MORE THAN MEETS THE EYE: SEMI-SUPERVISED LEARNING UNDER NON-IID DATA

Saul Calderon-Ramirez[*]
Centre for Computational Intelligence (CCI)
De Montfort University
Leicester, United Kingdom
sacalderon@itcr.ac.cr

Luis Oala[*]
AI Department
Fraunhofer HHI
Berlin, Germany
luis.oala@hhi.fraunhofer.de

ABSTRACT

A common heuristic in semi-supervised deep learning (SSDL) is to select unlabelled data based on a notion of semantic similarity to the labelled data. For example, labelled images of numbers should be paired with unlabelled images of numbers instead of, say, unlabelled images of cars. We refer to this practice as semantic data set matching. In this work, we demonstrate the limits of semantic data set matching. We show that it can sometimes even degrade the performance for a state of the art SSDL algorithm. We present and make available a comprehensive simulation sandbox, called *non-IID-SSDL*, for stress testing an SSDL algorithm under different degrees of distribution mismatch between the labelled and unlabelled data sets. In addition, we demonstrate that simple density based dissimilarity measures in the feature space of a generic classifier offer a promising and more reliable quantitative matching criterion to select unlabelled data before SSDL training.

Training an effective deep learning solution typically requires a considerable amount of labelled data S_l. In specific areas, like medical imaging, annotations can be expensive to obtain. Several approaches have been developed to address this data constraint, including data augmentation, self-supervised and semi-supervised deep learning (SSDL), among others (Weiss et al., 2016; Perez & Wang, 2017; Zhou, 2018). Semi-supervised learning leverages unlabelled data S_u, thus offering a promising approach for problems where little labelled data is available or a range of labels is lacking (Van Engelen & Hoos, 2020). As with other learning paradigms, the transfer of SSDL techniques from lab to real-world is complicated by violations of the independently and identically distributed (IID) assumption, specifically where the distributions of labelled inputs P_l and unlabelled inputs P_u do not match up, i.e. $P_l \neq P_u$. In principle, we would like to exploit available unlabelled data as flexibly as possible. In practice, mismatches between the labelled and unlabelled data sets can lead to serious performance degradation (Oliver et al., 2018). This begs the question how we can systematically select labelled and unlabelled data in non-IID settings such that performance on the downstream task is increased. A common recourse are what we call semantic matching heuristics. Under such a scheme anthropogenic semantic concepts are used to delineate distribution mismatches between data sets. For example, different classes of the same data set have been viewed as out-of-distribution to one another because one class displays animals and the other class displays vehicles (Chen et al., 2020; Zhao et al., 2021). Practices of semantic matching can be traced to other fields of machine learning, too, including out-of-distribution detection (Zisselman & Tamar, 2020) or the domain adaptation (Zhang et al., 2020) literature. Insights from generative modelling hint at the limits of semantic matching: a supposedly different data set like Street View House Numbers (SVHN) can have a higher likelihood under a generative model trained on ImageNet than ImageNet data itself (Nalisnick et al., 2018). Thus, in this work we take a close look at SSDL under varying non-IID conditions and demonstrate the limits of semantic matching. Specifically:

- We show that semantic heuristics can fail when selecting labelled and unlabelled data for the state-of-the-art SSDL algorithm MixMatch. To our knowledge, this is the first work to extensively evaluate the accuracy behaviour of MixMatch under different degrees of mismatch between the

[*] Equal contribution

Table 1: Results for the class distribution mismatch experiment, best OOD performance in bold per configuration. Each result entry in the table represents the mean and variance of accuracy across ten random experimental runs per entry. For a detailed description of symbols and the experiment see Section 2. Numbers in parentheses next to $S_{\text{u,OOD}}$ entries reflect the preference ranking provided by disimilarity matching with d_C.

$\mathbf{S_{IOD}}$	$\mathbf{T_{OOD}}$	$\mathbf{S_{u,OOD}}$	$\mathbf{\%_{uOOD}}$		$\mathbf{n_l}$	
				60	100	150
MNIST	Fully supervised baseline			0.457 ± 0.108	0.559 ± 0.125	0.645 ± 0.101
	SSDL baseline (no OOD data)			0.704 ± 0.096	0.781 ± 0.065	0.831 ± 0.0626
	OH	MNIST	50	$\mathbf{0.679 \pm 0.108}$	$\mathbf{0.769 \pm 0.066}$	0.802 ± 0.054
			100	0.642 ± 0.111	0.746 ± 0.094	0.798 ± 0.07
	Sim	SVHN (3)	50	0.637 ± 0.098	0.745 ± 0.081	0.801 ± 0.0699
			100	0.482 ± 0.113	0.719 ± 0.058	0.765 ± 0.072
	Dif	TI (2)	50	0.642 ± 0.094	0.739 ± 0.074	0.809 ± 0.066
			100	0.637 ± 0.097	0.732 ± 0.074	0.804 ± 0.071
		GN (5)	50	0.606 ± 0.0989	0.713 ± 0.087	0.786 ± 0.065
			100	0.442 ± 0.099	0.461 ± 0.073	0.542 ± 0.062
		SAPN (4)	50	0.631 ± 0.102	0.735 ± 0.082	$\mathbf{0.813 \pm 0.057}$
			100	0.48 ± 0.0951	0.524 ± 0.09	0.563 ± 0.095

labelled and unlabelled data. We make available the comprehensive simulation sandbox, called *non-IID-SSDL*, from our experiments. The sandbox incorporates eight data sets and different configurations for mixing data sets and the number of labelled data points.

- In addition, we demonstrate that simple dissimilarity measures in the feature space of a generic deep ImageNet classifier offer a promising and more reliable matching criterion. We corroborate these results with a comprehensive statistical analysis.

