

# Unified Adversarial Training for Bias Mitigation and Privacy Preservation

Rémy Vuagniaux<sup>[0009-0006-8587-4963]</sup>, Mohamad Dia<sup>[0000-0003-1028-787X]</sup>,  
Sareh Saeedi<sup>[0000-0002-6492-5657]</sup>, Simon Narduzzi<sup>[0000-0003-1762-199X]</sup>, Engin  
Türetken<sup>[0009-0006-6967-9674]</sup>, and Nadim Maamari<sup>[0000-0002-8910-4759]</sup>

CSEM, Rue Jaquet-Droz 1, 2002 Neuchâtel, Switzerland  
remy.vuagniaux@csem.ch  
<https://www.csem.ch/en/>

**Abstract.** Human-Centered Machine Learning (HCML) models often face challenges due to inherent biases related to population variability and limited access to large datasets. This results in algorithms that fail to generalize and accommodate out-of-distribution samples, thereby hindering real-world applications. Additionally, standard training procedures tend to make neural networks vulnerable to privacy risks such as reconstruction attacks. To address these issues, we propose a novel training method based on an adversarial network that aims to reduce the representation bias induced by the lack of diversity among training samples. Unlike similar approaches that use a known bias predictor as the adversarial signal, our method mitigates multiple unknown biases, acting as an effective regularization term that reduces the validation gap while also removing non-essential features. This feature selection further improves privacy by preventing the model from being repurposed or used to retrieve information about training or inferred samples, as demonstrated on the IMDb-Face dataset where the method achieves approximately a 6.7% improvement in accuracy and enhances robustness against reconstruction attacks by about 174%.

**Keywords:** HCML · Privacy · Ethic · Deep Learning · Bias Mitigation

## 1 Introduction

Human-Centered Machine Learning (HCML) has emerged as a major field in both industry and research in recent years [3], offering significant benefits for human well-being.

In many cases, the solutions used in this field rely mainly on Deep Learning (DL). However, these algorithms often suffer from biases inherent in the data.

This problem arises from the fact that data collection processes involving human subjects can be complex and challenging. In addition to ethical concerns and increasing regulations on sensitive data collection, gathering a diverse dataset for vision tasks is tedious due to the limited availability of samples from different subjects (i.e., the sampled population is too small to provide sufficient

diversity). This lack of data can make data stratification difficult and induce **sampling bias**, as well as generate **pseudo replication bias** when different training samples are taken from the same subject.

Another crucial aspect of HCML is the sensitivity associated to the data. For example, when handling medical data, it is important to make sure that the data is anonymized. Given a standard training paradigm, even if the data do not leave the infrastructure on which they are trained on, remembrance of it can be found in the model weights itself.

For example, it has been proven that it is possible to even reconstruct the subject appearance from model feature maps, as shown by [6,10]. The attackers might also be able to obtain sensitive information about the subject data being inferred. This problem has already been addressed by several works [4, 8, 10]. However, the majority of these methods rely on explicitly preidentified bias attributes (gender, age,...). Additionally, in most cases, these approaches are limited to classification tasks only.

We propose a novel adversarial training method that extends beyond classification tasks, supporting a wide range of applications while simultaneously mitigating multiple biases—even when these biases are unlabeled. To demonstrate the potential of our approach within an HCML context, we apply our method to the IMDb-Face dataset, showing that filtering out irrelevant features enhances the model’s generalization. Moreover, by suppressing sensitive information, our method also improves model privacy. Finally, we showcase the versatility of our technique by applying it to our own remote photoplethysmography (rPPG) dataset, particularly in scenarios with a limited number of subjects.

## 2 Related Work

In this section, we survey related works relevant to our study, focusing on two complementary areas: bias mitigation in machine learning and privacy-preserving techniques. We first examine approaches that reduce inherent biases in data and models, and then review methods that protect sensitive information during training.

