Semi-supervised Learning in Camera Surveillance Image Classification

Matúš Tuna, Kristína Malinovská, Igor Farkaš
Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava
Bratislava, Slovak Republic
Email: {matus.tuna,kristina.malinovska,igor.farkas}@fmph.uniba.sk

Svatopluk Kraus, Pavel Krsek
Czech Institute of Informatics, Robotics, and Cybernetics
Czech Technical University
Prague, Czech Republic
Email: {svatopluk.kraus,pavel.krsek}@cvut.cz

***Abstract*—Recognizing pedestrian attributes in camera surveillance images is a very hard problem, due to the lack of high-quality labeled data. In the field of deep learning the semi-supervised learning paradigm provides a possible answer to this problem. We propose a novel semi-supervised model that we call Binary Mean Teacher, tailored for binary classification task of detecting the presence of wearable objects. We train our model in a traditional scenario with a randomly initialized model, but we also explore fine-tuning a model pretrained on a large-scale image dataset. The performance of our model is compared to strong supervised baselines trained or fine-tuned using our dataset and the same augmentation strategy as in our model. We evaluate the impact of various augmentation strategies commonly used in deep learning on the performance of models in our binary classification task. Using only 1000 labeled training images, randomly initialized Binary Mean Teacher model achieves roughly 90% classification accuracy compared to 75% accuracy of randomly initialized supervised model that does not use any augmentations. When both Binary Mean Teacher and the supervised model are pretrained using the ImageNet dataset, and augmentations are used for both models, the Binary Mean Teacher achieves 92% accuracy compared to 90% accuracy of the supervised model.**

***Index Terms*—Image classification, Camera surveillance, Deep learning, Transfer learning, Semi-supervised learning**

I. Introduction

There are many open challenges for machine learning in the domain of camera surveillance. Apart from identification of individual people from photos or videos, there are other vision tasks such as detection of wearable objects or outfits of the observed people in cases that the identity is unknown or cannot be distinguished.

The task of recognizing the attributes of people is difficult. Moreover, it is very difficult in the case of camera surveillance, where the quality of images is usually low. The cameras used to collect the data have low resolution and high variability in view angles. Since the task inherently spans across time, the lighting and weather conditions vary significantly.

In our work we focus on the pedestrian attribute recognition problem. To be more specific, we focus on recognizing just one important attribute in question, therefore we aim at training a binary classification system. An ensemble of well-trained classifiers for detecting crucial features from the images can be used for finding matches among multiple-camera images, which can be very useful in forensic search. For instance if a suspect of whom the identity is not precisely known, leaving no space for face verification, was carrying a special type of baggage or clothing, that can be used as a clue for finding the person in the huge amount of surveillance camera footage.

Although the video data are not hard to obtain via continuous recording, as well as automatic segmentation of images of people, the task of collecting a large enough dataset is still limited by making proper annotations. Given the poor quality of the data, automation possibilities in this respect are limited and the task is highly dependent on human operators. The problem of having an abundance of unlabeled data together with a small amount of well labeled data is currently well studied in the field of deep learning. The modern methods for the semi-supervised learning (SSL), e.g. learning from labeled and unlabeled data, can be used to leverage some problems of the camera surveillance image classification.

Here we present our explorations of deep learning techniques for binary classification of images. In order to evaluate our models we used a dataset proposed in our previous work. Our work unfolds in two directions, namely the use of data augmentation to reach better performance in the supervised learning paradigm as a continuation of our previous work, but also a new path in which we experiment with the semi-supervised learning models. We propose a novel semi-supervised model for binary classification of wearable objects that we call Binary Mean Teacher, based on the Mean Teacher model [1].

II. Related work

A. Semi-supervised learning

Generally, the SSL makes use of a large set of unlabeled examples from the same distribution as the real training set with proper labels. There are many SSL algorithms, but currently the most successful are the methods based on consistency regularization. These methods rely on applying various perturbations to the input of the model that do not change the label of the input and then force the output of the model to stay consistent across these inputs. Examples of this approach are Ladder networks [2], Pi model [3], Temporal Ensembling [3],

978-1-6654-0976-6/21/$31.00 ©2021 IEEE

Virtual Adversarial Training [4] and Mean Teacher model [1]. Another family of SSL approaches are proxy-label methods. These methods aim to expand the labeled portion of the dataset by labeling the unlabeled portion of the dataset using a neural network model trained using a smaller labeled portion of the dataset. These predictions can be generated using a single model or an ensemble of different models and data augmentation techniques. Examples of proxy-label methods are Pseudo-label [5], Deep clustering for unsupervised learning of visual features [6], Billion-scale semi-supervised learning for image classification [7], MixMatch [8] or Noisy student training [9].

