Dense FixMatch: a simple semi-supervised learning method for pixel-wise prediction tasks

Miquel Martí i Rabadán[1,2], Alessandro Pieropan[2], Hossein Azizpour[1], and Atsuto Maki[1]

[1]KTH Royal Institute of Technology, Stockholm, Sweden
[2]Univrses AB, Stockholm, Sweden

Abstract

We propose Dense FixMatch, a simple method for online semi-supervised learning of dense and structured prediction tasks combining pseudo-labeling and consistency regularization via strong data augmentation. We enable the application of FixMatch in semi-supervised learning problems beyond image classification by adding a matching operation on the pseudo-labels. This allows us to still use the full strength of data augmentation pipelines, including geometric transformations.

We evaluate it on semi-supervised semantic segmentation on Cityscapes and Pascal VOC with different percentages of labeled data and ablate design choices and hyper-parameters. Dense FixMatch significantly improves results compared to supervised learning using only labeled data, approaching its performance with 1/4 of the labeled samples.

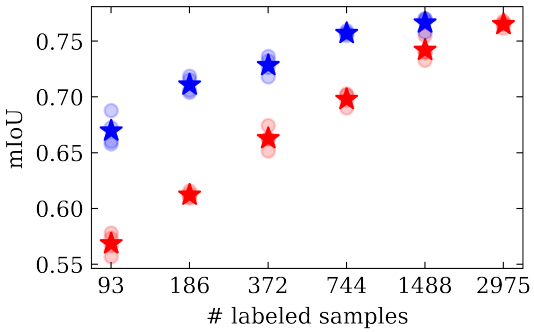

Figure 1: Dense FixMatch (blue) on unlabeled data improves the performance of semi-supervised semantic segmentation on Cityscapes `val` set using DeepLabv3+ with ResNet-101 backbone over supervised baselines (red) across different amounts of labeled samples. $\star$ represents the mean over four different runs with random labeled data splits. Results for individual runs are shown with circles.

1 Introduction

Semi-supervised learning (SSL) has shown great potential to reduce the annotation costs of training deep learning models. Modern methods achieve competitive results at a fraction of the amount of annotated samples required for standard supervised learning [1, 2, 25]. The potential cost savings are even larger for structured or dense prediction tasks, such as object detection, instance or semantic segmentation since the annotation cost for such tasks is much larger than for image classification.

However, SSL methods have been mainly developed and studied with image-level classification in mind [27, 19, 30, 1, 25]. Only more recently, methods have appeared adapting or proposing solutions to structured or dense tasks such as object detection [37, 13, 26, 16, 31] or semantic segmentation [9, 20, 14, 38, 5, 12]. Still, most works have focused on improving performance on specific tasks and not aimed at finding methods that could be applied to different tasks. Only a handful of methods are generic enough to be used for multiple tasks with no or few changes [9, 30, 37, 5, 26]. Designing task-generic methods is important for ease of portability to new tasks and the goal of our work, as well as a must in multi-task learning scenarios.

To this end, we perform simple but effective modifications to FixMatch [25] to adapt it for a larger class of dense or structured task, staying as close as possible to the original formulation. We call our approach Dense FixMatch and summarise it in Figure 2. We align the reference frame of the pseudo-labels obtained from the weakly-augmented view with that of the predictions obtained from the