### 2.1 Bias Mitigation

The diverse biases inherent in data, as outlined by [7], underscore the importance of developing robust mitigation strategies. In DL, biases may be introduced at various stages, especially during data collection and selection. A wide range of methods have been proposed for bias mitigation during training [4, 7, 8, 10, 14], with specific applications in HCML [9, 13]. According to this paper, debiasing methods can be classified into three main categories:

1. **Data Stratification:** The simplest method for bias mitigation is to stratify data based on known biases. While this approach can be effective for straightforward Machine Learning (ML) models, it becomes impractical in

cases where data collection is challenging since the diversity of data is already limited. Additionally, extensive bias factors (gender, ethnicity, age, weight) complicate stratification due to the number of possible combinations between those categories.

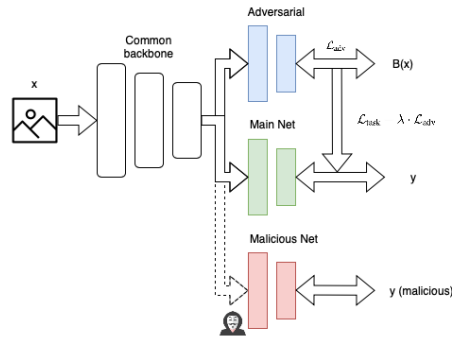
2. **Known Bias Debiasing:** This method aims to train models to disregard biased features by using techniques that force the model to "unlearn" the bias. Commonly, adversarial networks [4, 10, 13, 14] are employed to detect and mitigate biases within the model. Acting as a regularization term, the adversarial network's loss is set to increase while the model's loss decreases. However, explicit bias mitigation requires prior knowledge of the different biases, which in some cases is not trivial. For example, the authors in [4] used fur color as an explicit bias in animal classification (Cat vs. Dog) tasks, but other biases (such as fur length) could also be relevant in this case. This shows that the model can still be biased.
3. **Unknown Bias Debiasing:** This approach addresses both known and unknown biases without explicit specification of them. For instance, authors in [8] used a combination of Categorical Cross-Entropy (CCE) and Generalised Cross Entropy (GCE) to mitigate biases by encouraging the adversarial model to identify biased features while forcing the main network to unlearn them. This effectively addresses both known and unknown biases. However, this approach is limited to classification tasks, as the regularization term applied to the main model's loss relies on GCE and CCE.

## 2.2 Privacy-Preserving Machine Learning

Machine learning models are increasingly susceptible to privacy attacks that can extract sensitive information from training data [2, 6, 10, 11]. These include **adversarial attacks**, which seek to manipulate model predictions by injecting noise into inputs, and **model inversion attacks**, which aim to reconstruct input data or even retrieve samples used during training.

Tanuwidjaja et al. [11] illustrated that malicious networks can reconstruct input data provided to a trained model. Additionally, as DL become more computationally expensive, many developers and researchers are shifting model training and inference to cloud-based environments, using Machine Learning as a Service (MLaaS). This setup is sensitive to classification attack or reconstruction attack which operates as follows: the attacker gains access to the backbone model. Using known samples, they train a malicious network to classify a target label of interest or even reconstruct the input sample via inversion techniques. Once this malicious network is trained, the attacker places the model on the network and monitors the edge device which contains the backbone, being, for instance, a camera, an IoT device, etc. With this setup, the attacker can extract information about the current sample being processed.

In [10], the authors demonstrate how to counter this type of attack using an adversarial training method similar to bias mitigation, combined with adding noise to the features extracted by the edge model backbone. While they show that this approach can prevent a sensitive feature from being extracted by the



**Fig. 1.** Training process presented in this work. The adversarial and malicious network’s weights updates do not affect the backbone weights during their back-propagation. The main loss combines task loss and adversarial loss, encouraging the main model to unlearn biased features. The malicious network has no impact on the main model or backbone.

edge backbone, they do not address cases where an unknown feature could still be exploited by an attacker.

This work builds on previous research to address both bias mitigation and privacy preservation across different tasks. Our goal is to improve model performance while protecting privacy, advancing methods that enhance fairness and security in real-world applications.

### 3 Method

#### 3.1 Adversarial Setup

We adopt the Adversarial Network framework proposed in [4, 10], where the main model and the adversary share a common backbone (feature extractor). To simulate a potentially malicious cloud environment, as described by [10], we add an additional network which uses the backbone as well, referred to as the malicious network. This malicious network will track the model’s ability to hide private features. The complete setup is illustrated in Figure 1.