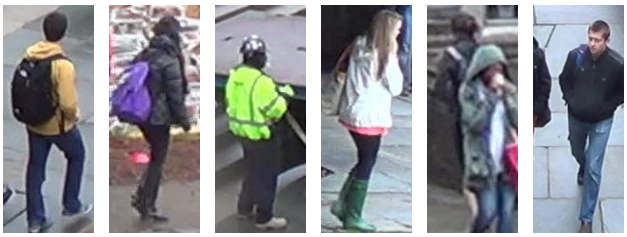
B. Classification of surveillance camera images

There are two main tasks related to the classification in the domain of camera surveillance image processing. First the re-identification task in which the identity of the observed persona has to be identified in multiple-view images from different cameras recorded over some time. The pioneers of transfer learning in the re-identification domain are Li et al. [10] who proposed the so-called metric transfer approach, which outperformed the state of the art based on features extraction. Further research showed that learning an additional task, along with the re-identification leads to better performance of the models. Examples of this approach is the addition of person classification [11], ranking task [12] or feature recognition task [13]. Wang et al. [14] combined the transfer learning and semi-supervised learning paradigms and made use of the attribute annotations on a dataset with sparse labels. Singh et al. [15] combined deep transfer learning with unsupervised learning via clustering of feature vectors using k-means clustering.

Another task is the so-called pedestrian attribute recognition task that aims to identify the physical attributes of the persona in the image, such as age (old/young), clothing or wearable items. Adding the attribute recognition subtask to re-identification has been shown to improve performance [13]. PETA [16] and RAP [17] are two of the recent benchmark datasets for pedestrian attribute recognition, which essentially is a multi-label classification task. Generally, the multi-label classification is learned as a sum of binary cross-entropy (BCE) losses from individual attributes. Yu et al. [18] showed improvement of performance on PETA and RAP datasets and demonstrated that weakly-supervised learning suffices to classify pedestrian attributes without the need of bounding box annotations. Ji et al. [19] enhanced the performance on PETA dataset using contextual information processed via LSTM modules in a hybrid CNN-LSTM neural architecture. Li et al. [20] proposed the deep hierarchical contexts model, a precisely engineered architecture which uses deep representations of humans in various poses together with the scene descriptions and is able to derive the attributes which are not even recognizable by a naked eye. Xiang et al. [21] tackled the problem of the lack of the labeled data via incremental few-shot learning. The first semi-supervised multi-label classification architecture was introduced by Cevikalp et al. [22].

III. Dataset and task

Our primary interest is in automated attribute recognition for camera surveillance systems. In this work, we build upon our previous work in the domain of pedestrian attribute recognition. Our dataset, called DukeMTMC-backpack dataset, consists of a reannotated version of the DukeMTMC-attribute [13] dataset that focuses on the backpack attribute. We reannotated the original dataset by hand to fix the label mismatches present in the original dataset, which led to substantially higher classification accuracy. These mismatches were caused by the fact that the original attributes had been assigned based on the entities observed throughout the whole video footage, rather than visual assessment of the attributes in the individual images. An illustration of our data is in Fig. 1.



Yes Yes No No N/A N/A

Fig. 1: Example of data classes in our dataset, from left to right: with backpack, without backpack and uncertain (N/A).

For experiments with semi-supervised learning, we took our DukeMTMC-backpack and used either the full labeled training set or kept only a portion (2%, 8%, 10%) of the labels and removed the labels from the remainder of the dataset. The case in which there is only data from DukeMTMC-backpack with portion of labels removed will be referred to as DS0. In order to explore the impact of the size of unlabeled portion of our dataset, we expanded the size of the unlabeled portion by using two additional datasets. First we expanded the unlabeled portion by including the data that we had previously removed due to uncertain labels in our previous work. We refer to this portion of the data as DS1.We decided to use a completely different, yet a similar dataset, namely both training and testing sets of the Market-1501 dataset [23]. The combination of DukeMTMC-backpack with Market-1501 will be referred to as DS2.

For completeness, we include the dataset statistics in Table. I. Note that the case of DS0 with 100% labels is actually not a real case of semi-supervised learning since there are no unlabeled data. However, it can work nicely as another supervised baseline that uses the same augmentation technique and distance-based regularization, yet with no contribution of the consistency regularization.

