Domain-based semi-supervised learning: exploiting label invariance in unlabeled data from distributed cameras

Leonardo Taccari
Verizon Connect Research, Florence, Italy
leonardo.taccari@verizonconnect.com

Abstract

In several practical supervised learning problems where we have a large amount of data from distributed cameras or sensors, we can use domain knowledge to identify subsets of unlabeled examples with the same (unknown) label. Under this assumption, we propose a straightforward way to exploit label invariance in unlabeled data within a domain-aware semi-supervised learning framework (DSSL). Our approach exploits such invariance to generate higher quality pseudolabels to be used in a consistency loss term.

We report experiments and ablation studies on three practical cases on data from real-world fleets of connected vehicles that naturally exhibit the required assumption: an image classification problem, a semantic segmentation task, and a time series classification one. We show that our approach is extremely effective, especially when few labeled samples are available, and can be easily adapted to tasks of different nature.

1. Introduction

In several real-world tasks commonly solved with supervised learning, collecting a large dataset of annotations is too costly, time-consuming, or simply infeasible, while a vast amount of unlabeled data is often available. In the last few years, a large number of methods have been proposed that carried to a renewed interest in self- and semi-supervised learning.

While these methods have obtained great success, they are often rather complex and brittle. We argue that, even in unsupervised data, there often are properties obtained from knowledge of the domain or the data generation process that one can directly exploit to make similar semi-supervised techniques more effective.

In this article we show a way to do so for the case of networks of sensors or cameras installed on moving vehicles: we propose a simple and practical domain-aware semi-supervised method, that leverages known invariants in unlabeled data, bringing about a significant increase in accuracy with minimal implementation effort. The idea is to exploit the existence of equivalence classes in unlabeled data where the target label (albeit unknown) is invariant. For example, this would be possible in situations where we can obtain multiple images of the same object from different point of views, or from different moments in time. This would allow us to naturally have multiple views of the same sample, rather than relying only on artificial data augmentations.

The proposed approach can be applied on a vast array of tasks that exhibit some fairly common assumptions in real-world scenarios where we have a network of distributed sensors that generate data. We show how the idea can be easily applied in practice on three applications, with a simple setup where we employ standard training procedures with no bells and whistles. In the first task, weather classification from videos, we exploit temporal consistency in single video clips; in the other two examples, ego-vehicle segmentation and vehicle type classification, we exploit the consistency of the target across different examples obtained from the same camera or sensor.

2. Related work

Semi-supervised learning refers to techniques that attempt to use jointly labeled and unlabeled data for learning [3]. In the last few years, a large number of new methods have been proposed that apply semi-supervised learning to neural networks. Recent approaches typically rely on some flavor of self-training, where a model produces artificial labels (pseudolabels) used to train itself: for instance, Temporal Ensembling [10] uses predictions averaged across different epochs to produce better artificial targets; similarly, MeanTeacher [21] uses and exponential average of the model weights to refine pseudolabels. Other works [28, 15, 4] use a separate teacher network to produce labels used for a student model.

Among the most successful approaches, let us briefly highlight approaches that bear most similarities with ours, based on the assumption that a model should output simi-

lar predictions on differently perturbed versions of the same example. UDA [27] introduces a set of advanced augmentation strategies that are used to produce predictions over different versions of the same image. Then, a divergence metric between the two distributions (perturbed vs unperturbed) is minimized. MixMatch [2] aggregates predictions on K differently transformed version of an image to obtain an artificial label. Labeled and pseudolabeled samples are then combined via *mixup* [30]. In FixMatch [18], the authors use an even simpler combination of pseudolabeling and consistency regularization. Pseudolabels are obtained thresholding predictions on weakly augmented data, and then are used in a loss against the prediction obtained from a heavily augmented view of the same input. In all these methods the choice of augmentations is crucial for their effectiveness, as remarked in [27, 2]. In Meta Pseudo Labels [15], a teacher network is used to generate pseudolabels to teach a student network. The teacher is constantly adapted by the feedback of the student's performance on the labeled dataset.