The objective of the adversarial network is to predict the bias label associated with each sample. By incorporating the adversarial loss as a regularization term in the main model’s loss, we want to encourage the main model to unlearn the biased features identified by the adversary.

The losses for the two models are defined as follows:

$$\mathcal{L}_{\text{main}} = \mathcal{L}_{\text{task}}(f(\mathbf{x}; \theta), \mathbf{y}) - \lambda \cdot \mathcal{L}_{\text{adv}}(f_{\text{adv}}(\mathbf{x}; \theta, \phi), B(\mathbf{x}))$$

in which  $\mathbf{x}$  is a batch of training samples and  $\mathbf{y}$  is the batch of corresponding training labels. Also,  $f(\mathbf{x}; \theta)$  denotes the main network, with  $\theta$  as the shared

backbone parameters, and  $\mathcal{L}_{\text{task}}$  indicates the task loss, which can vary based on the problem.

In Eq. 1,  $\mathcal{L}_{\text{adv}}$  is defines as follows:

$$\mathcal{L}_{\text{adv}}(f_{\text{adv}}(\mathbf{x}; \theta, \phi), B(\mathbf{x})) = - \sum_j B_j(\mathbf{x}) \log f_{\text{adv},j}(\mathbf{x}) \quad (1)$$

where  $B(\mathbf{x})$  is the associated subject ID and  $f_{\text{adv}}(\mathbf{x}; \theta, \phi)$  denotes the adversarial network,  $\phi$  is a set of parameters specific to the adversarial head but shares the backbone parameter  $\theta$  with the main network. The subject ID represents the individual identity associated with each sample (e.g., in an experiment with three subjects, 300 samples are collected, the adversarial network task is to classify the subject ID of each sample). The adversarial loss function follows a structure similar to CCE. For  $\mathcal{L}_{\text{adv}}$ , we also incorporate GCE, as it has been shown to better capture biased features, as demonstrated by [8].

For the malicious network we simply use the following loss function, which is in this case a simple CCE

$$\mathcal{L}_{\text{mal}}(f_{\text{mal}}(\mathbf{x}; \theta, \phi), y_{\text{mal}}) = - \sum_j y_{\text{mal},j} \log f_{\text{mal},j}(\mathbf{x}) \quad (2)$$

where  $y_{\text{mal}}$  is the corresponding malicious label, and  $f_{\text{mal}}(\mathbf{x}; \theta, \phi)$  is the malicious network.

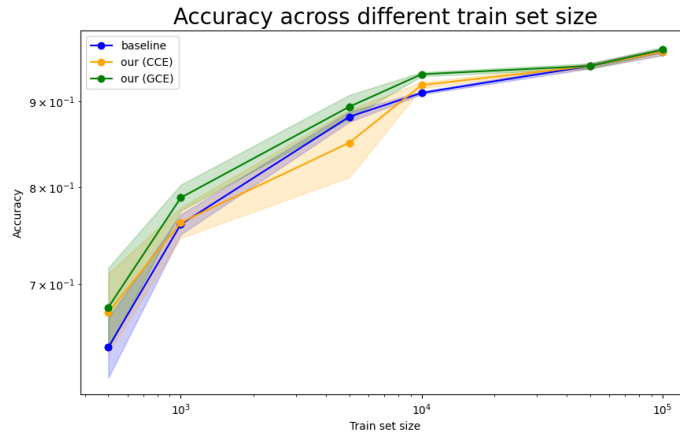
Our approach introduces a novel adversarial network that helps the main model to unlearn features which contain unnecessary information about the sample. By forcing the model to unlearn identity-related features, we force it to not learn non-essential attributes (i.e., removing features that allow the adversarial network to identify the subject id, such as skin color, hair, or facial shape, will also eliminates unintended biases or private feature).

Unlike prior methods such as [8], our approach selectively unlearns biased features while maintaining greater flexibility, as the main network is not restricted to classification tasks. Only the adversarial network needs to perform a classification. This balance allows our method to effectively mitigate bias, enhance privacy, and preserve both task adaptability and predictive accuracy.