https://doi.org/10.7557/18.6794

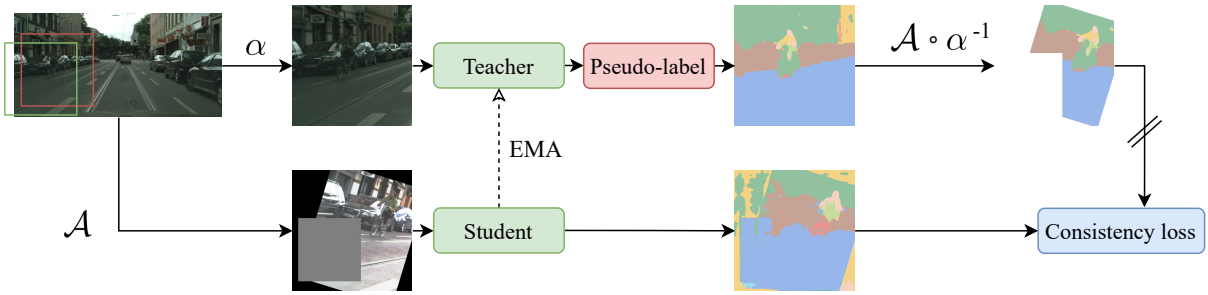

Figure 2: Dense FixMatch diagram for semantic segmentation. From an input image (top-left), two different views are created via α and $\mathcal{A}$, the weak and strong augmentation pipelines respectively. The squares represent the crops used for obtaining both views. Top: the first view is used by the teacher in the Mean Teacher [27] framework to generate pseudo-labels, updated via the exponential moving average (EMA) of the student weights. These are matched to the second view after applying the inverse of the weak augmentation α^{-1}, and then the strong one $\mathcal{A}$. Bottom: the second view is passed to the student model to obtain predictions to train against the pseudo-labels via the consistency loss, possible to define thanks to the shared structure and reference frame between both.

strongly-augmented view. This way, we can define a consistency loss at each output location while still using the full set of possible augmentations. Using strong and varied augmentations has been identified as a key component of self-training with input-consistency [29, 9] since they allow exploring larger neighbourhoods of the training data points in the input data manifold as well as in different directions. In addition, we incorporate the Mean Teacher (MT) framework for robustness to noisy pseudo-labels and imbalanced class size [16, 32].

We evaluate our approach on semantic segmentation with Cityscapes and Pascal VOC 2012 datasets and show the results in Figure 1 and Table 1, outperforming supervised baselines across different labeled data regimes by a large margin and achieving comparable results to other works in the literature. We also compare different mini-batch sampling approaches to assess whether it is feasible to use our method for semi-supervised multi-task learning where separate labeled and unlabeled data sampling is not possible [17].

Our contributions are as follows:

- We propose Dense FixMatch, a simple method that adds a matching operation between pseudo-labels and predictions to FixMatch thereby enabling its use on semi-supervised learning for any dense or structured task.
- We study its performance on semi-supervised semantic segmentation on Cityscapes and Pascal VOC 2012, showing improvements across multiple labeled data regimes over supervised baselines. For Cityscapes, we get improvements of up to +0.1 mean Intersection-over-Union (mIOU) for 93 and 186 labeled samples reaching **0.6697** mIoU and **0.7110** mIoU respectively, and +0.04 mIoU when using all labeled samples and extra unlabeled data reaching **0.8082** mIoU.
- We ablate our design choices and hyperparameters to give practitioners insights on how to tune it for new tasks and datasets.

2 Related work

The success of semi-supervised learning [34] has come mainly from its application to image classification with deep learning. Wei *et al.* [29] proved that SSL methods based on (a) self-training and (b) consistency regularization will achieve high accuracy with respect to ground-truth labels with the key to their success being to explore large enough neighbourhoods of the pseudo-labeled examples in the input data manifold, for example via aggressive data augmentation. Self-training or pseudo-labeling methods rely on bootstrapping current model predictions on unlabeled data and using them as labels. Consistency regularization relies on the assumption that small perturbations

of the data points in either input or latent space should not change the output.

FixMatch [25] and Noisy Student self-training [30] are two methods combining such building blocks: the former follows an online approach where pseudo-labels are generated during training, and the latter has subsequent pseudo-labeling and re-training phases. Both use strong data augmentation to train against the pseudo-labels. Other works have also used other kinds of perturbations for the consistency objective, such as adversarial examples [19], network perturbations [15, 27, 30] or MixUp [36, 1] as well as other techniques to tackle distribution misalignment between true- and pseudo-labels [2].

SSL applied to tasks other than image classification has also seen significant developments in recent years. For semi-supervised object detection, multiple works have used consistency regularization and perturbations via data augmentation [13, 26, 16, 31]. For semi-supervised semantic segmentation, the work in [9] found that strong and varied perturbations are required and proposed CutMix [35] as the strong augmentation. CCT [20] enforces consistency between predictions perturbing latent features. GCT [14] uses two differently initialized networks for co-training and a flaw detection module. CPS [5] instead enforces consistency against hard pseudo-labels. Pseudoseg [38] uses strong augmentation and fuses pseudo-labels from decoder predictions with ones from GradCAM [24]. ST++ [33] does self-training with strong data augmentation in the re-training phase while selecting and prioritizing reliable images. AEL [12] focuses on balancing the performance between classes via different task-specific strategies. U2PL [28] uses unreliable pseudo-labels for negative learning.

