Deep Face Detector Adaptation without Negative Transfer or Catastrophic Forgetting

Muhammad Abdullah Jamal
University of Central Florida
Orlando, FL 32816
a_jamal@Knights.ucf.edu

Haoxiang Li
Adobe Research
San Jose, CA 95110
lhxustcer@gmail.com

Boqing Gong
Tencent AI Lab
Bellevue, WA 98004
boqinggo@outlook.com

Abstract

Arguably, no single face detector fits all real-life scenarios. It is often desirable to have some built-in schemes for a face detector to automatically adapt, e.g., to a particular user's photo album (the target domain). We propose a novel face detector adaptation approach that works as long as there are representative images of the target domain no matter they are labeled or not and, more importantly, without the need of accessing the training data of the source domain. Our approach explicitly accounts for the notorious negative transfer caveat in domain adaptation thanks to a residual loss by design. Moreover, it does not incur catastrophic interference with the knowledge learned from the source domain and, therefore, the adapted face detectors maintain about the same performance as the old detectors in the original source domain. As such, our adaption approach to face detectors is analogous to the popular interpolation techniques for language models; it may opens a new direction for progressively training the face detectors domain by domain. We report extensive experimental results to verify our approach on two massively benchmarked face detectors.

1. Introduction

Face detection is often the very first step in analyzing faces. Recent literatures [3, 4, 5, 6] demonstrate the effectiveness of deep learning for face detection. However, as a massively data-driven method, the deep learning based face detectors are inevitably biased accordingly to the training data distribution. Collecting a comprehensive dataset for training can be highly expensive, if not impossible. Besides, considering the limited computational budget in real-world applications, arguably, there is no single face detector that fits all scenarios.

To address the discrepancy between the data distribution in training and the deployment of the face detector, it is highly desirable to have some adaptation mechanism built for the face detectors. When there are labeled *or unlabeled* images available from a particular target domain, one can adapt the detectors to achieve better performance in the target domain than the original ones do.

In this paper, we propose a novel face detector adaptation approach that is applicable whenever the target domain supplies many representative images, no matter they are labeled or not. It entails some very interesting properties which we contend are missing or not explicitly discussed in the previous works of adapting face detectors [7, 8, 9].

First of all, our approach is designed to *avoid negative transfer*, i.e., the adapted detector is supposed to perform better than or at least on par with the original one in the target domain. It is worth noting that the negative transfer frequently occurs in domain adaptation [10, 11, 12], being a notoriously hard problem to solve. Moreover, this problem is likely more severe in the face detector adaptation since the room to improve the state-of-the-art face detectors is actually very small — for the same reason, we argue that it is vital for a face detector adaptation algorithm to explicitly take account of the negative transfer caveat.

Besides, we do not rely on the source data to conduct the adaptation, in a sharp contrast to most domain adaptation methods for generic visual recognition [13, 14, 15]. Indeed, the face detector adaptation is supposed to be done *without accessing the source data* because the source datasets are often extremely large and contain sensitive identity information. We note that some existing works on face detector adaptation [9] actually follow this protocol.

At last but not the least, we strive to *prevent our approach from catastrophic forgetting* or the so called interference [16, 17, 18] with the source domain. In this sense, our method is analogous to the well-known language model interpolation [19] where one extends the old language model by interpolating it with the one trained for a new domain such that, in expectation, the resulting model performs well on all old domains as well the new domain. As such, our approach may also open an alternative direction for training