Semi-supervised techniques that, similar to what we propose, attempt to exploit specific invariants from the domain or the data at hand have also been proposed in the past. A line of research has explored the use of temporal coeherence to help learning in an unsupervised or semi-supervised manner, see [14, 13]. In [26], the authors exploit spatial and temporal consistency as loss regularization terms to develop a semi-supervised method for pedestrian counting. In [12], the authors use domain knowledge on specific relations between different examples. They propose a technique that attempts to enforce consistency of such relations, rather than of the unknown label. The method we propose is also similar in spirit to [1], where the authors exploit geographical invariance in the data in a self-supervised framework, and [8] where the authors exploit the knowledge that images from different medical sensors are aligned.

3. Domain-aware Semi-Supervised Learning

Let us consider a supervised learning problem, with a set $X_s = \{(x_i, y_i)\}$ of labeled pairs, where $x_i \in \mathcal{X}$ is the i-th training example and y_i is the i-th target or label. Let $f(\cdot\,;\theta) : \mathcal{X} \to \mathcal{Y}$ the prediction function of our current model parametrized by the weights θ.

Let $\mathcal{U} \subset \mathcal{X}$ be a set of unlabeled examples. Let $\sim$ be an equivalence relation, that can be derived from domain knowledge, which is known to be label-consistent.

Definition 1. A relation $\sim$ over $\mathcal{U}$ is said to be *label-consistent* if for each pair $u, v \in \mathcal{U}$, $u \sim v \implies g(u) = g(v)$, where $g : \mathcal{X} \to \mathcal{Y}$ is the function that maps each example to its ground truth label.

In other words, $\sim$ induces a partitioning of $\mathcal{U}$ into equivalence classes $G^j \subset \mathcal{U}$ that have homogeneous (albeit *unknown*) label. As an example, assume we have clips containing footage of a single animal each. If our task entails learning to classify animals in an image, we can exploit the domain knowledge that two frames a, b belonging to the same video ($a \sim b$) are going to belong to the same category ($g(a) = g(b)$).

In Domain-aware Semi-Supervised Learning (DSSL), we leverage the knowledge of a label-consistent relation $\sim$ over our unlabeled data, arising naturally from the data collection or generation process in the context of network of distributed sensors. Knowing that we can partition unlabeled data in groups of examples with the same (unknown) label allows us to:

- use different data points in the same label-consistent group effectively as different *views* of the same sample (i.e., multiple views across time), reducing the need of finding the correct mix/recipe of data augmentation for the task at hand;
- improve the quality of the pseudolabels aggregating predictions from multiple examples within a group;
- use the same high-quality pseudolabel for multiple examples within the same group;
- measure the quality of the pseudolabel of a group.

In DSSL, we construct a loss function as the sum of two terms: a standard supervised loss $\mathcal{L}_s$, applied to labeled data, and an unsupervised one, $\mathcal{L}_u$, which is a sort of *consistency loss* computed with pseudolabels obtained from unlabeled data. The two loss terms are combined with a weight term:

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_u. \tag{1}$$

In order to compute the unsupervised loss term (see Figure 1), at each iteration we consider B_u groups of examples that are equivalent under the relation $\sim$. Our assumption is that a model should output similar predictions on all examples in the same group. For each group $G^j, j \in 1, \ldots, B_u$, we aggregate the model output computed over the examples in G^j to obtain a single pseudolabel $\tilde{y}^j$ and a pseudolabel score σ^j:

$$\tilde{y}^j, \sigma^j = agg(\{f(u;\theta) : u \in G^j\}) \tag{2}$$

In practice, a pseudolabel will typically not be computed on *all* examples in G^j, that might be infinitely many, but only using a subset $U' \subset G^j$ of cardinality K'. Finally, if the score σ^j is greater or equal than a quality threshold τ, the resulting pseudolabel $\tilde{y}^j$ will be used to compute a loss against K'' other examples $U'' \subset G^j$, on which one can also apply a stochastic augmentation function $\alpha(\cdot)$. The total unsupervised loss term $\mathcal{L}_u$ is obtained summing up the terms $\mathcal{L}_u^j$ for all groups $j \in 1, \ldots, B_u$.

