m \geq \psi)$ {mean of conf. mask}
$\mathbf{y}_m = f_\theta(\hat{\mathbf{x}}_m)$ {stu. pred. on mixed image}
$d = q||\mathbf{y}_m - \mathbf{z}_m||_2^2$ {cons. loss}
Return d

5 EXPERIMENTS AND RESULTS

We first evaluate CowMix for semi-supervised consistency regularization on the challenging ImageNet dataset, where we are competitive with the state of the art. Next, we examine CowOut and CowMix further and compare with previously proposed methods by trying multiple versions of our approach combined with multiple models on three small image datasets: CIFAR-10, CIFAR-100 and SVHN. The training regimes used for both ImageNet and the small image datasets are sufficiently similar that we used the same codebase for all of our experiments.

Our results are obtained by using the teacher network for evaluation. We report our results as error rates presented as the mean $\pm$ 1 standard deviation computed from the results of 5 runs, each of which uses a different subset of samples as the supervised set. Supervised sets are consistent for all experiments for a given dataset and number of supervised samples.

5.1 ImageNet 2012

We contrast the following scenarios: a supervised baseline using 10% of the dataset, semi-supervised training with the same 10% of labelled examples using CowMix consistency regularization on all unlabeled examples, and fully supervised training with all 100% labels.

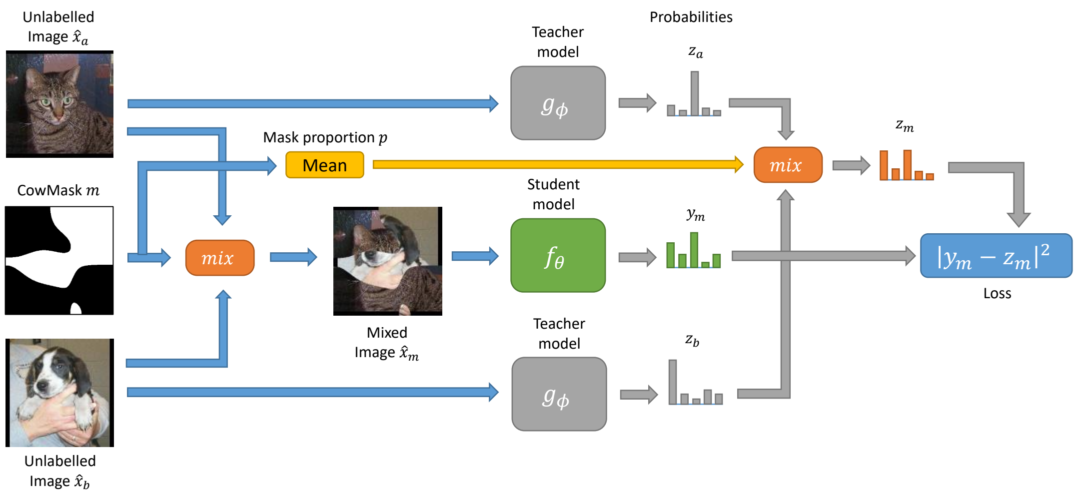


Figure 3: Illustration of the unsupervised masked based mixing loss component of semi-supervised image classification. Blue arrows carry image or mask content, grey arrows carry probability vectors and yellow carry scalars. Please note that confidence thresholding is not illustrated here.

5.1.1 Setup

We used the ResNet-152 architecture. We adopted a training regime as similar as possible to a standard ImageNet ResNet training protocol. We used a batch size of 1024 and SGD with Nesterov Momentum (Sutskever et al., 2013) set to 0.9 and weight decay (via L2 regularization) set to 0.00025. Our standard augmentation scheme consists of inception crop, random horizontal flip and colour jitter, as in (Tarvainen and Valpola, 2017).We found that the standard learning rate of 0.1 resulted in unstable training, but were able to stabilise it by reducing the learning rate to 0.04 (Tarvainen and Valpola, 2017). We found that our approach benefits from training for longer than in supervised settings, so we doubled the number of training epochs to 180 and stretched the learning rate schedule by a factor of 2, reducing the learning rate at epochs 60, 120 and 160 and reduced it by a factor of 0.2 rather than 0.1. We used a teacher EMA momentum α of 0.999.