In contrast, we follow FixMatch as close as possible to keep the benefits of using online pseudo-labeling and consistency regularization between predictions on weakly and strongly augmented images. We add only a spatial matching operation to enable its use in dense and structured tasks and the MT framework for improving pseudo-label quality.

3 Dense FixMatch

We adapt FixMatch [25] for its use in structured and dense prediction tasks in SSL.

Our method assumes the standard framework of semi-supervised learning where labeled samples X_L contribute to the supervised objective $\mathcal{L}_s$ and unlabeled samples X_U are used in an unsupervised objective $\mathcal{L}_u$, with the option to use the labeled samples also for the latter. The unsupervised loss weight λ trades off the contribution of both objectives to the final loss: $\mathcal{L} = \mathcal{L}_s(\mathbf{x}, \mathbf{y}, \theta) + \lambda\mathcal{L}_u(\mathbf{x}, \theta)$.

To define the unsupervised or consistency objective, FixMatch uses image-level pseudo-labels obtained from a weakly-augmented version of the unlabeled images (via augmentation pipeline α) to supervise learning on the strongly-augmented version of the same images (via $\mathcal{A}$). For image classification, the output is expected to be invariant to the applied transformations and so the obtained pseudo-label can be directly used for this purpose. In contrast, this is not possible when the output of the task at hand has a spatial structure related to that of the input and thus will vary depending on the applied augmentations. This is the case for dense or structured tasks such as semantic segmentation or object detection, among others. For those tasks, any geometric transformation of the input equivariantly transforms its corresponding output. Therefore, when using geometric transformations as part of the weak and strong augmentation pipelines, the obtained pseudo-labels will not generally match pixel-to-pixel or at each location.

We adopt a simple approach to align the predictions of one view (e.g. weak augmentation α) to the reference frame of the other view (e.g. strong augmentation $\mathcal{A}$). Specifically, we first apply the inverse geometric transformation of the first view to the predictions obtained on it and then apply the geometric transformation of the second view so that predictions on both views end up in the same reference frame. This simple mechanism enables to define a consistency objective between the two views for any dense or structured task, including semantic segmentation, object detection, and instance segmentation, while still being able to use different geometric transformations in both augmentation pipelines. Figure 2 illustrates our approach for the case of semantic segmentation.

For our experiments on semantic segmentation, we define the supervised objective $\mathcal{L}_s$ as a per-sample, location-wise cross-entropy loss between predictions on a weakly-augmented view of labeled samples and their corresponding (aug-

mented) ground-truth labels, and averaging over all valid locations. The unsupervised or consistency objective $\mathcal{L}_u$ is defined by applying the same location-wise cross-entropy objective between the predictions of a strongly-augmented view and the pseudo-label obtained from a weakly-augmented view of the same sample after spatially matching the latter. Note that, for the unsupervised loss $\mathcal{L}_u$, gradients are back-propagated only through the predictions on strongly-augmented samples, and not through the pseudo-labels.

From prediction to pseudo-label. For our experiments in semantic segmentation, the model outputs normalized classification probabilities, using softmax operation, for each possible class at each output location. Obtaining a "hard" pseudo-label means retaining only the most likely class at each location, achieved by applying the argmax operation. In addition, it is common to use only high-confidence predictions as pseudo-labels [30, 25] via a confidence threshold τ above which to retain the labels. Locations with prediction confidence below the threshold do not contribute to the loss.

Matching operation. In order to spatially align or match predictions and pseudo-labels, we bring the pseudo-label to the prediction reference frame by first applying the inverse of the geometric transforms in the weak augmentation pipeline, α^{-1}, and then the transforms of the strong augmentation pipeline, $\mathcal{A}$, to the pseudo-label $\bar{\mathbf{y}}_i$. In this way, we avoid back-propagating gradients through the matching operation. However, some locations of the matched pseudo-label will end up with invalid values which should not contribute to the loss.

$$\text{match}(\bar{\mathbf{y}}_i; \alpha, \mathcal{A}) = \mathcal{A}(\alpha^{-1}(\bar{\mathbf{y}}_i)). \tag{1}$$

Augmentation pipelines. We adapt the augmentations of FixMatch [25] to semantic segmentation. We use flip-and-shift as the weak augmentation α and RandAugment [7] for the strong augmentation $\mathcal{A}$ for simplicity, including both geometric and color transforms. We use random crops for both pipelines before applying the rest of the transforms to ensure the same input size and instead of the shift operations. Other works focusing on dense tasks and inspired by FixMatch avoid the misalignment by dropping the geometric transforms [38, 16], applying the same to both views [18], or applying them only as part of $\mathcal{A}$ [31].