A summary of the batch update of DSSL can be found in Algorithm 1. It is worth noting that DSSL can also be seen

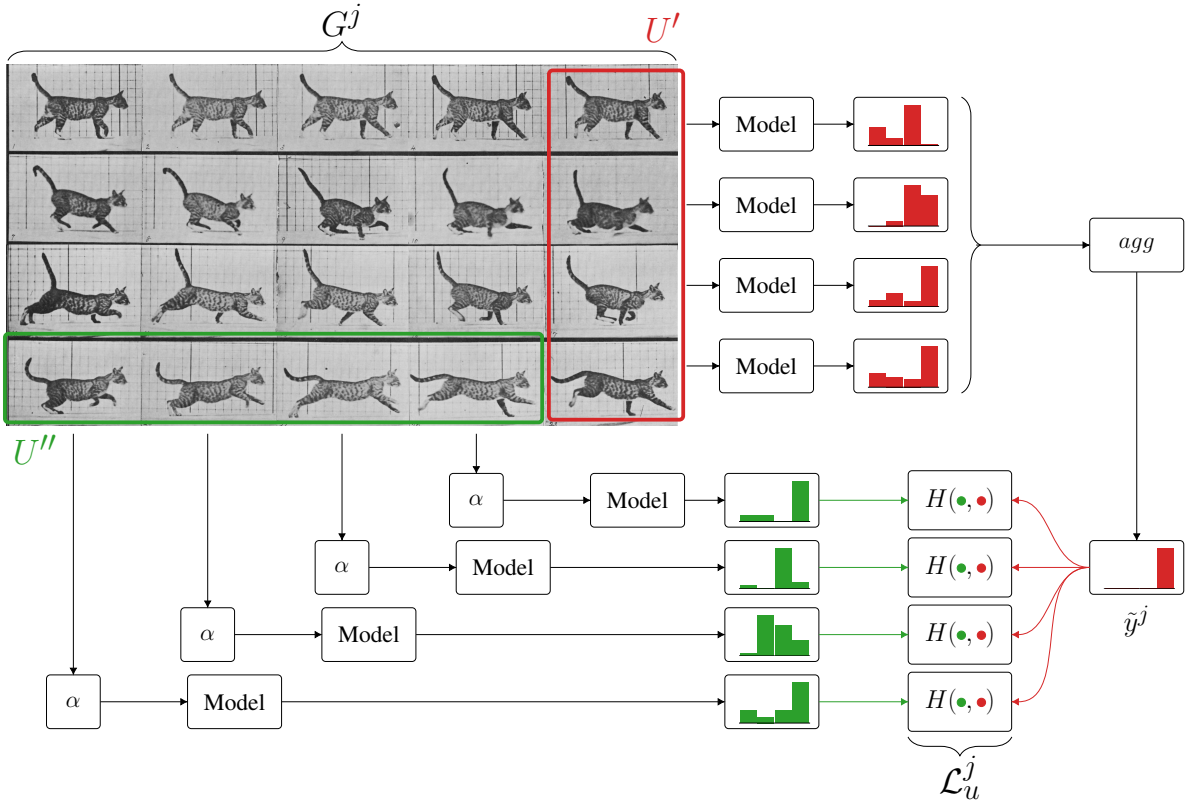

Figure 1. **Domain-aware Semi-supervised Learning**. Given a group G^j of unlabeled examples, known to belong to the same target class, DSSL builds a high-quality pseudolabel $\tilde{y}^j$ aggregating the output of multiple samples U' sampled from G^j. Then, the resulting pseudolabel is used as a target to compute the loss $\mathcal{L}_u^j$ over another set of examples $U'' \subset G^j$ that potentially underwent a data augmentation α. *(Picture adapted from a photographic study by Eadweard Muybridge.)*

as an extension/modification of existing methods based on self-consistency: if we take $U' = U''$ with size 1, DSSL essentially reduces to the core idea of FixMatch [18], while the idea of aggregating K different predictions to produce a pseudolabel was used, among others, in [27, 2].

4. Experiments

We provide a set of experiments on three real-world tasks in the context of connected vehicles, where we show how the use of our approach leads to a significant performance improvement.