We obtained our CowMix results using a mix loss weight of 100 and and a confidence threshold of 0.5. We drew the CowMask σ scale parameter from the range $(32, 128)$.

5.1.2 Results

Our ImageNet results are presented in Table 1. The CoMatch (Li et al., 2020) and Meta Pseudo Labels (Pham et al., 2021) approaches (both more recent than our CowMix work) uses the smaller ResNet-50 architecture and are able beat our top-5 error result and are slightly behind our top-1 error result. We match the S^4L MOAM (Zhai et al., 2019) top-5 error result and beat their top-1 error result, with a simple end-to-end approach and a significantly smaller model. By comparison the S^4L MOAM result is obtained using a 3-stage training and fine-tuning procedure. Recent self-supervised approaches have achieved impressive semi-supervised results on ImageNet by first training a model self-supervised fashion followed by fune-tuning using a subset of the labelled data. The recent SimCLR (Chen et al., 2020) approach (concurrent work) beats our result when using a much larger model. The more recent TWIST (Wang et al., 2021) approach beats our result using a double-width ResNet-50 that has only 50% more parameters than the ResNet-152 that we use. We tested our approach with wider models (e.g. ResNet-50$\times$2) but obtained our best results from the deeper and commonly used ResNet-152.

5.2 Small Image Experiments

Alongside CowOut and CowMix we implemented and evaluated Mean Teacher, CutOut/RandErase and CutMix, and we compare our method against these using the CIFAR-10, CIFAR-100, and SVHN datasets.

We note the following differences between our implementation and those of CutOut and CutMix: 1. Our boxes are chosen so that they entirely fit within the bounds of the mask, whereas CutOut and CutMix use a fixed or random size respectively and centre the box anywhere within the mask, with some of the box potentially being outside the bounds of the mask. 2. CutOut uses a fixed size box, CutMix randomly chooses an area but constrains the aspect ratio to be that of the mask, we choose both randomly.

Table 1: Results on ImageNet with 10% labels. Note that S^4L involves three steps with different training procedures, while CowMix involves a single training run. SimCLR is able to beat CowMix, but only when using a very large model.

Approach	Architecture	Params.	Top-5 err.	Top-1 err.
	Our baselines			
Sup 10%	ResNet-152	60M	22.12%	42.91%
Sup 100%	ResNet-152	60M	5.67%	21.33%
	Other work: self-supervised pre-training then fine-tune			
SimCLR (Chen et al., 2020)	ResNet-50	24M	12.2%	34.4%
SimCLR	ResNet-50×2	94M	8.8%	28.3%
SimCLR	ResNet-50×4	375M	7.4%	25.6%
TWIST (Wang et al., 2021)	ResNet-50	24M	9.0%.	28.3%
TWIST	ResNet-50×2	94M	**7.2%**	**24.7%**
	Other work: semi-supervised			
Mean Teacher (Tarvainen and Valpola, 2017)	ResNeXt-152	62M	9.11% $\pm$ 0.12	–
UDA (Xie et al., 2019)	ResNet-50	24M	11.2%	31.22%
FixMatch (Sohn et al., 2020)	ResNet-50	24M	10.87 $\pm$ 0.28%	28.54 $\pm$ 0.52%
S^4L Full (MOAM) (Zhai et al., 2019)	ResNet-50×4	375M	8.77%	26.79%
CoMatch (Li et al., 2020)	ResNet-50	24M	8.4%	26.4%
Meta Pseudo Labels (Pham et al., 2021)	ResNet-50	24M	8.62%	26.11%
	Our results			
CowMix	ResNet-152	60M	8.76 $\pm$ 0.07%	26.06 $\pm$ 0.17%

5.2.1 Setup

For the small image experiments we use a 27M parameter Wide ResNet 28-96x2d with shake-shake regularization (Gastaldi, 2017). We note that as a result of a mistake in our implementation we used a 3×3 convolution rather than a 1×1 in the residual shortcut connections that either down-sample or change filter counts, resulting in a slightly higher parameter count.