Mean Teacher. To obtain cleaner and more stable pseudo-labels [32, 16], we generate them using the teacher in MT [27] instead of the same model. The teacher is updated via the exponential moving average (EMA) of the student weights.

4 Experiments

Datasets. We use Cityscapes [6] and Pascal VOC 2012 [8] datasets for evaluating our method on semi-supervised semantic segmentation. Cityscapes consists of 2975 samples for training, with fine annotations for 19 classes, and 500 samples for evaluation. In addition, further 20000 samples are available in the `extra` set with coarse annotations, but we will use them later as unlabeled samples only. Pascal VOC consists of 1464 samples for training in the original set with annotations for 21 classes including background, and 1449 samples for evaluation in the validation set. Moreover, there are further 9118 labeled samples in the augmented set from SBD [10]. As is common in the literature, we simulate the semi-supervised setting with labeled and unlabeled splits of the training set with different labeled data regimes or ratios of labeled samples. For each split, we generate four different random splits with no guarantees of class balance. For Pascal VOC, we split the original set and use the augmented set as unlabeled data.

Model. Following the literature, we use DeepLabv3+ [4] on ResNet-50 or ResNet-101 backbones [11] to define our semantic segmentation model, giving predictions at the same spatial resolution as the input. The model is initialized with ImageNet [23] pre-trained weights.

Implementation details. We implement and train our models using PyTorch [21] with distributed training and mixed precision on up to four NVIDIA A100 GPUs, depending on the total batch size and the backbone used. We use the computer vision library Kornia [22] for implementing the data augmentation pipelines for its support of invertible geometric transforms and differentiable augmentations for multiple tasks, including semantic segmentation. We follow EMAN [3] and use the exponential moving average of the Batch Normalization statistics of the student to update the teacher.

Evaluation. We evaluate the performance of

Method	Backbone	Sampling	**93**$_{(1/32)}$	**186**$_{(1/16)}$	**372**$_{(1/8)}$	**744**$_{(1/4)}$	**1488**$_{(1/2)}$	**2975**$_{All}$
Supervised	RN-50		.5579$_{\pm.0091}$	.6004$_{\pm.0012}$	.6550$_{\pm.0051}$	.6943$_{\pm.0065}$	.7332$_{\pm.0095}$	.7608$_{\pm.0054}$
	RN-101		.5686$_{\pm.0080}$	.6122$_{\pm.0025}$	.6628$_{\pm.0081}$	.6979$_{\pm.0049}$	.7421$_{\pm.0081}$	.7652$_{\pm.0023}$
Dense FixMatch	RN-50	Explicit	.6581$_{\pm.0202}$	.7013$_{\pm.0079}$	.7243$_{\pm.0049}$	.7504$_{\pm.0063}$	.7599$_{\pm.0063}$	
		Implicit	.6554$_{\pm.0158}$	.7065$_{\pm.0065}$	.7339$_{\pm.0055}$	.7547$_{\pm.0070}$	.7637$_{\pm.0079}$	
	RN-101	Explicit	.6697$_{\pm.0119}$	.7110$_{\pm.0061}$	.7283$_{\pm.0069}$	.7572$_{\pm.0015}$	.7666$_{\pm.0050}$	
		Implicit	.6481$_{\pm.0178}$	.7083$_{\pm.0098}$	.7391$_{\pm.0028}$	.7565$_{\pm.0058}$	.7635$_{\pm.0088}$	

Table 1: Results of Dense FixMatch for semantic segmentation (mIoU) on **Cityscapes** `val` set with few labeled samples on different amounts of labeled data, ResNet-50/101 backbones and DeepLabv3+, and either explicit or implicit mini-batch sampling settings. Dense FixMatch significantly improves over the baselines for both settings. Results highlighted with color match those reported in Figure 1.