4.1. Image classification: weather condition

The BDD100K dataset[1] [29] includes thousands of 40-second videos recorded by dashcams around the US. In each video a single frame is annotated with a label related to the weather condition of the image: *Clear*, *Partly Cloudy*, *Overcast*, *Rainy*, *Snowy*. Assuming that it is highly unlikely that weather conditions would change during a short time span, each video can be seen as a label-consistent group of images, sharing the same (unknown) label. Examples can be seen in Fig. 2.

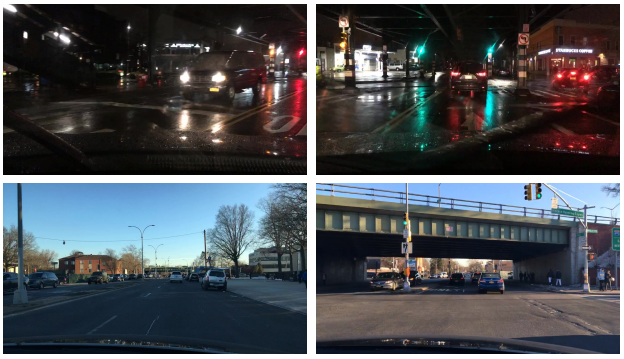

Figure 2. Pairs of frames from the same BDD100K video. The two images on top are extracted from a *rainy* night video (notice the wet road); the images on the bottom are extracted from a *clear* video.

For our experiments, similarly to common semi-supervised learning (SSL) benchmarks, we build a num-

[1]BDD100K is freely available at https://bdd-data.berkeley.edu/ for educational, research, and not-for-profit purposes.

Algorithm 1. Batch update for Domain-aware Self-Supervised Learning

Labeled data: B_s pairs of example x_i and corresponding target y_i
Unlabeled data: B_u groups $G^j \subset \mathcal{U}$, where each group contains unlabeled examples $u_k^j \in G^j$ with consistent (yet unknown) target
for $j = 1, \ldots, B_u$ **do**
  Sample a subset of examples $U' \subset G^j$ of cardinality K'
  Compute candidate pseudolabel and score as: $\tilde{y}^j, \sigma^j = agg(\{f(u;\theta) : u \in U'\})$
  where $agg(\cdot)$ is an aggregation function over the set of predictions computed on U'.
  if $\sigma^j \geq \tau$ **then**
    Sample a subset of examples $U'' \in G^j$ of cardinality K''
    Compute loss for group j as: $\mathcal{L}_u^j = \frac{1}{n''_u} \sum_{u \in U''} H(f(\alpha(u);\theta), \tilde{y}^j)$
    where α is a stochastic augmentation function and H is the loss for a single example.
  end if
end for
Compute supervised loss as: $\mathcal{L}_s = \frac{1}{B_s} \sum_{i=1}^{B_s} H(f(\alpha(x_i);\theta), y_i)$
Return total batch loss: $\mathcal{L} = \mathcal{L}_s + \frac{\lambda_u}{B_u} \sum_{j=1}^{B_u} \mathcal{L}_u^j$

ber of datasets in a hierarchical way, with increasing number of labeled examples. We indicate each of these training datasets as `BDD-Weather-[N]`, where $N \in \{250, 500, 1000, 2500\}$ is the number of total annotated images. For the unsupervised part, we use unlabeled videos as label-consistent groups from which we extract frames that are 4 seconds apart, to ensure images in the same group are not too similar. We keep a separate validation set, with 500 examples per class, and train and test at 320×180.

Our model consists of an EfficientNet-B1 [20] architecture trained with Cross-Entropy Loss and Adam [9] as optimizer. The problem is balanced, so we use accuracy to measure the performance of the approaches. For DSSL, we use $B_u = 8$ groups of size 4 both for pseudolabel aggregation (U') and consistency loss (U''). As aggregation function, we average the predictions over U' and take their $\arg\max$ (hard pseudolabel). The score σ^j is computed as the maximum confidence after the averaging.

In Figure 3 and Table 1 we show the results we obtain compared to a purely supervised training and an SSL baseline. Our baseline SSL is a barebone implementation of FixMatch with RandAugment (RA): we implement the core idea of FixMatch, self-training with consistency regularization using RA, but we do not employ the full arsenal of tricks employed in the original article. On the contrary, for all the methods in this experiment we use the same simple setup and standard choices for optimizer (Adam) and hyperparameters, allowing for our experiments to run on a single GPU Nvidia P100.