## 4 Experiment

### 4.1 IMDB-face Dataset

**Experiment setup** To align with HCML applications, we start our experiments with the IMDB-Face dataset [5]. This dataset contains 285946 images of individuals, with labels indicating the age, identity and the gender of the person. Similar to [4], the main task is to classify the gender of the person in the image. However, instead of using age as a bias to mitigate, we use the individual’s identity (denoted  $ID(x)$ ). if the adversarial network can correctly identify a person based on the embedding provided by the backbone, this indicates that the model



**Fig. 2.** Figure showing the results on the IMDB-Clean dataset with different training set sizes. The standard deviation for each model is computed across multiple seeds.

has failed to preserve privacy and has learned unnecessary features. To track the presence of unnecessary features, we introduce a malicious network tasked with predicting the age of the individual from the sample. Again if our method work as intended, the malicious network should have more difficulty predicting the age compare to the baseline model.

Three models are compared: a baseline model, another model trained with adversarial using CCE, and one using the GCE method. Unlike previous works that split the dataset based on attributes, we instead subsample the dataset based on the ID of a person. In particular, we define train and validations sets of sizes between 500 and 100000 samples. The validation set contains the same amount of data as the training set. Also the same individual (same individual *ids*) are present in the validation set that the one present in the train set. This introduces pseudo-replication in the validation set, but it is necessary to monitor the adversary, as it cannot identify items it has not previously seen and therefore cannot predict individual items that were never encountered. We keep the test set as default across all the train set size for fair comparison. The test set contain unseen individuals.

The model (backbone + main) is a ResNet-18 architecture [1]. Those kind of architecture already demonstrated good performance on this dataset [15]. We choose to use a non-pretrained version to ensure that feature learning is specific to the dataset used, allowing an unbiased evaluation of our method. In this experiment, each model is trained for 200 epochs, with a adversial lamdba of 0.1. The learning rate is set at  $1e - 4$  for the main model and  $1e - 5$  for the adversary.

**Results** Figure 2 shows the generalization effect of the three models with respect to the dataset size. As the size of the dataset increases, individual samples are less

**Table 1.** Comparison of the accuracies and loss of the different set for a train/validation set of 500 samples.

Method	Train Loss	Train Acc	Val Loss	Val Acc	Test Acc
Baseline	$7.1e-6 \pm 2.9e-6$	$0.97 \pm 0.021$	$0.517 \pm 0.008$	$0.798 \pm 0.035$	$0.641 \pm 0.026$
CCE (ours)	$0.001 \pm 0.0001$	$0.981 \pm 0.02$	$0.51 \pm 0.01$	$0.808 \pm 0.042$	$0.673 \pm 0.037$
GCE (ours)	$0.0002 \pm 0.0001$	$0.98 \pm 0.025$	$0.47 \pm 0.038$	$0.834 \pm 0.059$	<b><math>0.677 \pm 0.037</math></b>

biased and the network has more examples to generalize, even with adversarial settings, reaching a similar accuracy of 96% for all three models.

For small datasets (500 samples), both CCE and GCE training lead to better generalization, with +3.6% with respect to the baseline (see details in table 1). The addition of an adversarial network acts as a regularizer, reducing the gap between training and validation losses. Furthermore, as anticipated, some important features are discarded when the model is trained only with CCE, resulting in lower overall performance than its GCE counterpart. These observations suggest that our method is particularly beneficial when dealing with smaller datasets that suffer from limited diversity (representation bias).

**Classification attack** Additionally, using our method successfully removes features associated with bias in the data (i.e. the age), only keeping the features important for the classification of the gender. In parallel of the training, for this experiment, we trained a malicious network whose task is to predict the age of the individual in the sample.