The standard Wide ResNet training regime (Zagoruyko and Komodakis, 2016) is very similar to that used for ImageNet. We used the optimizer, but with weight decay of 0.0005 and a batch size of 256. As before, the standard learning rate of 0.1 had to be reduced to ensure stability, this time to 0.05. The small image experiments also benefit from training for longer; 300 epochs instead of the standard 200 used in supervised settings. The adaptations made to the Wide ResNet learning rate schedule were nearly identical to those made to the ImageNet schedule. We doubled its length and reduced the learning rate by a factor of 0.2 rather than 0.1. We did however remove the last step; the learning rate is reduced at epochs 120 and 240 rather than epochs 60, 120 and 160 as used in supervised settings. For erasure experiments we used a teacher EMA momentum α of 0.99 and for mixing experiments we used 0.97.

When using CowOut and CowMix we obtained the best results when the CowMask scale parameter σ is drawn from the range $(4, 16)$. We note that this corresponds to a range of $(\frac{1}{8}, \frac{1}{2})$ relative to the 32×32 image size and that the σ range used in our ImageNet experiments bears a nearly identical relationship to the 224×224 image size used there. For erasure experiments using CowOut we obtained the best results when drawing p; the proportion of pixels that are retained from the range $(0.25, 1)$. Intuitively it makes sense to retain at least 25% of the image pixels as encouraging the network to predict the same result for an image and a blank space is unlikely to be useful. For mixing experiments using CowMix we obtained the best results when drawing p from the range $(0.2, 0.8)$.

We performed hyper-parameter tuning on the CIFAR-10 dataset using 1,000 supervised samples and evaluating on 5,000 training samples held out as a validation set. The best hyper-parameters found were used as-is for CIFAR-100 and SVHN.

5.2.2 Results

Our results for CIFAR-10, CIFAR-100 and SVHN datasets are presented in Tables 2, 4 and 3 respectively. Considering the techniques we explore we find that mix-based regularization outperforms erasure based regularization, irrespective of the mask generation method used.

We would like to note that our 27M parameter model is larger than the 1.5M parameter models used for the majority of results in other works, so we cannot make an apples-to-apples comparison in these cases. Our CIFAR-10 results are competitive with recent work, except in small data regimes of less than 500 samples where EnAET (Wang et al., 2019) and Fix-

Table 2: Results on CIFAR-10 test set, error rates as $mean \pm std - dev$ of 5 independent runs.