IV. Our approach

To overcome the problem of a small amount of labeled data in our task, we have explored the possibilities of transfer learning with the conclusion that the ImageNet-pretrained models with fine-tuned weights yielded the best performance so far, which was about 92% of accuracy on our test set.

TABLE I: Naming and statistics of the datasets for SSL.

Labels	DS0: DukeMTMC-backpack	DS1: DS0 + uncertain	DS2: DS0 + Market-1501
10133	100.00%	61.33%	28.00%
1000	10.00%	6.05%	2.76%
800	8.00%	4.84%	2.21%
200	2.00%	1.21%	0.55%

In the current work, we further explore the transfer learning and the influence of data augmentation on learning. Note, that our DukeMTMC-backpack dataset is not typical in terms of train-to-test set size ratio, as we kept the split by the particular entities as in the original DukeMTMC-attribute. With uncertain samples left out, we ended up having a slightly larger testing set with a similar, yet not precisely the same data distribution as the training set.

In the domain of camera surveillance it is crucial to overcome the problem of the lack of labels for the abundance of collected data. The aim of the novel semi-supervised learning paradigm is to leverage a high number of unlabeled examples which are more easily obtainable than labelled examples. In our work, we explore the SSL approach, namely the well established Mean Teacher (MT) model by Tarvainen and Valopola [1] and adapt it for our dataset and task. The novel Binary Mean Teacher (BMT) model and its components are presented in this section.

A. Transfer learning

Transfer learning [24] works by adapting the source predictive function (neural classifier) to the target domain [25] assuming that the target and source domains share some common low-level structure, for instance natural world images. The low-level layers of a convolutional architecture are assumed to act as low-level feature extractors, sensitive to basic lines and shapes. A convolutional classifier trained on a large and variable dataset such as ImageNet [26] is altered for a different target function using the target dataset, which is usually much smaller. This practice has become very common in image processing tasks.

The fine-tuning of the model for the new task can be done locally by separately training the topmost fully connected layer on a new task using the existing feature extractor. However, as we also confirmed in our experiments with our dataset, the model fine-tuning, i.e. training the whole architecture with an appropriately small learning rate yields much better results. In our current work, we compare the performance of the pretrained and unpretrained model using the ImageNet dataset.

B. Semi-supervised learning

The goal of SSL can be broadly characterized as an usage of a large number of unlabeled examples from the same distribution as a small number of labeled examples for the purpose of informing the learning algorithm about the distribution of the overall dataset. The infusion of unlabeled data together with a mechanism for learning from them yields higher prediction accuracy than the identical algorithm trained only using a small number of training examples.

The Mean Teacher model [1] leverages a combination of various semi-supervised learning techniques to dramatically improve generalization accuracy in the semi-supervised learning context. Inspired by the Ladder Networks [2], MT predicts the class of the unsupervised data points using two sets of noise augmentations applied to unsupervised data points.

Noisy batches are evaluated using two neural networks called the student θ and the teacher θ'. Predictions for every data point in each of the noisy batches should be consistent between the student and the teacher. This is achieved using a consistency cost J between the student and teacher predictions in the form of the Mean Squared Error (MSE):

$$J(\theta) = \frac{1}{n}\sum_{i}^{n} \|f(x_i, \theta', \eta') - f(x_i, \theta, \eta)\|^2 \tag{1}$$

of all samples x_i corrupted by random noise η and η'.

MT uses MSE as the consistency cost, although different cost functions, such as Cross Entropy (CE) or Kullback-Leibler (KL) divergence are also usable. In addition to consistency cost, supervised cost in the form of CE is also evaluated for the labeled portion of the input batch:

$$S(\theta) = \frac{1}{m}\sum_{j}^{m} [-\log\ P_f(y_j|x_j; \theta, \eta)] \tag{2}$$

of all labeled samples x_j corrupted by random noise η and their respective targets y_j.

Consistency cost in the MT model effectively acts as a regularization technique that enforces the outputs of the model to be consistent across similar data points, which leads to improvement in generalization accuracy of the model. Since the consistency cost uses self-generated targets and because in the semi-supervised setting there are usually many more unlabeled examples than labeled examples, during the training the consistency cost might come to dominate the training process. The Mean Teacher model mitigates this problem by introducing a dynamic consistency cost weight w_t that balances the contribution of consistency cost and supervised cost. The overall composite cost function is then

$$Loss(\theta) = S(\theta) + w_t J(\theta) \tag{3}$$

MT also incorporates model parameters ensembling technique similar to prediction ensembling technique from Pi model [3]. In order to improve targets produced by the teacher network, the parameters of the teacher network θ' are

computed as an exponential moving average of the student model θ

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t, \tag{4}$$

where α represents an additional hyperparameter that controls the smoothing of the averaging.