Method	Backbone	Sampling	**92**$_{(1/16)}$	**183**$_{(1/8)}$	**366**$_{(1/4)}$	**732**$_{(1/2)}$	**1464**$_{Original}$	**10582**$_{Augmented}$
Supervised	RN-50		.4075$_{\pm.0114}$	.5361$_{\pm.0257}$	.6183$_{\pm.0153}$	.6788$_{\pm.0056}$	.7214$_{\pm.0029}$	.7522$_{\pm.0038}$
	RN-101		.4482$_{\pm.0256}$	.5771$_{\pm.0247}$	.6534$_{\pm.0081}$	.7059$_{\pm.0041}$	.7454$_{\pm.0023}$	.7722$_{\pm.0017}$
Dense FixMatch	RN-50	Explicit	.5215$_{\pm.0246}$	.6249$_{\pm.0374}$	.6902$_{\pm.0045}$	.7169$_{\pm.0010}$	.7391$_{\pm.0012}$	
		Implicit	.4928$_{\pm.0284}$	.5892$_{\pm.0334}$	.6729$_{\pm.0076}$	.7031$_{\pm.0028}$	.7432$_{\pm.0045}$	
	RN-101	Explicit	.5485$_{\pm.0501}$	.6582$_{\pm.0334}$	.7204$_{\pm.0059}$	.7473$_{\pm.0008}$	.7716$_{\pm.0020}$	
		Implicit	.4984$_{\pm.0300}$	.6133$_{\pm.0313}$	.7047$_{\pm.0053}$	.7414$_{\pm.0044}$	.7710$_{\pm.0046}$	

Table 2: Results of Dense FixMatch for semantic segmentation (mIoU) on **Pascal VOC 2012** `val` set with few labeled samples on different amounts of labeled data, ResNet-50/101 backbones and DeepLabv3+, and explicit or implicit settings.

our method using as the main metric the mean Intersection-over-Union (mIOU) over all classes. For simplicity, we use full-resolution, single-scale, and single-pass evaluation in contrast to the sliding evaluation approach used in other works [5, 12, 28]. For stability, the model used for evaluation has weights following the EMA of the weights obtained during training. For most experiments, this means exactly using the teacher in the MT framework for evaluation. For each labeled data regime, we train our model on each of the four random data splits using also a different random seed for each training run. We take the best checkpoint of each run according to the mIOU and compute the mean and standard deviation over the four runs.

Training details. We follow the training details of [28]. The baselines use a standard augmentation pipeline with random resized crops and horizontal flips. We employ two different mini-batch sampling strategies for SSL: (a) the common *explicit* setting in which labeled and unlabeled data are sampled separately, and (b) the alternative *implicit* setting in which all data is sampled uniformly regardless of labels [17]. We train each setting for the equivalent of 80 or 240 epochs on the full train set for the supervised baseline, i.e. 52910 or 89250 updates, for Pascal VOC and Cityscapes, respectively.

4.1 Results on few labeled samples

In Figure 1 and Tables 1 and 2, we show results for the common splits of 1/32 to 1/2 of all the full `train` set for Cityscapes and 1/16 to the full original `train` set for Pascal VOC 2012, respectively. We compare few-supervision baselines using only the labeled data with Dense FixMatch, which also uses the unlabeled samples in either the explicit or implicit settings. Our method performs better than the baselines for both mini-batch sampling approaches. The explicit setting is slightly better for more label-scarce regimes but both give similar results with more labeled data.

4.2 Results on full labeled set and extra unlabeled samples

In addition, we evaluate Dense FixMatch in the more realistic setting where all labeled samples are used and extra unlabeled data is available in Cityscapes. We use all samples from the `extra` set

Method	Backbone			mIoU	
Supervised	RN-50			$.7608_{\pm.0054}$	
	RN-101			$.7652_{\pm.0023}$	
		`train`	`extra`	Explicit	Implicit
Dense FixMatch	RN-50	✓		$.7869_{\pm.0018}$	
			✓	$.7916_{\pm.0026}$	$.7935^{\dagger}_{\pm.0017}$
		✓	✓	$\mathbf{.7998_{\pm.0020}}$	$.7948_{\pm.0016}$
	RN-101	✓		$.7907_{\pm.0020}$	
			✓	$.8005_{\pm.0010}$	$.7974^{\dagger}_{\pm.0020}$
		✓	✓	$\mathbf{.8082_{\pm.0024}}$	$.8012_{\pm.0013}$

Table 3: Results on Cityscapes full labeled set and `extra` unlabeled samples for the supervised baselines and semi-supervised Dense FixMatch in both the explicit and implicit mini-batch sampling settings for SSL. We also compare using our method as a regularization on the labeled data only and using it on both labeled and unlabeled data. †is our setting for the main experiments.