Labeled samples	40	50	100	250	500	1000	2000	4000	ALL
Other work: uses smaller Wide ResNet 28-2 model with 1.5M parameters									
EnAET		**16.45%**	**9.35%**	7.6% $\pm$ 0.34	7.27%	6.95%	6.0%	5.35%	
UDA				8.76% $\pm$ 0.90	6.68% $\pm$ 0.24	5.87% $\pm$ 0.13	5.51% $\pm$ 0.21	5.29% $\pm$ 0.25	
MixMatch				11.08% $\pm$ 0.87	9.65% $\pm$ 0.97	7.75% $\pm$ 0.32	7.03% $\pm$ 0.15	6.24% $\pm$ 0.06	
ReMixMatch	14.98% $\pm$ 3.38			6.27% $\pm$ 0.34		5.73% $\pm$ 0.16		5.14% $\pm$ 0.04	
FixMatch (RA)	**13.81% $\pm$ 3.37**			**5.07% $\pm$ 0.65**				4.26% $\pm$ 0.05	
Other work: uses 26M parameter models									
EnAET								4.18% $\pm$ 0.04	**1.99%**
UDA								3.7% / **2.7%**	
MixMatch								4.95% $\pm$ 0.08	
Our results: uses 27M parameter Wide ResNet 28-96x2d with shake-shake									
Supervised		76.01% $\pm$ 1.53	69.74% $\pm$ 2.09	58.41% $\pm$ 1.60	47.12% $\pm$ 1.78	36.61% $\pm$ 1.11	24.53% $\pm$ 0.80	14.81% $\pm$ 0.43	3.57% $\pm$ 0.09
	Augmentation / erasure based regularization								
Mean teacher		75.68% $\pm$ 3.72	67.77% $\pm$ 4.17	47.95% $\pm$ 4.52	29.72% $\pm$ 5.74	14.14% $\pm$ 0.56	8.79% $\pm$ 0.16	6.92% $\pm$ 0.15	3.04% $\pm$ 0.07
RandErase		74.67% $\pm$ 2.13	62.86% $\pm$ 3.61	37.63% $\pm$ 7.20	19.22% $\pm$ 3.34	11.87% $\pm$ 0.73	7.05% $\pm$ 0.14	5.27% $\pm$ 0.17	2.59% $\pm$ 0.10
CowOut		72.55% $\pm$ 3.80	56.72% $\pm$ 3.90	28.45% $\pm$ 7.03	14.00% $\pm$ 1.84	8.98% $\pm$ 1.11	6.27% $\pm$ 0.40	4.97% $\pm$ 0.12	2.50% $\pm$ 0.10
	Mix based regularization								
ICT		80.08% $\pm$ 2.57	72.96% $\pm$ 4.46	44.92% $\pm$ 7.85	17.10% $\pm$ 2.15	10.40% $\pm$ 0.63	7.75% $\pm$ 1.23	5.97% $\pm$ 0.11	3.45% $\pm$ 0.06
CutMix		66.06% $\pm$ 15.82	34.05% $\pm$ 6.19	9.01% $\pm$ 3.60	6.81% $\pm$ 1.04	5.44% $\pm$ 0.39	4.62% $\pm$ 0.15	4.11% $\pm$ 0.19	2.78% $\pm$ 0.14
CowMix		**55.46% $\pm$ 15.23**	23.00% $\pm$ 3.95	7.56% $\pm$ 0.94	**5.34% $\pm$ 0.80**	**4.73% $\pm$ 0.37**	**4.13% $\pm$ 0.16**	3.61% $\pm$ 0.07	2.56% $\pm$ 0.06

Table 3: Results on SVHN test set, error rates as $mean \pm stdev$ of 5 independent runs.

Labeled samples	40	100	250	500	1000	2000	4000	ALL
Other work: uses smaller Wide ResNet 28-2 model with 1.5M parameters								
EnAET		16.92%	**3.21% $\pm$ 0.21**	**3.05%**	2.92%	**2.84%**	2.69%	
UDA					2.55% $\pm$ 0.99			
MixMatch			3.78% $\pm$ 0.26	3.64% $\pm$ 0.46	3.27% $\pm$ 0.31	3.04% $\pm$ 0.13	2.89% $\pm$ 0.06	
ReMixMatch	**3.55% $\pm$ 3.87**	3.10% $\pm$ 0.50		2.83% $\pm$ 0.30		**2.42% $\pm$ 0.09**		
FixMatch (RA)	3.96% $\pm$ 2.17		**2.48% $\pm$ 0.38**		**2.28% $\pm$ 0.11**			
Other work: uses 26M parameter models								
EnAET					2.42%			
Our results: uses 27M parameter Wide ResNet 28-96x2d with shake-shake								
Supervised		71.24% $\pm$ 5.40	37.02% $\pm$ 6.15	18.85% $\pm$ 1.49	11.71% $\pm$ 0.55	8.23% $\pm$ 0.38	6.01% $\pm$ 0.46	2.82% $\pm$ 0.08
	Augmentation / erasure based regularization							
Mean teacher		62.16% $\pm$ 10.92	8.23% $\pm$ 4.62	3.84% $\pm$ 0.15	3.75% $\pm$ 0.10	3.61% $\pm$ 0.15	3.47% $\pm$ 0.12	2.73% $\pm$ 0.04
RandErase		52.55% $\pm$ 22.03	7.61% $\pm$ 1.71	6.17% $\pm$ 1.25	4.81% $\pm$ 0.46	3.66% $\pm$ 0.15	3.21% $\pm$ 0.22	2.36% $\pm$ 0.04
CowOut		66.66% $\pm$ 19.71	12.11% $\pm$ 1.82	5.94% $\pm$ 0.38	4.36% $\pm$ 0.29	3.59% $\pm$ 0.25	3.04% $\pm$ 0.04	2.42% $\pm$ 0.09
	Mix based regularization							
CutMix		**9.54% $\pm$ 2.53**	5.62% $\pm$ 0.93	4.32% $\pm$ 0.52	3.79% $\pm$ 0.41	3.26% $\pm$ 0.27	2.92% $\pm$ 0.09	2.29% $\pm$ 0.09
CowMix		9.73% $\pm$ 4.01	3.59% $\pm$ 0.30	3.80% $\pm$ 0.32	3.72% $\pm$ 0.60	3.13% $\pm$ 0.11	2.90% $\pm$ 0.19	2.18% $\pm$ 0.06