C. Data augmentations

Common image augmentation techniques were found to increase generalization accuracy of both fully supervised models [27] as well as unsupervised representation learning models [28] and semi-supervised learning models [1]. These may include augmentations such as rotation, translation, adding random noise or color jitter, random cropping and aspect ratio changes. More advanced techniques include augmentations using generative adversarial networks [29] or randomly selected and algorithmically optimized strategies [30]. We have drawn inspiration from the MT model [1] when selecting data augmentation techniques. We applied random rotation, random crop and resize, random horizontal flip and random color jitter to all training data in our experiments with Binary Mean Teacher, as well as in the supervised baselines. We provide the parameters for each augmentation used in Table II. We explored the importance of random augmentations and also the ways they influence the performance separately via supervised learning with a limited number of labels. The experimental setups and results are presented in the following section.

Fig. 2: Examples of used augmentations. Top left corner shows the original image.

D. Network architecture

In our previous work, the best results were achieved with the DenseNet model [31]. Therefore we used it consistently throughout all our experiments. Similarly to the ResNet model [32] used by the original MT model, the DenseNet model also adds a novel type of skip connections between layers. DenseNet, as well as other skip-connection models, alleviates the vanishing gradient problem and enables the individual convolutional layers to exploit information from different parts of the model. Even though the best results achieved in previous work were obtained with DenseNet161, we chose the DenseNet121 variation due to a smaller demand on computational resources and only a very small decrease in the model accuracy.

E. Binary Mean Teacher

To change the MT model for our purposes, we needed to change the shape of the last fully connected layer. Besides that, since our task is binary, we only have one output neuron and our primary (supervised) cost function is the Binary Cross Entropy (BCE) so our supervised cost is

$$S(\theta) = -\sum_{j}^{m} \left[y_j \log \hat{y_j} + (1-y_j) \log(1-\hat{y_j}) \right], \tag{5}$$

where y_j stands for the desired value (yes/no) and $\hat{y_j}$ represents the network output $f(x_j, \theta, \eta)$ for labeled input x_j given noise η.

In our preliminary experiments, we realized that the MSE cost used for the unsupervised part of the model given just one output neuron did not lead to successful learning. Therefore we altered the model to compute the unsupervised consistency cost directly as MSE from the last convolutional layers of the student and the teacher networks as displayed in Fig. 3. Our unsupervised cost is then computed as

$$J(\tau) = \frac{1}{n} \sum_{i}^{n} \left\| g(x_i, \tau', \eta') - g(x_i, \tau, \eta) \right\|^2, \tag{6}$$

where, as in Eq. 1, x_i are all training samples corrupted by noise η for the student network and η' for the teacher network, $\tau \subset \theta$ and $\tau' \subset \theta'$ are all weights of the student and teacher up to the last convolutional layer and the whole network $f = h(g(x))$ where g is the convolutional part and h is the fully connected layer atop of the architecture. We combined the two losses $S(\theta)$ and $J(\tau)$ in the same manner as the original MT model according to Eq. 3.

V. Experiments

We evaluate our models in several different setups: simple supervised learning with no pretraining and also model fine-tuning with the model pretrained using ImageNet. First, we describe the used hyperparameters. Second, we evaluate the influence of data augmentations in the supervised setup with fully labeled set, but also with a very limited small subset of the training data. Finally, we explore and evaluate our Binary Mean Teacher model that we adapted for our task and the dataset and compare the performance of BMT in the limited data case, i.e. just one thousand labels. We evaluate this semi-supervised model with varying quantity and quality of the unlabeled data as described below.