DSSL significantly outperforms the supervised training counterpart, especially in scarce-data regime, where it is much more label efficient. DSSL accuracy is often better than a supervised training with *twice* the amount of annotations – as an example, with 500 labeled examples, DSSL

	BDD-Weather			
Num. labels	250	500	1000	2500
Supervised	64.9	68.2	72.3	75.9
SSL baseline	68.1	71.1	75.5	77.3
DSSL	**69.0**	**73.0**	**76.0**	**78.4**

Table 1. DSSL vs supervised and semi-supervised baselines.

yields an accuracy of 73.0, compared to 72.3 obtained for $N = 1000$ with supervised training.

The SSL baseline, in all its simplicity, is very effective: this is a testament to the quality of the core idea of FixMatch. Yet, exploiting knowledge of label-invariance in the domain, DSSL outperforms it in all cases.

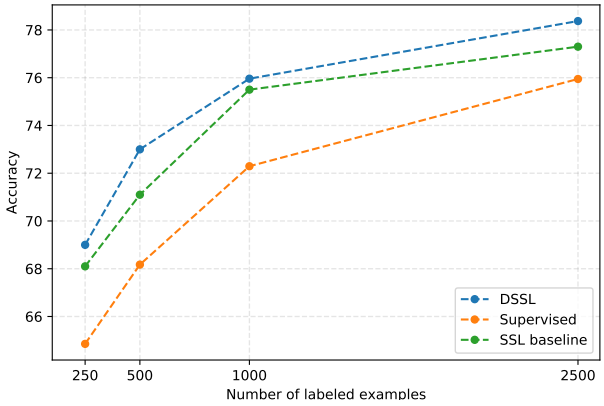

Figure 3. Performance of DSSL vs baselines on `BDD-Weather` with different amount of labeled data. Average over 5 runs.

We perform an ablation study on `BDD-Weather-500` modifying DSSL in two ways. In Table 4 we consider the

case where we change the augmentation policy. Replacing the aggressive RandAugment policy with weaker augmentation strategies causes a drop for both DSSL and the baselines. However, it is interesting to note that the SSL baseline, that relies heavily on the quality of the augmentation strategy for the consistency loss, suffers a greater drop. In the case were we completely remove the augmentation, the SSL baseline collapses (-7.3), with self-training becoming detrimental. DSSL is still rather effective even with no augmentation at all, outperforming the supervised baseline by more than 5 points. This suggests that, when applicable, DSSL might be significantly easier to use effectively than other available semi-supervised and self-supervised learning approaches.

In Table 5 we show the effect of changing the group size used in DSSL. Group size is probably the single parameter with most impact on the effectiveness. The topmost results show that groups are useful both when generating pseudolabels (U'), than when computing the loss from different views (U''). The table at the bottom suggest that, while larger groups should provide better quality pseudolabels, using a rather small number of items from each group seems ideal: we argue that the lower diversity contained in a batch might have an adverse effect. In addition, larger groups are also less practical, as they slow down the computation for marginal improvements.

	B_u	K', K''	Augmentation	Accuracy
DSSL	8	$(4, 4)$	RandAugment	**73.0**
SSL baseline	32	-	RandAugment	71.1
Supervised	-	-	RandAugment	68.2
DSSL	8	$(4, 4)$	Weak	**70.9** (-2.1)
SSL baseline	32	-	Weak	68.2 (**-2.9**)
Supervised	-	-	Weak	66.1 (-2.1)
DSSL	8	$(4, 4)$	None	**70.0** (-3.0)
SSL baseline	32	-	None	63.8 (**-7.3**)
Supervised	-	-	None	64.6 (-3.6)

Figure 4. Impact of augmentation strategy. DSSL is still rather effective even using no artificial augmentation at all, while the SSL baseline collapses.

4.2. Ego-vehicle semantic segmentation

As a second practical use case, we consider a semantic segmentation task: the detection of the ego-vehicle in video frames. Footage captured by dashcams, as BDD100K or Cityscapes [6], often contains a visible part of the ego-vehicle. This can be an issue for any downstream task, as mentioned also in [4]. Then, a common preliminary step entails the identification of the ego-vehicle within a frame.