Match (Sohn et al., 2020) outperform CowMix. Our CIFAR-100 and SVHN results are competitive with recent approaches but are not state of the art. We note that we did not tune our hyper-parameters for these datasets.

6 DISCUSSION

We explain the effectiveness of CowMix by considering the effects of CowMask and mixing based semi-supervised learning separately.

(DeVries and Taylor, 2017) established that Cutout – that uses a box shaped mask similar to RandErase - encourages the network to utilise a wider variety of features in order to overcome the varying combinations of parts of an image being present or masked out. In comparison to a rectangular mask the more flexibly shaped CowMask provides more variety and has less correlation between regions of the mask. Increasing the range of combinations of image regions being left intact or erased enhances its effectiveness.

The MixUp (Zhang et al., 2018) and CutMix (Yun et al., 2019) regularizers demonstrated that encouraging network predictions vary smoothly between two images as they are mixed – using either interpolation or mask-based mixing – improved supervised performance, with mask-based mixing offering the biggest gains. We adapted CutMix – in a similar fashion to ICT – for semi-supervised learning and showed that mask based mixing yields significant gains when used as an unsupervised regularizer. CowMix adds the benefits of flexibly shaped masks into the mix.

Table 4: Results on CIFAR-100 test set, error rates as $mean \pm stdev$ of 5 independent runs.

# Labels	1000	5000	10000	ALL
Other work: uses 1.5M parameters Wide ResNet 28-2				
EnAET	58.73%	31.83%	26.93% $\pm$ 0.21	20.55%
MixMatch			25.88% $\pm$ 0.30	
FixMatch			**22.60% $\pm$ 0.12**	
Other work: uses 26M parameter models				
EnAET			22.92%	16.87%
Our results: 27M param WRN 28-96x2d				
Supervised	78.80% $\pm$ 0.22	49.24% $\pm$ 0.40	36.04% $\pm$ 0.26	18.82% $\pm$ 0.22
	Augmentation / erasure based regularization			
Mean teacher	76.97% $\pm$ 0.99	38.90% $\pm$ 0.48	30.04% $\pm$ 0.60	17.81% $\pm$ 0.17
RandErase	70.48% $\pm$ 1.05	35.61% $\pm$ 0.40	28.21% $\pm$ 0.16	16.71% $\pm$ 0.29
CowOut	68.86% $\pm$ 0.78	38.82% $\pm$ 0.44	27.54% $\pm$ 0.29	16.46% $\pm$ 0.22
	Mix based regularization			
CutMix	64.11% $\pm$ 2.63	30.15% $\pm$ 0.58	24.08% $\pm$ 0.25	16.54% $\pm$ 0.18
CowMix	**57.27% $\pm$ 1.34**	**29.25% $\pm$ 0.47**	23.61% $\pm$ 0.30	**15.73% $\pm$ 0.15**

7 CONCLUSIONS

We presented and evaluated *CowMask* for use in semi-supervised consistency regularization, achieving a result competitive with the state of the art on semi-supervised Imagenet, with a much simpler method than in previously proposed approaches, using standard networks and training procedures. We examined both erasure-based and mixing-based augmentation using CowMask, and find that the mix-based variant – which we call *CowMix* – is particularly effective for semi-supervised learning. Further experiments on small image data sets SVHN, CIFAR-10, and CIFAR-100 demonstrate that CowMask is widely applicable.