τ	λ	mIoU
0	0.2	0.6566
	0.5	0.6464
	1	**0.6594**
	2	*0.6581*
	5	0.5699
0.5	0.1	0.6447
	0.2	**0.6609**
	0.5	0.6495
	1	$0.6501^{\dagger}$
	2	*0.6554*
	5	0.6090
	10	0.5276
0.8	1	0.6388
0.95	1	0.6148

Table 4: Ablation study on the pseudo-label confidence threshold τ and the consistency loss weight λ. †values used in the main experiments.

Method				mIoU
Supervised				0.5409
Mean Teacher*				0.5820
	Crop relation	Augmentation	MT	
Dense FixMatch	Same	Crop+color	✓	0.5794
	Same	Crop+color+cutout	✓	0.6531
	Min. overlap	Crop+color+cutout	✓	0.6542
	Min. overlap	Crop+geom.	✓	0.6539
	Min. overlap	Crop+geom.+cutout	✓	0.6517
	Min. overlap	Crop+color+geom.	✓	0.6579
	Same	Crop+color+geom.	✓	0.6157
	Any	Crop+color+geom.+cut.	✓	0.6300
	Same	Crop+color+geom.+cut.	✓	**0.6660**
	Min. overlap	Crop+color+geom.+cut.	✓	*0.6594*†
	Min. overlap	Crop+color+geom.+cut.		0.6275

Table 5: Ablation on the use of Mean Teacher (MT) framework, the relation between crops in the two views of Dense FixMatch and the use of the geometric, color or all augmentations in RandAugment. *Using MT on its own with the augmentation pipeline of the supervised baseline. We use a logistic warm-up schedule for the consistency weight during the first 60 epochs. †is our setting for the main experiments.

but discard the coarse annotations and just treat them as unlabeled samples. We also compare to the fully-supervised baseline using only the labeled data and using the consistency loss as a regularization term only, i.e. computed on the same labeled data as the supervised loss. Results are shown in Table 3. Computing the loss on both labeled and unlabeled samples gives the best results.

4.3 Ablations

In Table 4, we ablate the main hyper-parameters of our method using the 93 labeled samples regime of semi-supervised semantic segmentation with Cityscapes, with a single data split and seed, and the explicit setting. For dense tasks such as semantic segmentation, other works have reported that using a low τ or removing it altogether gives better final results [38], hypothesizing that discarding predictions of lower confidence makes the loss dominated by easy classes [18, 12]. Moreover, low-confidence predictions in semantic segmentation tend to concentrate in truly ambiguous regions such as the boundaries between objects of different classes [28] so discarding them means removing supervision mainly from boundary pixels which are in fact the most informative. Values for λ between 0.2 and 2 give comparable results.

In Table 5 we ablate our design choices: the relation between crops in both views, the choice of augmentations, and the use of MT. The best results are obtained when using MT and all possible augmentations, although using the same crop instead of overlapping crops between views improves results. We hypothesize this is due to a larger number of valid locations in the matched pseudo-labels.

Method	Road	Side.	Build.	Wall	Fence	Pole	T.light	T.sign	Veg.	Terr.	Sky	Person	Rider	Car	Truck	Bus	Train	Motor.	Bic.	Mean
% pixels	36.02	7.06	25.61	0.41	0.81	1.16	0.18	0.59	14.08	1.22	3.80	1.08	0.09	6.73	0.25	0.23	0.21	0.14	0.34	
Supervised	.962	.722	.873	.293	.296	.500	.497	.563	.896	.469	.914	.734	.336	.888	.114	.276	.318	.273	.677	$.558_{\pm.262}$
Dense FixMatch (E)	**.976**	.810	**.902**	.414	.424	.574	**.619**	**.710**	.906	.575	**.935**	**.778**	**.544**	**.924**	**.560**	**.699**	.295	**.488**	**.715**	$.676_{\pm.195}$
Dense FixMatch (I)	.974	**.820**	.900	**.467**	**.432**	**.578**	.603	.688	**.913**	**.587**	.931	.769	.448	**.924**	.495	.667	**.471**	.475	.661	$.674_{\pm.184}$