If we have different images captured from the same vehicle over time, one can safely assume that the mounting position of the camera is constant, and as such, the portion

	K'	K''	Accuracy
DSSL	4	4	**73.0**
	4	1	72.4
	1	4	71.7
	1	1	70.9
DSSL	10	10	71.8
	8	8	72.4
	4	4	**73.0**
	1	1	70.9

Figure 5. Impact of group size for DSSL. Reducing the group size is detrimental both for U' and U''.

of visible ego-vehicle is going to be the same in all captured frames. We can take different frames from the same camera as a label-consistent group of examples.

We use a private dataset of images collected from dashcams installed on a large fleet of vehicles in the US. Vehicles in the dataset are heterogeneous, ranging from passenger cars to heavy-duty trucks, and there is a great variability in terms of camera position, weather and lighting conditions (see Figure 6 for some examples). We labeled a set of 2000 images, from which we extract smaller nested datasets of size $N = \{100, 200, 400, 1000, 2000\}$. In addition, we have a set of more than 70k unlabeled images, gathered by 7k vehicles that were not already in the annotated dataset.

Figure 6. Images with visible ego-vehicle and example of ground truth mask. The dataset contains a large variety of vehicle types, mounting positions, and lighting conditions.

We choose a model commonly used for segmentation, U-Net [16], with an EfficientNet-B0 [20] backbone. We optimize the Dice loss [19] with Adam and we evaluate the results in validation and test computing the Intersection over Union (IoU) between the prediction and the ground truth mask. We use a custom set of data augmentations. For DSSL, we use batches with $B_u = 5$ unlabeled groups. For each group, we take U' of size 4 for pseudolabel generation, and U'' of size 4 to compute the consistency loss. As pseudolabel aggregation function, we compute the pixelwise mean of the predictions over U' to produce the can-

didate mask. The score σ^j is computed as the average pixelwise standard deviation over the predictions; in this case, we keep a pseudolabel if the score is below the threshold $\tau = 0.05$, indicating that the variance of the predictions in the group U' is small.

We report the results compared to a supervised and SSL baseline in Figure 7 and Table 2. DSSL is extremely effective leveraging annotated data in this task. With only 200 labeled examples, our approach outperforms a standard supervised training that uses the entire dataset. When both DSSL and the supervised training use the entire dataset, DSSL reaches a IoU of 90.6, while the supervised baseline only scores 85.9. The simple SSL baseline appears to fail on this task, suggesting that the custom augmentation we used, though aggressive, are not sufficient.

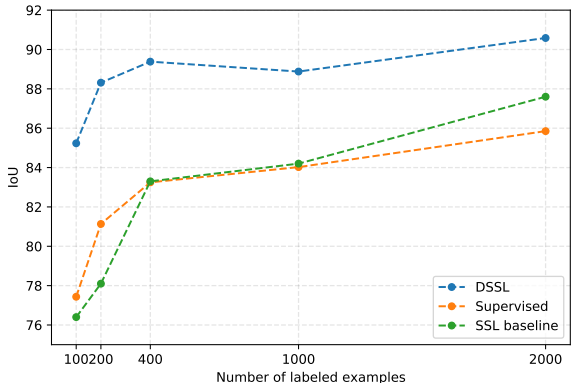

Figure 7. DSSL on ego-vehicle segmentation with different amount of labeled data. Average of 5 runs.

	Ego-vehicle segmentation				
Num. labels	100	200	400	1000	2000
Supervised	77.4	81.1	83.2	83.73	85.9
SSL baseline	76.4	78.1	83.3	84.02	87.6
DSSL	**85.2**	**88.3**	**89.4**	**89.24**	**90.6**

Table 2. DSSL vs supervised and SSL baselines. Average over 5 runs.

4.3. Vehicle type classification

As our third use case, let us consider a time series classification problem on data collected by navigation sensors (accelerometer, gyroscope, and GPS) from a large fleet of connected vehicles in the US. Given sensor data from a device mounted on a vehicle, a question that may arise is identifying the category of the vehicle from the dynamics captured in the data (see, e.g., [24, 17]). We specifically focus on the problem of classifying vehicles into one of 9 fine-grained GVWR (gross vehicle weight rating) classes, ranging from light-duty to heavy-duty trucks, and an additional class for passenger cars, for a total of 10 classes.