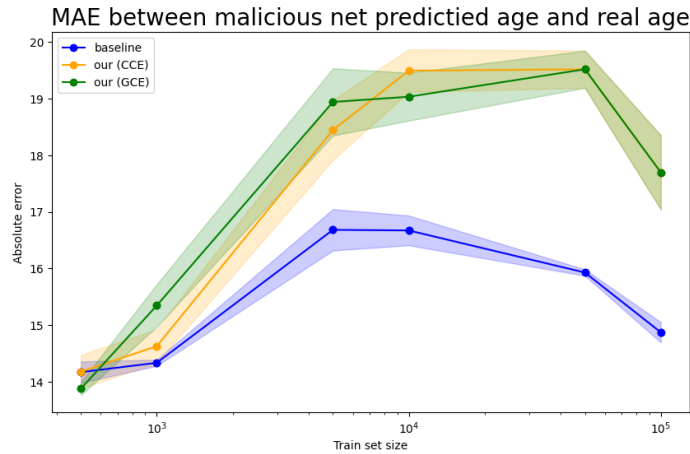
Figure 3 illustrates how the malicious network effectively predicts individual ages. We use the Mean Absolute Error (MAE) as a metric on the test set. For smaller training set sizes, our method performs as well as the baseline (with a mean error of 14 years). However, as the training set grows, the attacker has greater difficulty predicting individual ages with our method compared to the baseline, reaching a mean error of up to 19.5 years. This suggests that for smaller datasets, privacy mitigation is more effective as our model performs better, whereas with larger training sets, the privacy preservation becomes more advantageous.

## 4.2 rPPG Dataset

**Experiment Setup** For this last experiment we use our own dataset which consist of 100 subjects. The task is to estimate the rPPG signals from videos, with arm Photoplethysmogram (PPG) as ground truth.

The dataset was split into three sets: training, validation, and test. While the training and validation sets contain the same subjects (with different samples) as in our IMDb-Clean experiment, creating a pseudo-replication bias, the test set includes entirely new subjects and samples.

For task loss, we used the Pearson correlation coefficient combined with Mean Squared Error (MSE), following Unke et al. [12] and employed their PhysNet model as the experiment’s main model.



**Fig. 3.** Figure showing the mean absolute error between the malicious network prediction and the actual age of the person in the sample. We can see the privacy-enhancing effect of the adversarial network is more effective on bigger train set size.

The goal of this experiment is to assess the model’s performance in a real-world setting, with a focus on reducing **representation bias** in datasets with limited subject diversity. We trained the PhysNet model on five training set sizes 10, 20, 30, 50, and 79 subjects, keeping the test set constant across experiments. We tested two model types: a baseline model (control) and a model trained using GCE adversarial training, with hyperparameters set as follows: learning rate (adversarial, main)  $1e - 4$ , adversarial lambda 0.1, and 400 epochs.

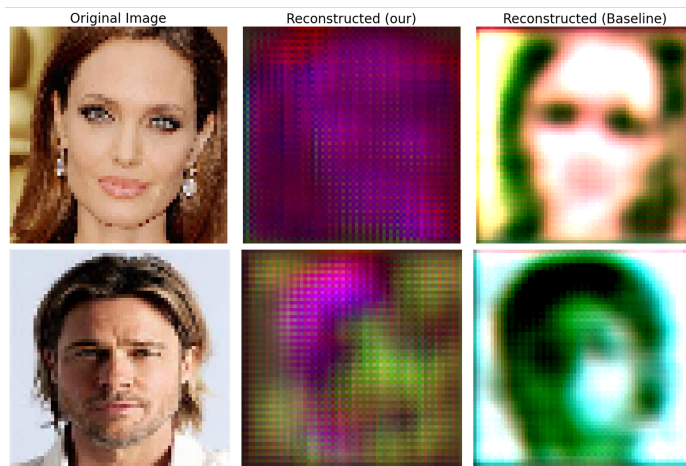
Lastly, we simulated a reconstruction attack on both the baseline and our models to evaluate the potential for face reconstruction attacks from samples, demonstrating the debiasing effect. For this, we used a transpose convolution network as a decoder to simulate the attack, training it for 100 epochs with an MSE loss and a learning rate of  $1e - 5$ .

**Results** In table 2, we observe that the model with adversarial training performs better overall. However, as seen in the previous experiment, the improvement diminishes as more distinct subjects are added to the training set, ultimately resulting in slightly worse performance than the baseline.

Furthermore, in table 2, we compare the loss of the attacker model trying to reconstruct the input image using the baseline and our model. We can see that the adversarial training reduce the possibility of reconstruction attacks. Figure 4 illustrates the potential of the method’s privacy-preserving capabilities.