Method	Bg.	Plane	Bicy.	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor.	Person	Pott.	Sheep	Sofa	Train	Tv	Mean
% pixels	72.45	1.09	0.83	0.15	0.94	1.57	3.46	2.60	0.58	0.65	1.25	2.08	1.13	0.58	0.90	5.27	0.56	0.34	0.49	1.72	1.35	
Supervised	.883	.702	.379	.227	**.574**	.443	.764	.639	.336	.114	.429	.215	.309	.332	.542	.647	.199	.511	.202	.610	.307	$.446_{\pm.206}$
Dense FixMatch (E)	.890	**.784**	**.523**	**.639**	.500	.555	**.843**	.684	**.776**	.188	**.514**	**.439**	.384	.388	**.752**	.622	.356	**.576**	**.352**	.743	**.553**	$.574_{\pm.180}$
Dense FixMatch (I)	**.899**	.740	.459	.010	.559	**.583**	.833	**.721**	.431	**.192**	.469	.347	**.418**	**.559**	.693	**.672**	**.364**	.356	.333	**.747**	.545	$.520_{\pm.213}$

Table 6: Class-wise IoU on `val` set of Cityscapes (top) when training on a 93 labeled samples data split and on `val` set of Pascal VOC 2012 (bottom) when training on a 92 labeled samples data split. We show in **bold** the best result for each class and in red the classes that perform worse than the supervised baseline for both the *explicit* (E) and *implicit* (I) mini-batch sampling settings.

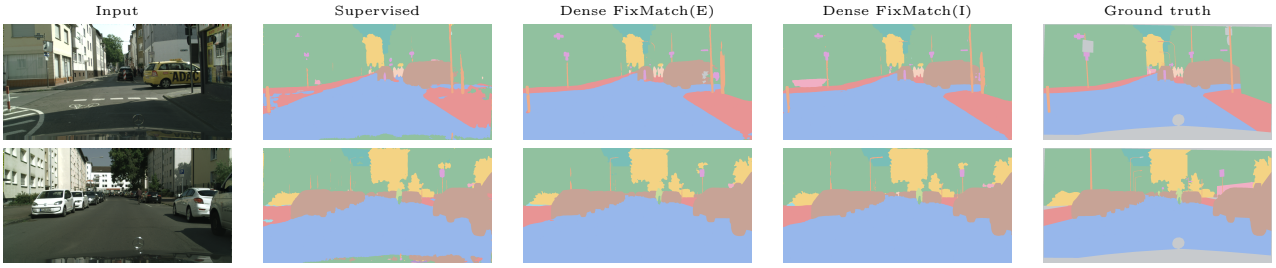

Figure 3: Qualitative results for semi-supervised learning on Cityscapes with 93 labeled samples for the supervised baseline and Dense FixMatch in the *explicit* (E) and *implicit* (I) mini-batch sampling settings, shown on samples in the validation set.

4.4 Class-wise analysis

In both Cityscapes and Pascal VOC, the predominant classes appear in orders of magnitude more pixels than the least frequent ones [6, 8]. In Table 6, we give per-class results comparing the supervised baseline to Dense FixMatch on a single data split of 93 or 92 labeled samples for each dataset respectively. For Cityscapes, our method improves significantly on average over the baseline, and does so while improving for all but one class. Importantly, it also reduces the effect of class imbalance since the gap between the best-performing classes and the worst-performing ones is reduced significantly, as shown by the lower standard deviation across classes. For Pascal VOC, Dense FixMatch improves the results on average, but up to 3 classes get lower IoU.

4.5 Qualitative results

In Figure 3, we give some examples for the 93 labeled samples experiments with Cityscapes comparing the baseline and Dense FixMatch for both the explicit and implicit settings. Both settings give similar results and outperform the supervised baseline. Some examples are the cleaner boundaries between road and side-walk and for poles.

5 Conclusions

We proposed Dense FixMatch, a simple method that puts together the most important components in modern deep semi-supervised learning and adds a matching operation on the pseudo-labels. In this way, it can be used for multiple dense or structured prediction tasks with the full strength of data augmentation pipelines, including strong geometric transformations. We evaluated it on semi-supervised semantic segmentation on Cityscapes and Pascal VOC and ablate design choices as well as hyper-parameters. This gives future practitioners insights on how to tune the proposed method for other datasets and tasks.

Acknowledgments

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. URL http://papers.neurips.cc/paper/8749-mixmatch-a-holistic-approach-to-semi-supervised-learning.pdf.

[2] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklkeR4KPB.