The classes are described in Table 3. We employ a fine-grained classification obtained merging a few common ones [25, 23], which is mainly based on the allowed weight limit.

Class	Duty classification	Weight limit
0	Cars	-
1	Light trucks	$<$ 6,000 lbs
2a	Light trucks	6,001–8,500 lbs.
2b	Light/medium trucks	8,501–10,000 lbs.
3	Medium trucks	10,001–14,000 lbs.
4	Medium trucks	14,001–16,000 lbs.
5	Medium trucks	16,001–19,500 lbs.
6	Medium trucks	19,501–26,000 lbs.
7	Heavy trucks	26,001–33,000 lbs.
8	Heavy trucks	$\geq 33,001$ lbs.

Table 3. Vehicle classes

The labels for our training and validation sets were obtained from the make and model, that was available for a small subset of the available data. For the rest of the data, no information is available other than an ID that uniquely identifies the vehicle. The assumption of DSSL are trivially satisfied: in unlabeled data, we know that all samples that are collected from the same vehicle clearly belong to the same target class.

Our dataset contains short segments (15 seconds) of temporally aligned sensor data $x = \{a(t), g(t), v(t)\}$: acceleration @100Hz from a tri-axial accelerometer, angular speed @100Hz from a gyroscope, and speed from GPS @1Hz. The training set contains 8000 labeled samples with one of the 10 classes, while we keep and held-out validation of 4000 samples. In addition, we have a set of over 50,000 unlabeled samples that are grouped based on the vehicle that generated it (whose class is unknown).

We use a simple custom architecture based on a CNN backbone followed by a layer of stacked Gated Recurrent Units (GRU [5]). We use Adam as optimizer, and DSSL with unlabeled batch size $B_u = 32$, $K' = 4$ and $K'' = 4$.

In this domain, there is clearly not the same number of well-studied artificial augmentations and policies as for image-related tasks. Moreover, it is not trivial to understand what kind of operations might affect the semantics of a signal: the input has three modalities (acceleration, angular speed, linear speed) that are correlated, but come from different sensors and have different sampling rates. In our experiments we use a custom set of transformations, inspired by those mentioned in [11, 22], that include axial rotations, additive noise, several kinds of filtering, warping, and cut-out of parts of the signal.

The results in Table 4 confirm that DSSL works very ef-

fectively compared both to the supervised and SSL baseline. The SSL baseline would likely need a more specific training procedure and augmentation strategy to work satisfactorily, while DSSL shows again that it is not too sensitive to the specific augmentation recipe, model architecture or training procedure, working essentially out of the box. This makes the core idea of DSSL especially suitable to be adapted to different domains, that might lack the broad set of established techniques developed for images.

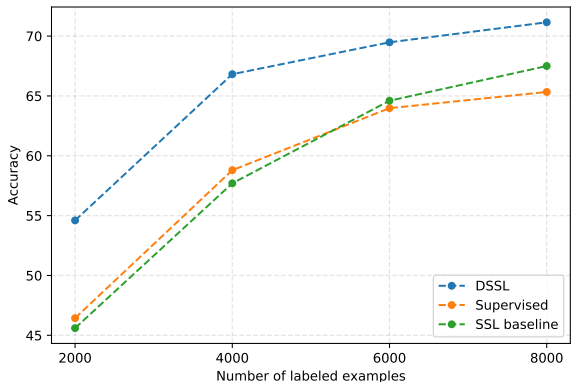

Figure 8. DSSL on the vehicle type classification problem with different amount of annotated data. Average of 5 runs.

	Vehicle type classification			
Num. labels	2000	4000	6000	8000
Supervised	45.5	57.5	63.9	65.6
SSL baseline	45.6	57.7	64.6	67.5
DSSL	**54.4**	**67.3**	**69.5**	**71.9**

Table 4. DSSL vs supervised and semi-supervised baselines. Average of 5 runs.