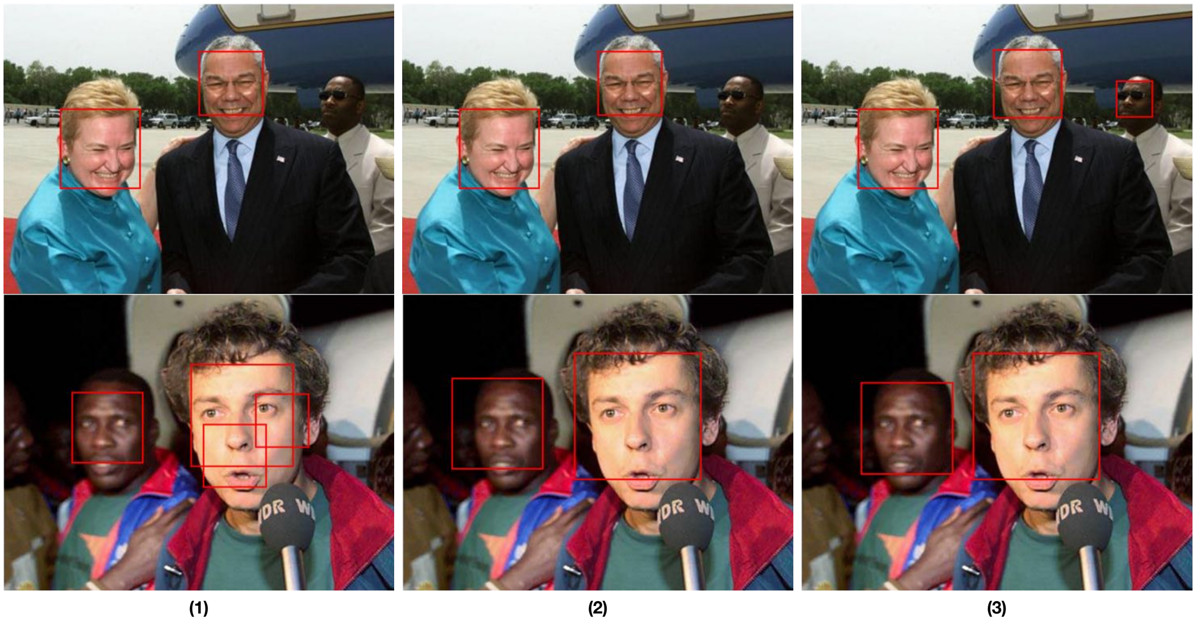


Figure 1. From left to right are face detection results on the FDDB dataset with **a state-of-the-art face detector (1)** [1, 2], the same detector but adapted by our method to the target domain (FDDB) **with no data annotation (2)**, and **with some data annotations (3)**.

the face detectors, namely, one can progressively improve the face detectors by growing the number of new domains without the need of keeping the images of the old domains.

Overview of our approach. We adapt a deep learning based face detector by fine-tuning [20, 21] it using both labeled and unlabeled images of the target domain. In order to avoid the negative transfer, we devise a loss function to approximate the expected performance improvement from the old detector to the new one. Since the hypothesis space — the set of networks specified by the weights — is the same for the two detectors, to minimize the loss does not change the old detector unless it finds another network that is expected to perform better than the old one in the target domain. While the expected performance gain of a network is mainly estimated by labeled data, we also augment it by deriving a closed form of the network's worst possible performance degradation that can be estimated by the unlabeled images of the target domain.

Our approach shares some spirits with AdaBoost [22] and residual learning [23] in the sense that the cost function of interest is a residual with respect to the source detector. Arguably, the residual loss is best captured by a residual detection score. Hence, we construct the target detector by an offset to the source one. Jointly, the residual loss and the offset detection score alleviate the urge of updating the weights of the old detector, effectively reducing the effect of catastrophic forgetting about the source domain.

The main contributions of this paper include both the novel adaptation approach and the three key properties of our method (cf. above) which we contend are missing from the previous works and yet are supposed to be possessed by a good face detector adaptation algorithm. We describe the approach in Section 3 for supervised, semi-supervised, and unsupervised settings after a review of the related works (Section 2). We present extensive experimental studies in Section 4 on two massively benchmarked face detectors.

2. Related work

Face detector adaptation. Jain and Learned-Miller use a Gaussian process to update the low detection scores by assuming smoothness of the detections and that the detected regions of high scores are more likely correct than the others [7]. Wang et al. [8] and Li et al. [9] make similar assumptions and yet use the regions of high detection scores to re-train a new detector for the target domain using vocabulary trees and probabilistic elastic part models, respectively. When the target domain comprises video sequences, the motion and tracking cues are usually very effective for adapting the detectors [24, 25, 26, 27, 28].

Domain adaptation. There has been a rich line of works on domain adaptation for generic visual recognition [13, 29], such as object recognition [14], action recognition [30], Webly-supervised learning [31, 32, 33], attribute detection [34], etc. They minimize the discrepancy between the source and target by exploring the data from both domains. However, the modern face detectors are often trained from an extreme-scale training set, making it hard to carry the source data to the adaptation stage. Domain adaptation in the absence of the source data [35, 36] is the most relevant to ours. Such methods use the source models either for regularization [36] or to augment the features of the target data [35], while we consider a different problem, deep face detectors, and refer to the source model in both the cost function and the classifier of the target face detector.

Negative transfer is a notorious caveat in domain adaptation [37, 38, 39, 40]. Whereas existing works attempt to solve this problem by defining intuitive statistical measures, we directly tackle it with a novel cost function motivated by the safe semi-supervised learning [41, 42, 43]. Nonetheless, we devise the cost function in such a way of seamlessly integrating it with the deep models. Besides, we derive an analytic form for the unsupervised adaptation, getting rid of the cumbersome EM style optimization.