1 METHOD

Note that from hereon out we refer to Inside of Distribution (IOD) data as the labelled, in-distribution data and to Out of Distribution (OOD) data as the unlabelled data that is non-IID relative to IOD. In order to systematically assess the impact of non-IID data on SSDL performance we created the flexible *non-IID-SSDL* simulation sandbox which we make openly available[1]. The ablation sandbox has five degress of freedom: base data S_{IOD} which constitutes the original task to be learned, the type of OOD data T_{OOD}, the OOD data source $S_{\text{u,OOD}}$, the relative amount of OOD data among the unlabelled data $\%_{\text{u,OOD}}$, the amount n_l of labelled observations and the SSDL algorithm to be used. More details on the configurations of these arguments are provided in Section 2 below. The algorithm under consideration for the *non-IID-SSDL* stress test in our analysis is MixMatch, a state of the art SSDL method (Berthelot et al., 2019). We provide a detailed description of our implementation in Appendices A.1 and B.1. As a simple quantitative baseline against the semantic matching heuristic we consider the dissimilarity between two data sets S_a and S_b in the feature space of a generic Wide-ResNet pre-trained on ImageNet. We consider four dissimilarity measures: two Minkowski based distance sets, $d_{\ell_2}\left(S_a, S_b, \tau, \mathcal{C}\right)$ and $d_{\ell_1}\left(S_a, S_b, \tau, \mathcal{C}\right)$, corresponding to the Euclidean and Manhattan distances, respectively. Full details on the computation of the measures can be found in Appendix A.2. We also perform additional qualitative analyses by comparing the distributions of individual features between different data sets.

Finally, we corroborate the results with statistical analyses. The distances were all calculated on smaller sub-samples of the full data sets. Hence, we also test if the measurements are statistically meaningful using a Wilcoxon test (Fagerland & Sandvik, 2009). In all sandbox simulations, we performed ten experimental runs and report the accuracy mean and variance of the models performing best on the test data from each run.

2 EXPERIMENTS

In order to evaluate semantic matching heuristic we consider the following settings for the *non-IID-SSDL* sandbox. Three configurations for S_{IOD} comprising MNIST, CIFAR-10 and FashionMNIST. A total of three configurations for T_{OOD} (Other-Half (OH), Similar (Sim) and Different (Dif)) are tested. In the OH setting half of the classes and associated inputs are taken to be the S_{IOD} data whereas the other half of classes are taken to be the $S_{\text{u,OOD}}$ data. Sim is a $S_{\text{u,OOD}}$ data set that is supposedly semantically related to S_{IOD}, e.g. MNIST and SVHN. Dif is a $S_{\text{u,OOD}}$ data set that is assumed to be semantically unrelated to S_{IOD}, e.g. MNIST and Tiny ImageNet. We note that a degree

[1] All code and experimental scripts, with automatic download of sandbox data for ease of reproduction, can be found at `https://github.com/luisoala/non-iid-ssdl`.

Table 2: Dissimilarity measures between the labelled and unlabelled datasets S_l and S_u. Numbers in italics correspond to results with $p > 0.05$ for the Wilcoxon test. For a detailed description of symbols and the experiment see Appendix A.2

$\mathbf{S_l}$	$\mathbf{S_u}$	$\%_{\mathbf{uOOD}}$	$\mathbf{d}_{\ell_2}$	$\mathbf{d}_{\ell_1}$	$\mathbf{d_{JS}}$	$\mathbf{d_C}$
MNIST	OH	50	*0.011 ± 0.006*	*0.459 ± 0.28*	*0.266 ± 0.221*	*0.811 ± 0.512*
		100	*0.014 ± 0.019*	*0.38 ± 0.507*	1.001 ± 0.725	1.263 ± 0.665
	SVHN	50	*0.09 ± 0.017*	1.569 ± 0.504	6.789 ± 0.924	12.021 ± 1.757
		100	0.25 ± 0.053	4.702 ± 1.04	52.349 ± 3.292	42.026 ± 4.31
	TI	50	0.008 ± 0.023	1.519 ± 0.223	3.663 ± 0.772	5.512 ± 0.767
		100	0.217 ± 0.04	4.3 ± 0.636	10.305 ± 1.667	15.18 ± 2.698
	GN	50	0.11 ± 0.0219	1.958 ± 0.534	14.785 ± 1.052	23.593 ± 1.859
		100	0.357 ± 0.081	5.987 ± 1.091	52.349 ± 4.253	86.21 ± 3.471
	SAPN	50	0.089 ± 0.0311	2.479 ± 0.7433	15.116 ± 1.475	20.151 ± 1.619
		100	0.323 ± 0.07	6.308 ± 1.366	53.397 ± 4.253	77.456 ± 4.474

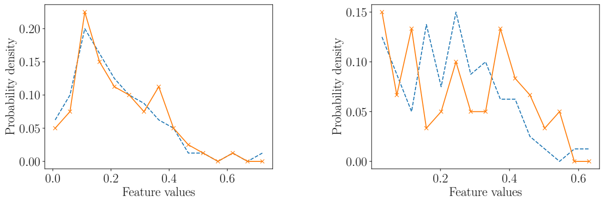
of subjectivity is involved in the selection what is to be considered Sim and Dif - a drawback of the semantic matching heuristic itself. There are five configurations for $S_{\text{u,OOD}}$ as explained above: the other half OH, a similar data set, and three different data sets including two noise baselines. They include Street View House Numbers SVHN, Tiny ImageNet (TI), Gaussian Noise (GN), Salt and Pepper Noise (SAPN) and Fashion Product (FP). Each configuration represents a multi-class classification task with $|\mathcal{Y}| = 5$, that is a random subset of half of the classes of base data S_{IOD}. This is so that metrics on all T_{OOD} settings, including OH, are comparable. We vary the relative amount of OOD data $\%_{\text{u,OOD}}$ between 0, 50 and 100, where the IOD data is a partition of the labelled data. For instance in MNIST, after splitting the classes for IOD and OOD, we use a portion of the IOD data as labelled and another one as unlabelled. We also vary the amount of labelled data points n_l between 60, 100 and 150. We study the behaviour of MixMatch under very limited number of labels settings, where the benefit of SSDL is usually higher. This makes the impact of distribution mismatch more evident. For each run we sampled non-overlapping subsets of data from S_{IOD} and $S_{\text{u,OOD}}$ to obtain the required number of labelled n_l and unlabelled n_u samples for the run. Descriptive statistics for standardization of the inputs were only computed from these subsets to keep the simulation realistic and not leak any information from the global data sets. All hyper-parameters are described in Table 3 in Appendix A.1 and kept constant across all the tests performed to isolate the effects of the variable sandbox parameters.