Research on semi-supervised learning is moving fast, and many new approaches have been proposed over the last year alone that use mask-based perturbation. In future work we would like to further explore the use of CowMask in combination with these other recently proposed methods.

REFERENCES

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., and Raffel, C. (2019a). Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. (2019b). Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5050–5060.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Cascante-Bonilla, P., Tan, F., Qi, Y., and Ordonez, V. (2020). Curriculum labeling: Self-paced pseudo-labeling for semi-supervised learning. *arXiv preprint arXiv:2001.06001*.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. (2019a). Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 113–123.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (2019b). Randaugment: Practical data augmentation with no separate search. *arXiv preprint arXiv:1909.13719*.

Dai, Z., Yang, Z., Yang, F., Cohen, W. W., and Salakhutdinov, R. R. (2017). Good semi-supervised learning that requires a bad gan. In *Advances in neural information processing systems*, pages 6510–6520.

DeVries, T. and Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552.

French, G., Laine, S., Aila, T., Mackiewicz, M., and Finlayson, G. (2020). Semi-supervised semantic segmentation needs strong, varied perturbations. In *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press.

French, G., Mackiewicz, M., and Fisher, M. (2018). Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*.

Gastaldi, X. (2017). Shake-shake regularization. *arXiv preprint arXiv:1705.07485*.

Harris, E., Marcu, A., Painter, M., Niranjan, M., Prügel-Bennett, A., and Hare, J. (2020). Understanding and enhancing mixed sample data augmentation. *arXiv preprint arXiv:2002.12047*.

Laine, S. and Aila, T. (2017). Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*.

Li, J., Xiong, C., and Hoi, S. C. (2020). Semi-supervised learning with contrastive graph regularization. *arXiv preprint arXiv:2011.11183*.

Lundh, F., Clark, A., et al. Pillow.

Luo, Y., Zhu, J., Li, M., Ren, Y., and Zhang, B. (2018). Smooth neighbors on teacher graphs for semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8896–8905.

Oliver, A., Odena, A., Raffel, C., Cubuk, E. D., and Goodfellow, I. J. (2018). Realistic evaluation of semi-supervised learning algorithms. In *International Conference on Learning Representations*.

Pham, H., Dai, Z., Xie, Q., and Le, Q. V. (2021). Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., and Raiko, T. (2015). Semi-supervised learning with ladder networks. In *Advances in neural information processing systems*, pages 3546–3554.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. (2016a). Mutual exclusivity loss for semi-supervised deep learning. In *23rd IEEE International Conference on Image Processing, ICIP 2016*.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. (2016b). Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 1163–1171.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242.

Shu, R., Bui, H., Narui, H., and Ermon, S. (2018). A DIRT-T approach to unsupervised domain adaptation. In *International Conference on Learning Representations*.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147.

Tarvainen, A. and Valpola, H. (2017). 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.

Verma, V., Lamb, A., Kannala, J., Bengio, Y., and Lopez-Paz, D. (2019). Interpolation consistency training for semi-supervised learning. *CoRR*, abs/1903.03825.

Wang, F., Kong, T., Zhang, R., Liu, H., and Li, H. (2021). Self-supervised learning by estimating twin class distributions. *arXiv preprint arXiv:2110.07402*.

Wang, X., Kihara, D., Luo, J., and Qi, G.-J. (2019). EnAET: Self-trained ensemble autoencoding transformations for semi-supervised learning. *arXiv preprint arXiv:1911.09265*.

Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., and Le, Q. V. (2019). Unsupervised data augmentation. *arXiv preprint arXiv:1904.12848*.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. In Richard C. Wilson, E. R. H. and Smith, W. A. P., editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press.

Zhai, X., Oliver, A., Kolesnikov, A., and Beyer, L. (2019). S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1476–1485.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. (2020). Random erasing data augmentation.