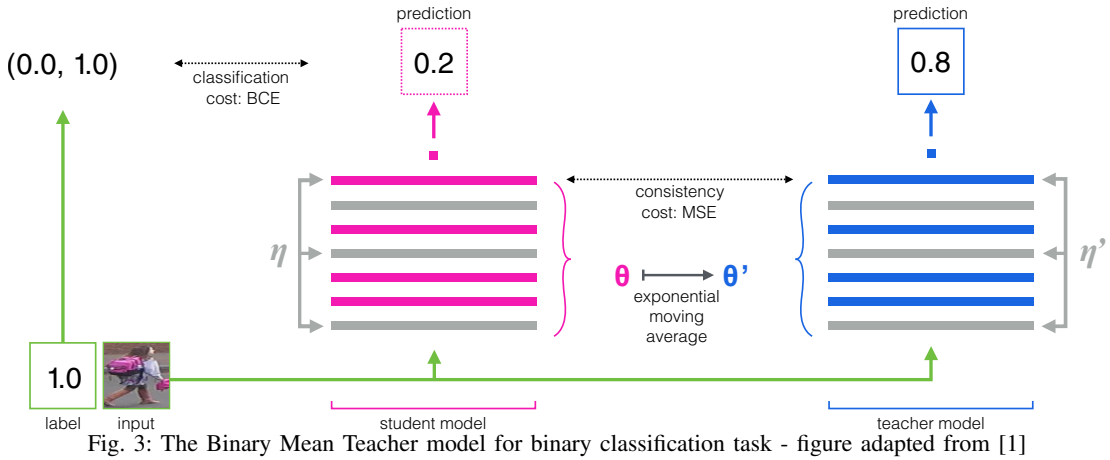


Fig. 3: The Binary Mean Teacher model for binary classification task - figure adapted from [1]

A. Model selection and hyperparameter tuning

In our experiments we used DenseNet121 adapted for our binary classification task. The output layer contains one output neuron predicting whether the person in the image is with a backpack as described in Sec. IV.

In our experiments, where there was no pretraining we used the Xavier weight initialization [33]. As described in Sec. IV-E, we used the composite loss function that contains supervised loss (Eq. 5) and unsupervised loss (Eq. 6) weighted by parameter w_t. We trained our networks using well established ADAM optimizer algorithm [34] with standard Adam hyper parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and no weight decay. Due to a high computational demand, we used exactly 100 training epochs in all experiments including the hyperparameter search. As the performance measure we use standard classification accuracy.

B. Supervised learning and augmentations

Data augmentation undoubtedly enhances the performance of the image classification models and is one of the core components of the BMT model. We explored the influence of the augmentations in the case of sole supervised learning with a limited dataset and with a full dataset. Our assumption behind this experiment was that even with a very limited amount of data, using a reasonable augmentation technique, we can achieve a performance comparable to the full dataset training. We explored the full model training as well as model fine-tuning with the model pretrained using ImageNet.

For our experiments, we used the same set of augmentations as the original MT [1] for the ImageNet (discussed in Sec. IV-C) and evaluated them separately and in combination. The only difference in our augmentation strategies compared to the original MT is in the crop and resize, which were slightly adapted to keep the backpack and the torso of the person in the image intact. For the sake of clarity we name the augmentations using arbitrary codes as shown in Table II.

In a search for optimal hyperparameters, we experimented with a varying learning rate and minibatch size for various sizes of the training data sets in both setups: full training and model fine-tuning. We tested batch sizes of 4–64 samples and

TABLE II: Augmentation types and parameters.

	Technique	**Parameters**
A	Random rotation	max. 10 degree
B	Varying aspect ratio crop & resize	224×224, scale=(0.8, 1), ratio=(0.8, 1.2)
C	Random horizontal flip	
D	Random color jitter	brightness=0.4, contrast=0.4, saturation=0.4, hue=0.1

initial learning rates from 10^{-2} to 10^{-5}. While for the learning rate we identified the best value over all tested combinations and learning setups, which was 10^{-4}, ideal minibatch size varied from 4 to 16. However, the differences in performance were rather small so we decided to settle with minibatch size 8 as the best value. Note that such a small minibatch size could not be feasible for the Binary Mean Teacher, where a certain ratio of unlabeled and labeled samples has to be kept in each minibatch. Unlike our smallish optimal minibatch size, the SSL uses rather large minibatches of approximately 128 or more samples. We assume this is necessary due to the nature of the dataset which can in some cases contain only several labeled samples per minibatch which would not be feasible to make a data split with such a small size.

The influence of the augmentation strategy on the accuracy of our models is shown in Fig. 4. The best performing models were the ones that used model fine-tuning and a combination of three and more augmentations. It seems that the cropping and resizing (B) as well as the color jitter (D) have the best overall influence on the task performance.