Table 1 shows the results of the data set matching experiments described in Section 2. We make a number of observations. First, in the majority of cases using IOD unlabelled data or a 50-50 mix of IOD and OOD unlabelled data beats the fully supervised baseline. The gains range from 15% to 25% for MNIST, 10% to 15% for CIFAR-10 and 7% to 13% for FashionMNIST across all $S_{\text{u,OOD}}$ and n_l. As expected, in most of the cases the accuracy is degraded when including OOD data in S_u, with a more severe degradation when noisy data sets (SAPN, GN) are used as OOD data contamination source. Second, it is not always the case that $T_{\text{OOD}} =$ OH (when $S_{\text{u,OOD}}$ is supposedly more semantically similar to S_{IOD}) yields the best MixMatch performance. This is observed when $S_l =$ CIFAR-10, $n_l = 100$ and $n_l = 150$, where OOD unlabelled data from Tiny ImageNet results in more accurate models than using the other half of CIFAR-10 as $S_{\text{u,OOD}}$. It is interesting that an $S_{\text{u,OOD}}$ data set of type Dif can be more beneficial than a $S_{\text{u,OOD}}$ data set of type Sim which is also the case for FashionMNIST and Tiny ImageNet versus Fashion Product at $n_l = 150$. This contradicts the belief underlying the semantic matching heuristic that unlabelled data that appears semantically more related to the labelled data is always the better choice for SSDL. Given length restrictions, the full results can be found in Table 6 of Appendix B.5.

Rather, as we demonstrate in the second set of experiments below, a notion of dissimilarity between data sets in the feature space of a generic deep classifier can offer a simple and reproducible comparison heuristic. The disimilarity scores for all OOD configurations from the ablation study can be found in Table 2. We can observe that these dissimilarity measures trace the accuracy results found in Table 6. This correlation is quantified in Figure 1a with the cosine based density measure d_c correlating particularly well with the accuracy results of Table 6. Also, the p-values are consistently lower for the density based distances, meaning that density based distances present more confidence as seen in Table 7. The results with italics correspond to p-values lower than 0.05. We suspect that this is related to the quantitative approximation of the feature distribution mismatch implemented both in the d_{JS} and d_{C} distances. In Table 6 we indicate the distance-based preference ranking provided by disimilarity matching with d_C. The non-IID configurations resulting in the best SSDL accuracy are contained in the top two d_C selections seven out of nine times. Note that with the dissimilarity measures we can do this selection *before* SSDL training and thus improve the overall result. For reasons of space, the full results can be found in Table 7 of Appendix B.6.

$\mathbf{S_l}$	$\mathbf{n_l}$	$\mathbf{d_{\ell_1}}$	$\mathbf{d_{\ell_2}}$	$\mathbf{d_{JS}}$	$\mathbf{d_C}$
MNIST	60	-0.876	-0.898	-0.969	-0.944
	100	-0.805	-0.83	-0.786	-0.948
	150	-0.794	-0.822	-0.81	-0.944
CIFAR-10	60	-0.823	-0.853	-0.944	-0.921
	100	-0.826	-0.878	-0.966	-0.947
	150	-0.808	-0.838	-0.952	-0.927
FashionMNIST	60	-0.2	-0.268	-0.735	-0.789
	100	-0.264	-0.326	-0.781	-0.824
	150	-0.286	-0.347	-0.785	-0.827

(a)



(b)

Figure 1: A summary of the data set dissimilarity results for the test bed. (a) Correlation results for the dissimilarity measures between S_l and S_u with OOD contamination and SSDL accuracy. (b) Distribution for feature 159 of a model trained with MNIST labelled data (orange and continuos in both plots), and ImageNet and SVHN unlabelled data (left and right column, respectively, with the blue dashed line in both).

As for the qualitative test results, Figure 1b shows an example of a feature density function approximated for randomly selected samples for the MNIST-ImageNet and MNIST-SVHN data set pairs. The plots reveal a stronger density based similarity between the MNIST and ImageNet than the MNIST and SVHN data sets. This is in spite of perhaps higher semantic similarity between SVHN and MNIST. This correlates well with the quantitative figures yielded in Table 7, where the MNIST data set is more dissimilar to the SVHN data set (MNIST contaminated by 100% with the SVHN data set) than the ImageNet data set (MNIST contaminated by 100% with the ImageNet data set). This also highly correlates with the final SSDL accuracy yielded with both unlabelled data sets (MNIST contaminated by 100% with SVHN and ImageNet) shown in Table 6. MixMatch shows a higher accuracy when using ImageNet as an unlabelled data set compared to using SVHN as unlabelled data. For reasons of space, the full results can be found in Appendix B.7.

3 Discussion

In this work, we introduced the *non-IID-SSDL* sandbox. We demonstrated that semantic similarity matching between labelled and unlabelled data is not a reliable recipe for successful SSDL. We also showed that simple dissimilarity measurements in the feature space of a generic classifier offer a promising and reproducible alternative.

When relaying these results back to application domains we consider the following implications. Real-world SSDL may come with different degrees of OOD data contamination: for instance, with a deep learning model trained for medical imaging analysis, unlabelled data can include images within the same domain, but capturing different pathologies not present in the labelled data. A close scenario has been simulated with the OH setting which resulted in a subtle accuracy degradation in most cases. However, the accuracy gain obtained in our tests vis-a-vis the fully supervised baseline is still substantial. Another plausible real-world scenario for SSDL is the inclusion of widely available unlabelled data sets, e.g. built with web crawlers, where the domain shift can be more substantial. Such scenarios have been simulated with the OOD types similar and different. We can observe that notions of semantic similarity between labelled and unlabelled data set pairings, e.g. (MNIST-SVHN) or (FashionMNIST-Fashion Product), do not necessarily imply an SSDL accuracy gain. Dissimilarity measures, in particular d_C, show promise as a simple proxy for SSDL accuracy, according to our test results. This is visible when comparing the accuracy and distance results of the previous pairings to (MNIST-Tiny ImageNet) and (FashionMNIST-Tiny ImageNet) which have higher accuracies and, also, surprisingly, lower distance measurements. We speculate how this can be related to the quality of the feature extractor that the usage of more diverse data can yield, as the positive results often yielded with self-supervised learning (Zhai et al., 2019).