Catastrophic forgetting or interference [17, 44, 45, 18] refers to that a pre-trained network cannot perform well on the old tasks after it is fine-tuned for a new task. Recent years witness an upsurge of interest in this problem, including the exploitation of a local winner-takes-all activation function [46], dropout [16, 47], a knowledge distillation loss [48, 49, 50], pathway connections [51], and progressive networks [52]. We argue that it is probably easier to deal with the catastrophic forgetting problem for domain adaptation which can be seen as a special case of sequential multi-task learning, due to that the source and target domains share the same semantic labels. We leverage exactly this idiosyncrasy to re-parameterize the target classifier as the source classifier plus an offset.

3. Approach

A face detector usually consists of two components: proposing candidate face regions from an image and classifying or scoring the regions. In this work, we adapt deep convolutional neural networks based face detectors to a given target domain by calibrating the second component, i.e., the classifiers. For simplicity, we express a deep face detector (e.g., [2]) as $\sigma(\mathbf{w}^T F(\mathbf{x};\theta))$, where $\sigma(z) = (1+\exp(-z))^{-1}$ is the sigmoid function indicating how likely the region proposal $\mathbf{x}$ out of an image is a face. The feature representations $F(\mathbf{x};\theta)$ of this region is extracted by a convolutional neural network, where θ collects all the network parameters except the classifier weights $\mathbf{w}$. Given such a detector pre-trained in the source domain, our goal is to adapt it to the target domain without using any source data and that the adapted face detector $\sigma(\widetilde{\mathbf{w}}^T F(\mathbf{x};\widetilde{\theta}))$ is not hurt by negative transfer or catastrophic forgetting.

In order to facilitate the adaptation to the target domain, we need the access to some representative images of that domain. We envision that a real use case of the face detector adaptation entails many unlabeled target images and yet only a small number or even none of labeled ones. Our approach takes account of both scenarios.

3.1. Unsupervised face detector adaptation

We first consider the unsupervised face detector adaption in which we have access to the proposed regions $\{\mathbf{x}_t\}_{t=1}^T$ of the target domain but not their labels — the labels $\{y_t \in \{0,1\}\}$ are unknown. The objective is to obtain a high-quality face detector $\sigma(\widetilde{\mathbf{w}}^T F(\mathbf{x};\widetilde{\theta}))$ for the target domain using the pre-trained face detector $\sigma(\mathbf{w}^T F(\mathbf{x};\theta))$ and the unlabeled images of the target domain.

Our approach is originally motivated by the works on safe semi-supervised learning [42, 41, 43], where the idea is to trust the classifier pre-trained from the labeled data as much as possible and to improve upon it only *relatively*. In our context, the relative performance change for any data point $(\mathbf{x}_t, y_t)$, $y_t \in \{0,1\}$, of the target domain is

$$\begin{aligned}\text{RES}_t(\widetilde{\mathbf{w}},\widetilde{\theta}) :=& \,\mathcal{C}\big(y_t, \sigma(\widetilde{\mathbf{w}}^T F(\mathbf{x}_t;\widetilde{\theta}))\big) \\ & - \mathcal{C}\big(y_t, \sigma(\mathbf{w}^T F(\mathbf{x}_t;\theta))\big), \qquad (1)\end{aligned}$$

where $\mathcal{C}(y,\hat{y})$ is a performance measure, which is implemented as the multi-class classification accuracy in [41], top-k precision, F-score, and area under the ROC curve in [42], and log-likelihood in [43]. We instead use the cross-entropy $\mathcal{C}(y,\hat{y}) = -y\log\hat{y} - (1-y)\log(1-\hat{y})$ in this paper. This choice seamlessly integrates it with the stochastic training procedure for deep neural networks.

When there are no labels available in the target domain, we find a robust target face detector that improves upon the source one under the worst case scenario,

$$\min_{\mathbf{u},\,\widetilde{\theta}} \frac{\lambda}{2}\|\mathbf{u}\|_2^2 + \mathbb{E}_t \max_{y_t \in \{0,1\}} \text{RES}_t(\mathbf{w}+\mathbf{u},\widetilde{\theta}), \qquad (2)$$