**Table 2.** Test Loss Comparison for Different Numbers of Subjects in the training set and Reconstruction Loss of an attacker model.

Subjects	Test Loss (Baseline) $\pm$ Std	Test Loss (Our GCE) $\pm$ Std
10	2.4643 $\pm$ 0.0084	2.0642 $\pm$ 0.1461
20	1.5589 $\pm$ 0.1295	1.3548 $\pm$ 0.0082
30	0.9734 $\pm$ 0.1524	0.6740 $\pm$ 0.0595
50	0.4008 $\pm$ 0.0855	0.3494 $\pm$ 0.0883
79	-0.2738 $\pm$ 0.0271	-0.1586 $\pm$ 0.0210
Subjects	Rec. Loss (Baseline) $\pm$ Std	Rec. Loss (Our GCE) $\pm$ Std
10	0.9 $\pm$ 0.42	2.21 $\pm$ 0.83
20	0.8 $\pm$ 0.42	2.42 $\pm$ 0.87
30	0.92 $\pm$ 0.46	2.23 $\pm$ 0.8
50	0.91 $\pm$ 0.44	2.62 $\pm$ 0.96
79	0.82 $\pm$ 0.41	2.40 $\pm$ 0.84



**Fig. 4.** Figure demonstrating the potential of our approach. The reconstruction using an adversarial network show impressive result in privacy preservation compare to the baseline network.

## 5 Conclusion

Data collection in HCML poses significant challenges, particularly in ensuring diversity and mitigating bias. This work introduces a straightforward yet effective approach that employs an adversarial Siamese network as a regularization term, compelling the main model to unlearn features that reveal individual identities.

Our method effectively reduces representation bias, enhances privacy, and improves robustness against attacks. Validated across two datasets, it outperforms baseline models, for instance, achieving a 6.7% accuracy improvement on the IMDb-Clean dataset.

This work has been partially funded by the EU Project dAIEDGE (GA Nr 101120726)

## References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
2. He, Z., Zhang, T., Lee, R.B.: Attacking and protecting data privacy in edge–cloud collaborative inference systems. *IEEE Internet of Things Journal* **8**(12) (2020)
3. Kaluarachchi, T., Reis, A., Nanayakkara, S.: A review of recent deep learning approaches in human-centered machine learning. *Sensors* **21**(7), 2514 (2021)
4. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning not to learn: Training deep neural networks with biased data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2019)
5. Lin, Y., Shen, J., Wang, Y., Pantic, M.: Fp-age: Leveraging face parsing attention for facial age estimation in the wild. *arXiv* (2021)
6. Liu, X., Xie, L., Wang, Y., Zou, J., Xiong, J., Ying, Z., Vasilakos, A.V.: Privacy and security issues in deep learning: A survey. *arXiv* **9** (2020)
7. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* **54**(6) (2021)
8. Nam, J., Cha, H., Ahn, S., Lee, J., Shin, J.: Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems* **33** (2020)
9. Noseworthy, P.A., Attia, Z.I., Brewer, L.C., Hayes, S.N., Yao, X., Kapa, S., Friedman, P.A., Lopez-Jimenez, F.: Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ecg analysis. *Circulation: Arrhythmia and Electrophysiology* **13**(3) (2020)
10. Sepehri, Y., Pad, P., Frossard, P., Dunbar, L.A.: Priphit: Privacy-preserving hierarchical training of deep neural networks. *arXiv* (2024)
11. Tanuwidjaja, H.C., Choi, R., Baek, S., Kim, K.: Privacy-preserving deep learning on machine learning as a service—a comprehensive survey. *arXiv* **8** (2020)
12. Unke, O.T., Meuwly, M.: Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of chemical theory and computation* **15**(6) (2019)
13. Yang, J., Soltan, A.A., Eyre, D.W., Yang, Y., Clifton, D.A.: An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ digital medicine* **6**(1), 55 (2023)
14. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)
15. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: S3fd: Single shot scale-invariant face detector. In: Proceedings of the IEEE international conference on computer vision. pp. 192–201 (2017)