In future work, we plan to investigate the relationship between generic feature similarity and SSDL downstream performance further. The fact that feature dissimilarity scores can be calculated before SSDL training and independent of the SSDL model poses an interesting profile for application in data augmentation (Hendrycks et al., 2020) or the auditing of machine learning algorithms (Oala et al., 2020). Finally, further exploration of the connections to OOD detection (Zisselman & Tamar, 2020), concept drift (Webb et al., 2016) and class distribution mismatch (Chen et al., 2020) offer interesting avenues for further research in non-IID SSDL.

ACKNOWLEDGMENTS

We would like to thank Jordina Torrents-Barrena and Sören Becker for fruitful discussions during the inception of this work. We would like to thank Shengxiang Yang, Armaghan Moemeni, Wojciech Samek and Miguel A. Molina-Cabello for feedback on the first manuscript.

REFERENCES

Param Aggarwal. Fashion product images (small). `https://www.kaggle.com/paramaggarwal/fashion-product-images-small`.

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

Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *AAAI*, pp. 3569–3576, 2020.

Daniel Cremers and Stefano Soatto. A pseudo-distance for shape priors in level set segmentation. In *2nd IEEE workshop on variational, geometric and level set methods in computer vision*, pp. 169–176. Citeseer, 2003.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Morten W Fagerland and Leiv Sandvik. The wilcoxon–mann–whitney test under scrutiny. *Statistics in medicine*, 28(10):1487–1497, 2009.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

WJ Krzanowski. Non-parametric estimation of distance between groups. *Journal of Applied Statistics*, 30(7):743–750, 2003.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Luis Oala, Jana Fehr, Luca Gilli, Pradeep Balachandran, Alixandro Werneck Leite, Saul Calderon-Ramirez, Danny Xie Li, Gabriel Nobis, Erick Alejandro Muñoz Alvarado, Giovanna Jaramillo-Gutierrez, Christian Matek, Arun Shroff, Ferath Kherif, Bruno Sanguinetti, and Thomas Wiegand. Ml4h auditing: From paper to practice. In Emily Alsentzer, Matthew B. A. McDermott, Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy, and Stephanie L. Hyland (eds.), *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136 of *Proceedings of Machine Learning Research*, pp. 280–317. PMLR, 11 Dec 2020. URL `http://proceedings.mlr.press/v136/oala20a.html`.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 3235–3246, 2018.

Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

Geoffrey I Webb, Roy Hyde, Hong Cao, Hai Long Nguyen, and Francois Petitjean. Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4):964–994, 2016.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):9, 2016.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 2017.

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, pp. 1476–1485, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Yabin Zhang, Bin Deng, Kui Jia, and Lei Zhang. Label propagation with augmented anchors: A simple semi-supervised learning baseline for unsupervised domain adaptation. In *European Conference on Computer Vision*, pp. 781–797. Springer, 2020.

Xujiang Zhao, Killamsetty Krishnateja, Rishabh Iyer, and Feng Chen. Robust semi-supervised learning with out of distribution data, 2021.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1): 44–53, 2018.

Ev Zisselman and Aviv Tamar. Deep residual flow for out of distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13994–14003, 2020.

A Method - Details

A.1 MixMatch Algorithm and Deep Learning Model

In MixMatch, the consistency loss term minimizes the distance of the pseudo-labels and the model predictions over the unlabelled data set X_u. Pseudo-label $\widehat{\boldsymbol{y}}_j$ estimation is performed with the average model output of a transformed input x_j, with K number of different transformations. $K = 2$ is advised in (Berthelot et al., 2019). The estimated pseudo-labels $\widehat{\boldsymbol{y}}$ might be too *unconfident*. To tackle this, pseudo-label sharpening is performed with a temperature ρ. The data set with the estimated and sharpened pseudo-labels was defined as $\widetilde{S}_u = \left(X_u, \widetilde{Y}\right)$, with $\widetilde{Y} = \{\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2, \ldots, \widetilde{\boldsymbol{y}}_{n_u}\}$.

Data augmentation is a key aspect in semi-supervised deep learning as found in (Berthelot et al., 2019). To further augment data using both labelled and unlabelled samples, they implemented the Mix Up algorithm developed in (Zhang et al., 2017). Linear interpolation of a mix labelled observations and unlabelled (with its corresponding pseudo-labels) observations.

$$\left(S'_l, \widetilde{S}'_u\right) = \Psi_{\text{MixUp}}\left(S_l, \widehat{S}_u, \alpha\right) \tag{1}$$

The Mix Up algorithm creates new observations from a linear interpolation of a mix of unlabelled (with its corresponding pseudo-labels) and labelled data. More specifically, it takes two labelled (or pseudo labelled) data pairs $(\boldsymbol{x}_a, y_a)$ and $(\boldsymbol{x}_b, y_b)$. The Mix Up method generates a new observation and its label $(\boldsymbol{x}', y')$ by following these steps:

1. Sample the Mix Up parameter λ from a Beta distribution $\lambda \sim \text{Beta}\,(\alpha, \alpha)$.
2. Ensure that $\lambda > 0.5$ by making $\lambda' = \max\left(\lambda, 1 - \lambda\right)$.
3. Create a new observation with a lineal interpolation of both observations: $\boldsymbol{x}' = \lambda' \boldsymbol{x}_a + \left(1 - \lambda'\right) \boldsymbol{x}_b$.