where $\mathbb{E}_t$ denotes the mean average $\frac{1}{T}\sum_{t=1}^T$. We introduce this notation to stress the fact that the expected performance change from the old face detector to the adapted one can be unbiasedly estimated by the mean average over the target examples. We overload the notation y_t a little and use the fact that the groundtruth labels are binary. We also decompose the classifier of the target detector by $\mathbf{w}+\mathbf{u}$, where $\mathbf{w}$ are the parameters of the source detector's classifier. This decomposition is mainly for two reasons. First, we can interpret Eq. (1) as the residual between the performances of the two face detectors. Arguably, this quantity is accordingly best captured by the residual detection score between the two detectors. Hence, we re-parameterize the binary classifier of the target face detector as $\widetilde{\mathbf{w}} = \mathbf{w}+\mathbf{u}$. Second, notice that the ℓ_2 regularization over the offset weights $\mathbf{u}$ effectively constrains the classifier ($\widetilde{\mathbf{w}}$) of the target face detector around that ($\mathbf{w}$) of the source detector. This prevents the classifier from shifting around, taxing less than otherwise over the network weights $\widetilde{\theta}$ for the overall target face detector to generate right predictions. Accordingly, the resultant representations $F(\mathbf{x};\widetilde{\theta})$ do not significantly deviate from the original representations $F(\mathbf{x};\theta)$ for the region proposal $\mathbf{x}$ of either source or target domain. In other words, the network *does not catastrophically forget* the knowledge extracted from the source domain.

To fit problem (2) to the existing deep learning tools (e.g., Tensorflow), we first note that there is an analytical solution to the inner maximization. Denote by $a_t = \sigma((\mathbf{w}+\mathbf{u})^T F(\mathbf{x}_t;\widetilde{\theta}))$, $\bar{a}_t = 1 - a_t$, $b_t = \sigma(\mathbf{w}^T F(\mathbf{x}_t;\theta))$, $\bar{b}_t = 1 - b_t$. We have the following,

$$\max_{y_t\in\{0,1\}} \quad \text{RES}_t(\mathbf{w}+\mathbf{u},\widetilde{\theta}), \quad \forall t \tag{3}$$

$$\begin{aligned} \Leftrightarrow \max_{y_t\in\{0,1\}} \quad & -y_t \log a_t - (1-y_t)\log \bar{a}_t \\ & + y_t \log b_t + (1-y_t)\log \bar{b}_t \end{aligned} \tag{4}$$

$$\begin{aligned} \Rightarrow \; y_t = 1 \text{ if } \log a_t + \log \bar{b}_t - \log \bar{a}_t - \log b_t < 0 \\ \text{and } y_t = 0 \text{ otherwise.} \end{aligned} \tag{5}$$

Next, we substitute the above back to Eq. (2) which then reduces to the canonical minimization problem and can be conveniently solved by programming the cost function using some off-shelf deep learning tools.

Remarks. Eq. (2) is interesting in a few ways. The residual term indicates the relative loss by the target face detector with respect to the source detector. If, for the ease of discussion, we assume the adapted face detector performs about the same on all the target examples, then the residual is large only when the source face detector does a good job and correctly classifies the data point $(\mathbf{x}_t, y_t)$ — incurring small cross-entropy loss. The data points with small cross-entropy loss values by the source detector would be penalized more, because of their relative large residuals, than the other data in the optimization process. As a result, the new face detector is enforced to imitate the source detector: if a data point is correctly classified by the source detector's classifier, so should it be by the target detector.

In our experiments, we initialize the weights of the target face detector $(\widetilde{\theta}, \mathbf{w}, \mathbf{u})$ by the source detector $(\theta, \mathbf{w}, \mathbf{0})$. Hence, after solving Eq. (2), the new detector gives rise to no higher loss than the source face detector; the residuals are either negative or zero. As a result, there is *no negative transfer* to the target domain in expectation. Moreover, since we seek to minimize the residual loss for the worst possible label assignments (cf. $\max_{y_t}$ in Eq. (2)), the obtained detector is not worse than the source one (i.e., no negative transfer) for *any* label assignments to the region proposals $\{\mathbf{x}_t\}$.

We note that the search space of the possible label assignments in Eq. (2) could be reduced by imposing similar assumptions as in [7, 8, 9]. In particular, for the region proposals whose prediction scores are high (low) by the source face detector, we may assign 1's (0's) to them. The worst case label assignment would then be applied only to the regions of which the source detector is unsure. We leave this to the future work.

3.2. Supervised face detector adaptation

In the supervised face detector adaptation, we are given a small set of labeled face images of the target domain $\{(\mathbf{x}_t, y_t)\}_{t=1}^T$ which is by itself insufficient for training a high-quality face detector. Following Eq. (2), it is now natural to write out the objective function under the supervised setting as below,

$$\min_{\mathbf{u},\,\widetilde{\theta}} \quad \frac{\lambda}{2}\|\mathbf{u}\|_2^2 + \mathbb{E}_t\, \text{RES}_t(\mathbf{w}+\mathbf{u},\widetilde{\theta}). \tag{6}$$

Note that the second cross-entropy term of Eq. (1) has no actual effect in the problem (6) — the minima of $(\mathbf{u},\widetilde{\theta})$ remain the same if we remove that term from Eq. (6). However, we keep it there for the ease of presentation.