In the context of our domain, the well-labeled images are not in abundance. Therefore, we were interested to see how well a standard supervised model can generalize over the whole testing set when trained with only a small amount of labeled data. We chose small portions in accordance with the SSL experiments which go up to one third of the dataset and compared it with the full dataset. Results are displayed in Fig. 5. Note that the testing set we used was the same as in any other task and contained more than 11000 images. The smallest portion of labels, which was 2% of our training set equals to approximately 200 images. This graph illustrates

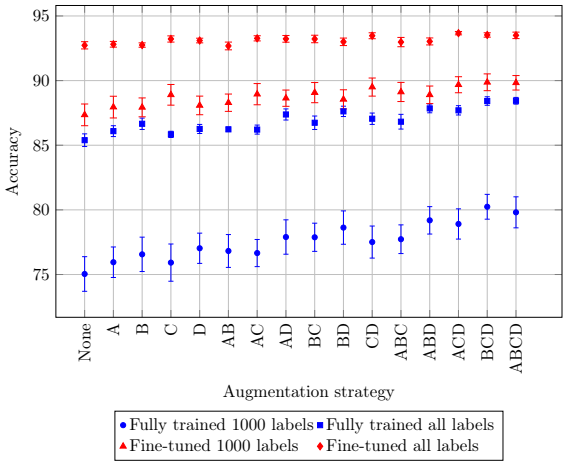


Fig. 4: The influence of the augmentation strategy on the accuracy. The legend that maps the augmentation strategies with their labels is displayed in Table II

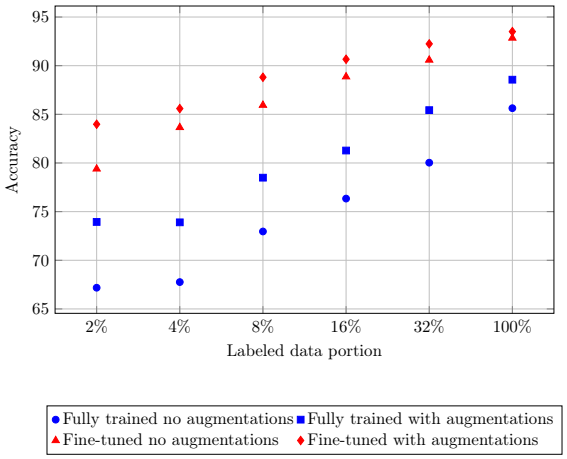


Fig. 5: The influence of the size of training dataset as a portion of original training dataset on the model accuracy.

how augmentations can compensate for the lack of data even when no unlabeled data are used. The effect of augmentations is much more visible in the case of full training which, however performs generally worse than the fine-tuned setup. If pretrained models are used, the weights already respond well to low-level features so the augmentations that generally increase variability of the training data have smaller influence on the model performance.

C. Semi-supervised learning with Binary Mean Teacher

We used our BMT model as described Sec. IV-E and evaluated both learning modes as in the supervised scenario. We used model fine-tuning with ImageNet, but also trained the models from the scratch. Adding more complexity, we tested the model on three variations of the labeled and unlabeled dataset listed in Table I, which start from basic DukeMTMC-backpack (DS0) and then add on different unlabeled data. We also took varying amounts of labeled data from our trainset (e.g. 1000 or the whole 10000) while the remainder was added to the unlabeled part of the training set. As described in Table I, DS1 combines DukeMTMC-backpack with uncertain category which contains images that cannot be conclusively categorized due to occlusions and the presence of multiple persons. Then in DS2 we combine two different, yet similar datasets which both contain images of people from surveillance cameras, DukeMTMC and Market-1501. Although there are strong similarities, the distribution of backpack vs. no-backpack cases in Market-1501 is different from that in our dataset. Nevertheless, we presume that a higher number of unlabelled data will improve the accuracy of our model despite the different dataset distributions. This can be very useful in practical applications where there is an abundance of data, but the labels for the data are hard to achieve. Therefore a reasonable small set of good examples along with well tailored augmentation strategies can help balance this problem.

We have found the Binary Mean Teacher model to be very sensitive to hyperparameter selection. Therefore we paid particular attention to hyperparameter tuning process. Because of the large number of hyperparameters that can influence the final accuracy, we focused on exhaustive parameter search only for a small subset of hyperparameters that most influenced the final accuracy. These include the learning rate, unsupervised weight, EMA decay rate and ramp-up length. We performed hyperparameter search on these parameters with range [0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00005, 0.00001] for the learning rate, [20, 30, 40, 50, 60, 70, 80] for unsupervised weight, [0.9, 0.95, 0.99, 0.995, 0.999, 0.9995, 0.9999] for EMA decay rate and [1,5,10,15,20] for ramp-up length. We performed the hyperparameter search using 1000-label subset of DS2 and then reused these hyperparameters in subsequent BMT experiments.