With the augmented data sets $\left(S'_l, \widetilde{S}'_u\right)$, the MixMatch training can be summarized as:

$$f_{\boldsymbol{w}} = T_{\text{MixMatch}}\left(S_l, S_u, \alpha, \gamma, \lambda\right) = \underset{\boldsymbol{w}}{\text{argmin}} \mathcal{L}\left(S, \boldsymbol{w}\right) \tag{2}$$

$$\mathcal{L}\left(S, \boldsymbol{w}\right) = \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in S'_l} \mathcal{L}_l\left(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{y}_i\right) + r(t)\gamma \sum_{(\boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j) \in \widetilde{S}'_u} \mathcal{L}_u\left(\boldsymbol{w}, \boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j\right) \tag{3}$$

For supervised and semi-supervised loss functions, the cross-entropy $\mathcal{L}_l\left(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{y}_i\right) = \delta_{\text{cross-entropy}}\left(\boldsymbol{y}_i, f_{\boldsymbol{w}}\left(\boldsymbol{x}_i\right)\right)$ and the Euclidean distance $\mathcal{L}_u\left(\boldsymbol{w}, \boldsymbol{x}_j, \widetilde{\boldsymbol{y}}_j\right) = \left\|\widetilde{\boldsymbol{y}}_j - f_{\boldsymbol{w}}\left(\boldsymbol{x}_j\right)\right\|$, are usually implemented, respectively. The regularization γ controls the direct influence on unlabelled data. Since in the first epochs, unlabelled data based predictions are unreliable, the function $r(t) = t/\rho$ increases the unsupervised term contribution as the number of epochs progress. The coefficient ρ is referred as the rampup coefficient. Unlabelled data also influences the labelled data term $\mathcal{L}_l$, since unlabelled data is used also to artificially augment the data set with the Mix Up algorithm. This loss term is used at training time, for testing, a regular cross entropy loss is implemented.

A.2 Dissimilarity Measures

In this work we implement a set of Deep data set Dissimilarity Measure (DeDiM)s based on data set subsampling, as image data sets are usually large, following a sampling approach for comparing two populations, as seen in (Krzanowski, 2003). We compute the dissimilarity measures in the feature space of a generic Wide-ResNet pre-trained on ImageNet, making our proposed approach non-specific to the SSDL model to be trained. This enables an evaluation of the unlabelled data before training the SSDL model. The proposed measures in this work are meant to be simple and quick to evaluate with practical use in mind. We propose and test the implementation of two Minkowski based distance sets, $d_{\ell_2}\left(S_a, S_b, \tau, \mathcal{C}\right)$ and $d_{\ell_1}\left(S_a, S_b, \tau, \mathcal{C}\right)$, corresponding to the Euclidean and Manhattan distances, respectively, between two data sets S_a and S_b. Additionally, we implement and test two non-parametric density based data set divergence measures; Jensen-Shannon (d_{JS}) and cosine distance

(d_C). For all the proposed dissimilarity measures, the parameter τ defines the sub-sample size used to compute the dissimilarity between the two data sets S_a and S_b, and $\mathcal{C}$ the total number of samples to compute the mean sampled dissimilarity measure. The general procedure for all the implemented distances is detailed as follows.

- We randomly sub-sample each one of the data sets S_a and S_b, with a sample size of τ, creating the sampled data sets $S_{a,\tau}$ and $S_{b,\tau}$.
- We transform an input observation $\boldsymbol{x}_j \in S_i$, with $\boldsymbol{x}_j \in \mathbb{R}^n$, being n the dimensionality of the input space, using the feature extractor f, yielding the feature vector $\boldsymbol{h}_j = f\left(\boldsymbol{x}_j\right)$.
- The feature vector $\boldsymbol{h}_i \in \mathbb{R}^{n'}$ has dimension n' dimensions, with $n' < n$. For instance, the implemented feature extractor f uses the ImageNet pretrained Wide-ResNet architecture, extracting $n' = 512$ features. This yields the two feature sets $H_{a,\tau}$ and $H_{b,\tau}$

For the Minkowski based distance sets $d_{\ell_2}\left(S_a, S_b, \tau, \mathcal{C}\right)$, $d_{\ell_1}\left(S_a, S_b, \tau, \mathcal{C}\right)$, we perform the following steps for the sets of features obtained in the previous description $H_{a,\tau}$ and $H_{b,\tau}$:

- For each feature vector $\boldsymbol{h}_j \in H_{a,\tau}$, find the closest feature vector $\boldsymbol{h}_k \in H_{b,\tau}$, using the ℓ_p distance, with $p = 1$ or $p = 2$ for the Manhattan and Euclidean distances, respectively: $\widehat{d_j} = \min_k \left\|\boldsymbol{h}_j - \boldsymbol{h}_k\right\|_p$. We do this for a number of $\mathcal{C}$ samples, yielding a list of distance calculations $d_{\ell_p}\left(S_a, S_b, \tau, \mathcal{C}\right) = \left\{\widehat{d}_1, \widehat{d}_2, ..., \widehat{d}_{\mathcal{C}}\right\}$.
- We compute a reference list of distances for the same list of samples of the data set S_a to itself (intra-data set distance), thereby computing $d_{\ell_p}\left(S_a, S_a, \tau, \mathcal{C}\right)$. This yields a list of reference distances $\check{d}_1, \check{d}_2, ..., \check{d}_{\mathcal{C}}$. In our case S_a corresponds to the labelled data set S_l, as the distance to different unlabelled data sets S_u is to be computed.
- To ensure that the absolute differences between the reference and inter-data set distances $d_c = \left|\widehat{d}_c - \check{d}_c\right|$ are statistically significant, we compute the p-value associated with a Wilcoxon test.
- After the distance set between two data sets $d_{\ell_p}\left(S_a, S_b, \tau, \mathcal{C}\right)$ is obtained, its average reference subtracted distance $\overline{d}$ and its corresponding statistical significance p-value are computed.

As for the density based distances implemented we follow a similar sub-sampling approach, with these steps:

- For each dimension $r = 1, ..., n'$ in the feature space, we compute the normalized histograms to approximate the density functions $p_{r,a}$, in the sample $H_{a,\tau}$. Similarly, we compute the normalized histograms to yield the set of approximate density functions $p_{r,b}$ for $r = 1, ..., n'$, using the observations in the sample $H_{b,\tau}$.
- For the Jensen-Shannon divergence (d_{JS}) and the cosine distance (d_C), we compute the sum of the dissimilarities between the density functions $p_{r,a}$ and $p_{r,b}$, to yield the estimated dissimilarity for the sample j: $\widehat{d_j} = \sum_{r=1}^{n'} \delta_g\left(p_{r,a}, p_{r,b}\right)$, where $g = JS$ and $g = C$ for the Jensen-Shannon divergence and the cosine distance, respectively. We do this for all the $\mathcal{C}$ samples, yielding the list of inter-data set distances: $\widehat{d}_1, \widehat{d}_2, ..., \widehat{d}_{\mathcal{C}}$. To lower the computational burden, we assume that the dimensions are statistically independent.
- Similar to the Minkowski distances, we compute the intra-data set distances for the data set S_a, in this context the labelled data set S_l, to obtain the list of reference distances $\check{d}_1, \check{d}_2, ..., \check{d}_{\mathcal{C}}$.
- Similarly, to verify that the inter- and intra-data set distance differences $d_c = \left|\widehat{d}_c - \check{d}_c\right|$ are statistically significant, we compute the p-value associated with a Wilcoxon test. The distance computation yields the sample mean distance $\overline{d}$ and its statistical significance p-value.