[3] Z. Cai, A. Ravichandran, S. Maji, C. Fowlkes, Z. Tu, and S. Soatto. Exponential moving average normalization for self-supervised and semi-supervised learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 194–203, 2021. doi: 10.1109/CVPR46437.2021.00026.

[4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Computer Vision – ECCV 2018*, pages 833–851, Cham, 2018. Springer International Publishing. doi: 10.1007/978-3-030-01234-2_49.

[5] X. Chen, Y. Yuan, G. Zeng, and J. Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2613–2622, 2021. doi: 10.1109/CVPR46437.2021.00264.

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. doi: 10.1109/CVPR.2016.350.

[7] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le. Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88 (2):303–338, June 2010. doi: 10.1007/s11263-009-0275-4.

[9] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *British Machine Vision Conference*, number 31, 2020. URL https://www.bmvc2020-conference.com/conference/papers/paper_0680.html.

[10] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998, 2011. doi: 10.1109/ICCV.2011.6126343.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[12] H. Hu, F. Wei, H. Hu, Q. Ye, J. Cui, and L. Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b98249b38337c5088bbc660d8f872d6a-Abstract.html.

[13] J. Jeong, S. Lee, J. Kim, and N. Kwak. Consistency-based semi-supervised learning for object detection. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL

https://papers.nips.cc/paper/2019/hash/d0f4dae80c3d0277922f8371d5827292-Abstract.html.

[14] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *European Conference on Computer Vision (ECCV)*, August 2020. doi: 10.1007/978-3-030-58601-0_26.

[15] S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=BJ6oOfqge.

[16] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda. Unbiased teacher for semi-supervised object detection. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=MJIve1zgR_.

[17] M. Martí i Rabadán, S. Bujwid, A. Pieropan, H. Azizpour, and A. Maki. An analysis of over-sampling labeled data in semi-supervised learning with fixmatch. In *Proceedings of the Northern Lights Deep Learning Workshop*, volume 3, 2022. doi: 10.7557/18.6269.

[18] L. Melas-Kyriazi and A. K. Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12430–12440, 2021. doi: 10.1109/CVPR46437.2021.01225.

[19] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019. doi: 10.1109/TPAMI.2018.2858821.

[20] Y. Ouali, C. Hudelot, and M. Tami. Semi-supervised semantic segmentation with cross-consistency training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12671–12681, 2020. doi: 10.1109/CVPR42600.2020.01269.

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[22] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3663–3672, 2020. doi: 10.1109/WACV45572.2020.9093363.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

[25] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. Raffel, E. D. Cubuk, A. Kurakin, and C. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html.

[26] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister. A simple semi-supervised learning framework for object detection, 2020. URL https://arxiv.org/abs/2005.04757.

[27] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1195–1204, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html.

[28] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4238–4247, 2022. doi: 10.1109/CVPR52688.2022.00421.

[29] C. Wei, K. Shen, Y. Chen, and T. Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=rC8sJ4i6kaH.

[30] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2020. doi: 10.1109/CVPR42600.2020.01070.

[31] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, X. Bai, and Z. Liu. End-to-end semi-supervised object detection with soft teacher. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3040–3049, 2021. doi: 10.1109/ICCV48922.2021.00305.

[32] Z. Xu, D. Lu, Y. Wang, J. Luo, J. Jayender, K. Ma, Y. Zheng, and X. Li. Noisy labels are treasure: Mean-teacher-assisted confident learning for hepatic vessel segmentation. pages 3–13, 2021. doi: 10.1007/978-3-030-87193-2_1.

[33] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao. St++: Make self-trainingwork better for semi-supervised semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4258–4267, 2022. doi: 10.1109/CVPR52688.2022.00423.

[34] X. Yang, Z. Song, I. King, and Z. Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20, 2022. doi: 10.1109/TKDE.2022.3220219.

[35] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019. doi: 10.1109/ICCV.2019.00612.

[36] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

[37] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. Le. Rethinking pre-training and self-training. In *Advances in Neural Information Processing Systems*, volume 33, pages 3833–3845, 2020. URL https://proceedings.neurips.cc/paper/2020/file/27e9661e033a73a6ad8cefcde965c54d-Paper.pdf.

[38] Y. Zou, Z. Zhang, H. Zhang, C.-L. Li, X. Bian, J.-B. Huang, and T. Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=-TwO99rbVRu.