We performed the hyperparameter tuning process separately for fine-tuning and full training setup. Other hyperparameters were fine-tuned using various heuristics with original parameters used in the original MT research [1] as a starting point. We used minibatch of size 64 due to the memory limitations of our hardware. In each minibatch we used 48 unlabeled and 16 labeled examples (8 examples with backpack and 8 without backpack). We have found that using a minibatch composition with a fixed number of labeled and unlabeled examples led to higher classification accuracy than sampling labeled and unlabeled examples randomly due to high imbalance between labeled and unlabeled examples. We also explored different minibatch compositions but found previously mentioned composition to work best across multiple experiments. We have fixed the ramp-down length at the value 20 in the initial exploratory experiments and then left it fixed because this parameter did not significantly alter the performance. We found that the optimal ramp up length was 10 for our setup. We used all the above mentioned data augmentations at once, which is the same as in original MT setup for the ImageNet dataset. In Table III, we summarize optimal hyperparameters that differ between setups.

TABLE III: Optimal BMT hyperparameters found for different setups.

Setup	Learn. rate	w_0	EMA α
Full training, all labels	0.001	50.0	0.995
Full training, 1000 labels	0.0005	70.0	0.999
Fine-tuning, all labels	0.0001	70.0	0.9999
Fine-tuning, 1000 labels	0.001	60.0	0.99

D. Overall results

To sum up the results of our experiments with supervised and semi-supervised learning with and without the use of model fine-tuning, we evaluated our best models on the above-mentioned datasets. The fine-tuned models start with the DenseNet architecture pretrained using the ImageNet dataset, which is a common practice in deep neural computer vision. In our case, we can confirm that the transfer learning paradigm is useful also for binary classification of images.

For the supervised learning, we just use the DS0 with varying portion of the training dataset. Our Binary Mean Teacher results are shown for DS0-DS2 also with a varying ratio of the labeled part of the training set. For evaluation, we always use the whole DukeMTMC-backpack testing set (roughly 12000 images). The results are shown in Tables IV and V.

TABLE IV: Overall results: full training

Learning	Dataset	Augment	All	10%	8%	2%
Supervised	DS0	None	85.39	74.96	72.35	63.10
Supervised	DS0	ABCD	88.44	78.47	77.95	69.38
BMT	DS0	ABCD	83.28	89.14	81.77	75.33
BMT	DS1	ABCD	83.73	88.53	82.45	75.56
BMT	DS2	ABCD	85.09	89.57	83.17	75.67

TABLE V: Overall results: model fine-tuning

Learning	Dataset	Augment	All	10%	8%	2%
Supervised	DS0	None	92.73	86.96	86.19	78.29
Supervised	DS0	ABCD	93.50	89.91	88.89	82.07
BMT	DS0	ABCD	92.85	90.72	90.04	79.21
BMT	DS1	ABCD	92.98	90.47	89.75	78.96
BMT	DS2	ABCD	93.53	91.75	90.83	77.70

VI. Discussion

In this work we propose a semi-supervised model for binary classification task of detecting the presence of wearable objects that we call the Binary Mean Teacher. Our model is based on our finding that computing the consistency cost using the final output of the network, as in the original Mean teacher model, yields suboptimal results in binary classification setting. In our model, we instead opted to compute consistency cost using the last convolutional layer of both the student and teacher models. Computing and combing the consistency cost on different layers of the network might further improve the classification accuracy but we leave this to future work.

We compared the classification accuracy of this model to strong transfer learning baseline based on model pretrained on the ImageNet dataset. We also tested the impact of various input augmentation methods on the final classification accuracy when applied to both the transfer learning baseline and Binary Mean Teacher model. Skipping the cropping augmentation (Resized crop) leads to slightly higher classification accuracy, due to the sometimes small distance of the object of interest (backpack) to the border of the image. This can lead to removal of the backpack object during the augmentation process rendering the example invalid.

The main advantage of the Binary Mean Teacher compared to the baseline manifests itself when we train both models from randomly initialized weights. In this setting we observe 5–10% difference in the classification accuracy in favor of the Binary Mean Teacher model. In the tests using 1000 labeled training images, randomly initialized Binary Mean Teacher model achieved 90% classification accuracy. Randomly initialized baseline without any image augmentations achieved 75% accuracy using the same number of training images. This demonstrates the superiority of our approach compared to the baseline in the setting with randomly initialized weights.