The proposed dissimilarity measures do not fulfill the conditions to be a mathematical pseudo-metric since the distance of an object to itself is not exactly zero and symmetry properties are not fulfilled for the sake of evaluation speed (Cremers & Soatto, 2003). Despite these relaxations, we will see that these dissimilarity measures, especially the two that are density based, are an effective proxy for estimating the $S_{\text{u,OOD}}$ accuracy gain.

B EXPERIMENTS - DETAILS

B.1 HYPERPARAMETERS

B.1.1 GLOBAL HYPERPARAMETERS

Table 3: Global hyperparameters which are kept constant throughout all experiments. This was done in order to isolate the effects of the changing OOD data configurations.

Description	Value
Model architecture used in all tasks	wide_resnet
Number of training epochs	50
Batch size	16
Learning rate	0.0002
Weight decay	0.0001
Rampup coefficient	3000
Optimizer	Adam with 1-cycle policy (Smith, 2018)
K, Number of augmentations	2
T, Sharpening temperature	0.5
α, Parameter for the Beta distribution	0.75
γ, Gamma for the unsupervised loss weight	25

B.1.2 MIXMATCH HYPERPARAMETERS

For supervised and semi-supervised loss functions, the cross-entropy and the Euclidean distance, are used, respectively. The regularization coefficient γ controls the direct influence on unlabelled data. Unlabelled data also influences the labelled data term since unlabelled data is used also to artificially augment the data set with the Mix Up algorithm. This loss term is used at training time, for testing, a regular cross entropy loss is implemented. We use the recommended hyperparameters documented in (Berthelot et al., 2019).

Table 4: MixMatch hyperparameters. All parameters were chosen following the recommendations by (Berthelot et al., 2019).

Symbol	Description	Name in code	Value
K	Number of augmentations	`K_TRANSFORMS`	2
T	Sharpening temperature	`T_SHARPENING`	0.5
α	Parameter for the Beta distribution	`ALPHA_MIX`	0.75
γ	Gamma for the loss weight	`GAMMA_US`	25
-	Whether to use balanced (5) or unbalanced (-1) loss for MixMatch	`BALANCED`	5

B.2 DATA SETS OVERVIEW

If you wish to reproduce any of the experiments data sets are automatically downloaded by the experiment script `ood_experiment_at_scale_script.sh` for your convenience based on which experiment you choose to run.

An overview of the different data sets can be found below. Note that we used the training split of each data set as the basis to construct our own training and test splits for each experimental run.

The Gaussian and Salt and Pepper data sets were created with the following parameters: a variance of 10 and mean 0 for the Gaussian noise, and an equal Bernoulli probability for 0 and 255 pixels, in the case of the Salt and Pepper noise.

Table 5: Information on the data sets used in the experiments. **Format** specifies the format the image files were provided in, **d** specifies the size of the images, **N** specifies the number of samples in the data set, $|\mathcal{Y}|$ specifies the number of classes in the data set, **Relative class distribution** specifies the relative class distribution in the data set.

data set	**Format**	**d**	**N**	$\|\mathcal{Y}\|$	**Relative class distribution**
MNIST (LeCun et al., 1998)	`.jpg`	28 × 28	42,000	10	Uniform
SVHN (Netzer et al., 2011)	`.png`	32 × 32	73,557	10	0.07/0.19/0.14/0.12/0.1/ 0.09/0.08/0.07/0.07/0.07
Tiny ImageNet (Deng et al., 2009)	`.jpg`	64 × 64	100,000	200	Uniform
CIFAR-10 (Krizhevsky et al., 2009)	`.jpg`	32 × 32	50,000	10	Uniform
FashionMNIST (Xiao et al., 2017)	`.png`	28 × 28	60,000	10	Uniform
Fashion Product (Aggarwal)	`.jpg`	60 × 80	44,441	5	0.48/0.25/0.21/0.05/0.01
Gaussian	`.png`	224 × 224	20,000	NA	NA
Salt and Pepper	`.png`	224 × 224	20,000	NA	NA

B.3 Preprocessing

Each data point was preprocessed in the following way. After a subset of labelled and unlabelled data for an experimental run had been constructed the means and standard deviations (one pair for labelled data, one pair for unlabelled data) were calculated for this specific subset. Then, the labelled and unlabelled inputs were standardized by subtracting the respective mean and dividing by the respective standard deviation.

In addition, in situations when the size of the unlabelled images differed from the size of the labelled images up- or downsampling was used to align the unlabelled image size.

B.4 Hardware

Experiments were run on three machines. Machine 1 has one 12GB NVIDIA TITAN X GPU, 24 Intel(R) Xeon(R) CPU E5-2687W v4 @ 3.00GHz and 32GB RAM. Machine 2 has four 16GB NVIDIA T4 GPUs, 44 CPUs from the Intel Xeon Skylake family and 150GB RAM. Machine 3 has one 12GB NVIDIA TITAN V GPU, 24 Intel(R) Xeon(R) E5-2620 0 @ 2.00GHz CPU and 32GB RAM. Experimental runs were parallelized using the ampersand option in `bash` executing 10 runs in parallel on a single GPU. With the current code base this requires up to 10 CPUs per GPU as well as approximately 25GB RAM per GPU. With this setup a single training epoch of 10 parallel experimental runs should last between 2 and 4 minutes per GPU, depending on which type of GPU is used.