5. Conclusion

In this work, we proposed a straightforward technique to exploit invariances based on domain knowledge in the context of cameras and sensors installed on connected vehicles. More specifically, we leverage the fact that we can easily identify groups of unlabeled examples with the same (unknown) label. This helps in devising a semi-supervised method (DSSL) that obtains high-quality aggregated pseudolabels, and uses different examples in the same group (for example, frames from the same camera) as differently perturbed versions with the same label, reducing the need of aggressive artificial augmentations.

We showed how DSSL can work in a very straightforward way on three real-world tasks, including a time series classification task on sensor data. In all three cases, DSSL significantly outperforms both a purely supervised counterpart and a baseline SSL method, that appears to be significantly more brittle. In an ablation study, we show that DSSL can even work with no artificial data augmentation at all. This suggests that in practical settings, cleverly using all the available domain knowledge could be much more effective than relying on more sophisticated methods, especially in domains other than images, that might not have the same abundance of specialized techniques.

We believe that there might be several other cases that could benefit from our approach. As an example, another kind of domain knowledge that could be readily used in the context of distributed cameras is the GPS location, which is typically known for each example. In several tasks (e.g., see [7]), a cluster of frames that are geographically close would be associated with the same ground truth label, thus DSSL could be easily applied.

A potential future avenue for research might be the integration of this kind of domain-based invariants into a self-supervised learning method, where, rather than minimizing the discrepancy between artificially perturbed versions of the same example, one would minimize dissimilarity between different examples belonging to the same group. More in general, we believe that there is ample potential for exploration of simple and practical algorithms that cleverly exploit domain knowledge to reach and even surpass state-of-the-art performance of more general approaches.

References

[1] Kumar Ayush, Burak Uzkent, Chenlin Meng, Marshall Burke, David Lobell, and Stefano Ermon. Geography-Aware Self-Supervised Learning. *arXiv preprint arXiv:2011.09980*, 2020. 2

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2, 3

[3] Olivier Chapelle, B Scholkopf, and A Zien. *Semi-Supervised Learning*. MIT Press, 2006. 1

[4] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin D. Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *Computer Vision – ECCV 2020*, pages 695–714. Springer International Publishing, 2020. 1, 5

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 6

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes

dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[7] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, and Li Fei-Fei. Fine-grained car detection for visual census estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 7

[8] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Transactions on Medical Imaging*, 2021. 2

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[10] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017. 1

[11] Arthur Le Guennec, Simon Malinowski, and Romain Tavenard. Data augmentation for time series classification using convolutional neural networks. In *ECML/PKDD workshop on advanced analytics and learning on temporal data*, 2016. 6

[12] Quande Liu, Lequan Yu, Luyang Luo, Qi Dou, and Pheng Ann Heng. Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE Transactions on Medical Imaging*, 39(11):3429–3440, 2020. 2

[13] Davide Maltoni and Vincenzo Lomonaco. Semi-supervised tuning from temporal coherence. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2509–2514. IEEE, 2016. 2

[14] Hossein Mobahi, Ronan Collobert, and Jason Weston. Deep learning from temporal coherence in video. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 737–744, 2009. 2

[15] Hieu Pham, Qizhe Xie, Zihang Dai, and Quoc V Le. Meta pseudo labels. *arXiv preprint arXiv:2003.10580*, 2020. 1, 2

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[17] Matteo Simoncini, Leonardo Taccari, Francesco Sambo, Luca Bravi, Samuele Salti, and Alessandro Lori. Vehicle classification from low-frequency gps data with recurrent neural networks. *Transportation Research Part C: Emerging Technologies*, 91:176–191, 2018. 6

[18] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Neurips*, 2020. 2, 3

[19] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017. 5

[20] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 4, 5

[21] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 1

[22] Terry T Um, Franz MJ Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 216–220, 2017. 6

[23] US Department of Energy. Vehicle weight classes & categories, 2021. 6

[24] Peter Widhalm, Philippe Nitsche, and Norbert Brändie. Transport mode detection with realistic smartphone sensor data. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 573–576. IEEE, 2012. 6

[25] Wikipedia. Truck classification, 2021. 6

[26] Wei Xia, Junping Zhang, and Uwe Kruger. Semisupervised pedestrian counting with temporal and spatial consistencies. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):1705–1715, 2014. 2

[27] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc., 2020. 2, 3

[28] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves ImageNet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 1

[29] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 3

[30] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 2