The difference is much smaller when we initialize the weights of both supervised baselines and Binary Mean Teacher using the weights pretrained on the ImageNet dataset. In this setting Binary Mean Teacher achieved 92% accuracy and the baseline achieved 90% accuracy. This suggests that the supervised model pretrained on a suitable large dataset should be used as a baseline when testing semi-supervised models. In this setting the Binary Mean Teacher model has only a slight accuracy advantage compared to supervised baseline, which demonstrates the effect of well-established fine-tuning paradigm. This finding is contrary to most semi-supervised research results where the gap between supervised models and semi-supervised models is much higher because of absence of baselines pretrained on large-scale image datasets such as ImageNet. We suspect that much larger unlabeled datasets would be needed to fully leverage the potential of semi-supervised models like our Binary Mean Teacher model in the binary classification task studied in this work.

In the 2% labeled examples setting the Binary Mean Teacher model achieved a slightly lower accuracy than the baseline. This can be attributed to the hyperparameter tuning procedure that used the 1000 label (8%) setting. This suggests that setups using a different number of labeled examples require different hyperparameter settings to achieve optimal results. We did not perform exhaustive hyperparameter tuning for every setting with a different number of labeled examples due to computational constraints.

VII. Conclusion

We presented our work on binary classification of pedestrian attributes in camera surveillance images using deep neural networks. This domain is characteristic with a lack of well labeled

data, which can be counterbalanced by special techniques that make use of data augmentations as well as unlabeled data. We proposed a semi-supervised model called Binary Mean Teacher for the task of detecting the presence of wearable objects in surveillance data. We compared the performance of our model to strong supervised baselines and found that although semi-supervised learning did lead to high performance gains in low training data regimes, these gains were limited when both semi-supervised model and supervised model were pretrained on large image dataset and when image augmentations are applied in both settings. We explored the influence of data augmentations, well-tailored random transformations of input data that increase generalization abilities of the networks. This work reveals that the lack of training data can be compensated for by both data augmentations in case of simple supervised learning as well as by using the semi-supervised learning paradigm.

ACKNOWLEDGMENT

This research was supported by the Ministry of the Interior of the Czech Republic Project VI20172019082, Smart Camera, EU project Inafym CZ.02.1.01/0.0/0.0/16.019/ and by KEGA project no. 042UK-4/2019.

REFERENCES

[1] 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*, 2017, pp. 1195–1204.

[2] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 3546–3554.

[3] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[4] 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*, vol. 41, no. 8, pp. 1979–1993, 2018.

[5] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896.

[6] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 132–149.

[7] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, "Billion-scale semi-supervised learning for image classification," *arXiv preprint arXiv:1905.00546*, 2019.

[8] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *arXiv preprint arXiv:1905.02249*, 2019.

[9] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 687–10 698.

[10] W. Li, R. Zhao, and X. Wang, "Human re-identification with transferred metric learning," in *Asian conference on computer vision*. Springer, 2012, pp. 31–44.

[11] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," *arXiv preprint arXiv:1611.05244*, 2016.

[12] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[13] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.

[14] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2275–2284.

[15] M. K. Singh, V. Laxmi, and N. Nain, "Unsupervised person re-id in surveillance feed using re-ranking," in *International Conference on Security & Privacy*. Springer, 2019, pp. 179–192.

[16] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Learning to recognize pedestrian attribute," *arXiv preprint arXiv:1501.00901*, 2015.

[17] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A richly annotated dataset for pedestrian attribute recognition," *arXiv preprint arXiv:1603.07054*, 2016.

[18] K. Yu, B. Leng, Z. Zhang, D. Li, and K. Huang, "Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization," *arXiv preprint arXiv:1611.05603*, 2016.

[19] Z. Ji, W. Zheng, and Y. Pang, "Deep pedestrian attribute recognition based on lstm," in *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 151–155.

[20] Y. Li, C. Huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *European Conference on Computer Vision*. Springer, 2016, pp. 684–700.

[21] L. Xiang, X. Jin, G. Ding, J. Han, and L. Li, "Incremental few-shot learning for pedestrian attribute recognition," *arXiv preprint arXiv:1906.00330*, 2019.

[22] H. Cevikalp, B. Benligiray, and O. N. Gerek, "Semi-supervised robust deep neural networks for multi-label image classification," *Pattern Recognition*, vol. 100, p. 107164, 2020.

[23] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[24] L. Y. Pratt, "Discriminability-based transfer between neural networks," in *Advances in Neural Information Processing Systems*, 1993, pp. 204–211.

[25] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, p. 9, 2016.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[28] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. Eslami, and A. v. d. Oord, "Data-efficient image recognition with contrastive predictive coding," *arXiv preprint arXiv:1905.09272*, 2019.

[29] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2017.

[30] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," *arXiv preprint arXiv:1909.13719*, 2019.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[33] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.