B.5 *non-IID-SSDL* Sandbox - Full Results

Table 6: Results for the class distribution mismatch experiment, best OOD performance in bold per configuration. Each result entry in the table represents the mean and variance of accuracy across ten random experimental runs per entry. For a detailed description of symbols and the experiment see Section 2. Numbers in parentheses next to $S_{\text{u,OOD}}$ entries reflect the preference ranking provided by the dissimilarity measure with d_C.

$\mathbf{S_{IOD}}$	$\mathbf{T_{OOD}}$	$\mathbf{S_{uOOD}}$	$\%_{\mathbf{uOOD}}$	$\mathbf{n_l}$		
				60	100	150
MNIST	Fully supervised baseline			0.457 ± 0.108	0.559 ± 0.125	0.645 ± 0.101
	SSDL baseline (no OOD data)			0.704 ± 0.096	0.781 ± 0.065	0.831 ± 0.0626
	OH	MNIST	50	$\mathbf{0.679 \pm 0.108}$	$\mathbf{0.769 \pm 0.066}$	0.802 ± 0.054
			100	0.642 ± 0.111	0.746 ± 0.094	0.798 ± 0.07
	Sim	SVHN (3)	50	0.637 ± 0.098	0.745 ± 0.081	0.801 ± 0.0699
			100	0.482 ± 0.113	0.719 ± 0.058	0.765 ± 0.072
	Dif	TI (2)	50	0.642 ± 0.094	0.739 ± 0.074	0.809 ± 0.066
			100	0.637 ± 0.097	0.732 ± 0.074	0.804 ± 0.071
		GN (5)	50	0.606 ± 0.0989	0.713 ± 0.087	0.786 ± 0.065
			100	0.442 ± 0.099	0.461 ± 0.073	0.542 ± 0.062
		SAPN (4)	50	0.631 ± 0.102	0.735 ± 0.082	$\mathbf{0.813 \pm 0.057}$
			100	0.48 ± 0.0951	0.524 ± 0.09	0.563 ± 0.095
CIFAR-10	Fully supervised baseline			0.380 ± 0.024	0.445 ± 0.042	0.449 ± 0.022
	SSDL baseline (no OOD data)			0.453 ± 0.046	0.474 ± 0.019	0.501 ± 0.033
	OH	CIFAR-10	50	$\mathbf{0.444 \pm 0.040}$	0.472 ± 0.039	0.525 ± 0.050
			100	0.443 ± 0.023	0.472 ± 0.047	0.499 ± 0.054
	Sim	TI (1)	50	0.435 ± 0.054	0.473 ± 0.039	$\mathbf{0.543 \pm 0.040}$
			100	0.417 ± 0.020	$\mathbf{0.480 \pm 0.039}$	0.498 ± 0.042
	Dif	SVHN (2)	50	0.419 ± 0.027	0.464 ± 0.044	0.469 ± 0.056
			100	0.385 ± 0.034	0.418 ± 0.035	0.440 ± 0.046
		GN (4)	50	0.409 ± 0.047	0.454 ± 0.048	0.491 ± 0.032
			100	0.297 ± 0.029	0.306 ± 0.034	0.302 ± 0.038
		SAPN (5)	50	0.438 ± 0.029	0.455 ± 0.037	0.485 ± 0.034
			100	0.236 ± 0.031	0.246 ± 0.032	0.232 ± 0.022
FashionMNIST	Fully supervised baseline			0.571 ± 0.073	0.704 ± 0.066	0.720 ± 0.093
	SSDL baseline (no OOD data)			0.715 ± 0.049	0.760 ± 0.044	0.756 ± 0.069
	OH	FashionMNIST	50	$\mathbf{0.714 \pm 0.049}$	0.721 ± 0.104	0.765 ± 0.053
			100	0.660 ± 0.061	0.711 ± 0.090	0.747 ± 0.061
	Sim	FP (4)	50	0.707 ± 0.039	0.724 ± 0.030	0.778 ± 0.078
			100	0.546 ± 0.101	0.542 ± 0.099	0.540 ± 0.105
	Dif	TI (2)	50	0.690 ± 0.065	$\mathbf{0.745 \pm 0.093}$	0.792 ± 0.058
			100	0.690 ± 0.073	0.728 ± 0.066	$\mathbf{0.794 \pm 0.056}$
		GN (5)	50	0.644 ± 0.061	0.689 ± 0.075	0.755 ± 0.055
			100	0.352 ± 0.025	0.366 ± 0.065	0.361 ± 0.057
		SAPN (3)	50	0.671 ± 0.072	0.708 ± 0.095	0.729 ± 0.088
			100	0.276 ± 0.069	0.297 ± 0.046	0.283 ± 0.059

B.6 DISSIMILARITY MEASURES - FULL RESULTS

Table 7: Distance measures between the labelled and unlabelled datasets S_l and S_u. Numbers in italics correspond to results with $p > 0.05$ for the Wilcoxon test. For a detailed description of symbols and the experiment see Section 2

$\mathbf{S_l}$	$\mathbf{S_u}$	$\%_{\mathbf{uOOD}}$	$\mathbf{d}_{\ell_2}$	$\mathbf{d}_{\ell_1}$	$\mathbf{d_{JS}}$	$\mathbf{d_C}$
MNIST	OH	50	*0.011 ± 0.006*	*0.459 ± 0.28*	*0.266 ± 0.221*	*0.811 ± 0.512*
		100	*0.014 ± 0.019*	*0.38 ± 0.507*	1.001 ± 0.725	1.263 ± 0.665
	SVHN	50	*0.09 ± 0.017*	1.569 ± 0.504	6.789 ± 0.924	12.021 ± 1.757
		100	0.25 ± 0.053	4.702 ± 1.04	52.349 ± 3.292	42.026 ± 4.31
	TI	50	0.008 ± 0.023	1.519 ± 0.223	3.663 ± 0.772	5.512 ± 0.767
		100	0.217 ± 0.04	4.3 ± 0.636	10.305 ± 1.667	15.18 ± 2.698
	GN	50	0.11 ± 0.0219	1.958 ± 0.534	14.785 ± 1.052	23.593 ± 1.859
		100	0.357 ± 0.081	5.987 ± 1.091	52.349 ± 4.253	86.21 ± 3.471
	SAPN	50	0.089 ± 0.0311	2.479 ± 0.7433	15.116 ± 1.475	20.151 ± 1.619
		100	0.323 ± 0.07	6.308 ± 1.366	53.397 ± 4.253	77.456 ± 4.474
CIFAR-10	OH	50	*0.056 ± 0.023*	*0.915 ± 0.934*	*0.338 ± 0.325*	0.892 ± 0.402
		100	*0.061 ± 0.04*	0.769 ± 0.461	*0.451 ± 0.41*	0.648 ± 0.407
	TI	50	0.082 ± 0.037	*0.928 ± 0.815*	*0.388 ± 0.243*	*0.423 ± 0.362*
		100	*0.056 ± 0.048*	*0.992 ± 0.517*	*0.469 ± 0.426*	0.415 ± 0.232
	SVHN	50	*0.055 ± 0.032*	*0.948 ± 0.699*	*0.665 ± 0.565*	*0.414 ± 0.357*
		100	*0.075 ± 0.036*	1.291 ± 0.925	0.736 ± 0.658	0.581 ± 0.343
	GN	50	*0.107 ± 0.083*	1.344 ± 1.156	1.708 ± 0.421	3.001 ± 0.696
		100	0.127 ± 0.087	1.531 ± 0.767	5.855 ± 0.552	8.703 ± 0.926
	SAPN	50	0.1146 ± 0.044	1.854 ± 0.894	2.299 ± 0.691	2.56 ± 0.762
		100	0.208 ± 0.05	5.502 ± 1.156	8.225 ± 0.866	9.554 ± 0.489
FashionMNIST	OH	50	*0.02 ± 0.012*	*0.34 ± 0.162*	*0.669 ± 0.566*	*0.575 ± 0.423*
		100	0.059 ± 0.032	0.801 ± 0.402	0.305 ± 0.237	0.774 ± 0.343
	FP	50	0.105 ± 0.0526	2.168 ± 0.774	7.263 ± 0.622	5.305 ± 0.405
		100	0.195 ± 0.0457	4.819 ± 1.077	9.056 ± 0.462	11.286 ± 0.751
	TI	50	*0.04 ± 0.03*	*0.798 ± 0.542*	*0.897 ± 0.516*	0.897 ± 0.516
		100	0.065 ± 0.03	1.66 ± 0.45	1.4 ± 0.488	1.912 ± 0.683
	GN	50	*0.047 ± 0.03*	*0.533 ± 0.347*	2.819 ± 0.703	3.843 ± 0.704
		100	0.074 ± 0.041	1.325 ± 0.631	9.042 ± 0.699	15.511 ± 0.445
	SAPN	50	0.036 ± 0.022	0.52 ± 0.303	2.799 ± 0.497	2.799 ± 0.497
		100	0.076 ± 0.044	1.411 ± 0.548	8.464 ± 0.553	8.464 ± 0.553

B.7 Feature Comparison - Additional Samples

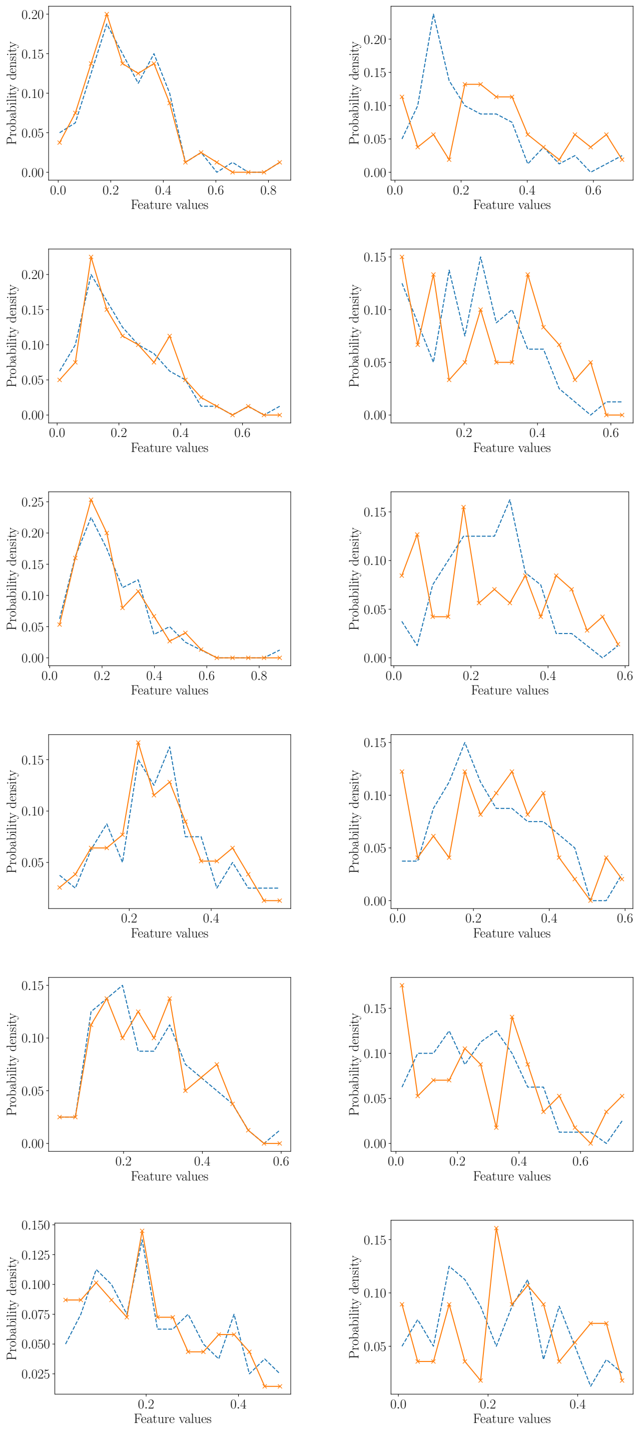


Table 8: Feature distribution for a model trained with MNIST labelled data (orange and continuos in both plots), and ImageNet and SVHN unlabelled data (left and right column, respectively, blue dashed line in both). For each plot a different dataset partition was used. From top to bottom, for each row: feature 372, 159, 7, 420, 82 and 491.