3.3. Semi-supervised face detector adaptation

Recall that we aim to adapt a pre-trained deep neural network based face detector to the target domain that supplies many unlabeled images and possibly some labeled ones. Indeed, a real use case of the face detector adaptation likely falls under this semi-supervised regime. In this case, we initialize the target detector by copying the weights from the source detector, and then alternate between the supervised and unsupervised adaptations in our training. In particular, we update the target face detector twice in each iteration by the gradients of eq. (6) and eq. (2), respectively.

4. Experiments

Our approach is model-agnostic, in the sense that it is readily applicable to different types of face detectors. In this section, we report extensive experimental results on two massively benchmarked deep face detectors.

Face detectors and source domains. We experiment with two deep learning based face detectors: CascadeCNN [53] and Faster-RCNN [1, 2]. The CascadeCNN face detector is fast but extracts relatively weaker features while the Faster-RCNN model runs slower due to its use of a bigger network and more discriminative features.

In particular, CascadeCNN is trained by 25,000 faces from the AFLW dataset [54]. The Faster-RCNN face detector is trained using the training set of WIDER FACE dataset [6], which provides 32,203 images and 393,703 labeled faces with a high degree of variability in scale, pose, occlusion, etc. Per the comparison experiments in [2], the open-sourced Faster-RCNN face detector model is superior over 11 other top-performing detectors, all of which are published after 2015. Finally, it is interesting to note that both AFLW and WIDER FACE strive to cover a wide spectrum of face appearance variations, making them effective sources to adapt from.

The target domain. The FDDB [55] dataset is a popular face detection benchmark. It contains 2,854 images and a total of 5,171 labeled faces. The images are randomly partitioned into 10 folds, of which we use the first six as our training set, the seventh for validation, and the remaining three for testing. We also evaluate our method on Caltech Occluded Faces in the Wild (COFW) dataset [56]. Due to limited space, we report the results on COFW in the supplementary materials.

We claim that this choice — WIDER FACE or AFLW as the source domain and FDDB as the target domain — well represents the real application scenarios of face detector adaptation. On the one hand, there is a large training set in the source domain for us to learn a generic face detector that performs very well on different testing sets. WIDER FACE relies on diverse data sources since it employs Google and Bing to acquire the images and AFLW is a large-scale dataset collected from Flicker. On the other hand, the target domain of FDDB images are relatively homogeneous, all sampled from the Yahoo! news website. They are mostly professional photos sharing some common idiosyncrasies.

Evaluation metrics. Both WIDER FACE and FDDB datasets have defined and released the code for standard evaluation metrics. The Precision-Recall curve is used by WIDER FACE. FDDB employs the ROC curves of discrete and continuous scores computed from a bipartite graph. We use their code to evaluate our results in order to have direct comparison with existing methods.

Competing methods. We compare our approach to the following competing baselines [1].

- **Source** refers to the detectors trained from the original training data and is the starting point for our method to fine-tune the neural network parameters.
- **Fine-tuning [20]** simply fine-tunes the models using the labeled data of the target domain, if they are available, following the same way the detectors are trained in their source domains yet with smaller learning rates.
- **GP [7]** is a Gaussian process based unsupervised face detector adaptation method which uses the regions of high detection confidence — far from $p = 0.5$ — to update the detection scores of the other regions.
- **LWF [57]** is a recent learning without forgetting (LWF) method that augments the conventional cross-entropy loss with the knowledge distillation loss [50] such that the adapted face detector preserves the response characteristics learned from the source domain.
- **GDSDA [58]** introduces the generalized distillation [59] into semi-supervised domain adaptation.
- **HTL [36]** is a representative hypothesis transfer method that transfers knowledge from the source domain to the target by augmenting the feature representations of the target domain.
- *Gradient Reversal* [60] is an effective method for the domain adaptation of deep neural networks. The main idea is to learn representations to fail the classifier that predicts from which domain a data point comes. Since it has to access the source domain data, it is actually not fair to compare this method with the other baselines or ours. Nonetheless, we still include its results in the FDDB experiment for reference.

Some experimental details. We freeze the first eight convolutional layers of the Faster-RCNN model for all the experiments. We fine-tune all parameters of the last 48-net detection net in the CascadeCNN model. The validation set of the target domain is used to determine the hyper-parameters of all the methods. For Faster-RCNN, we use $\lambda = 1e$-3 and the base learning rates $1e$-4 and $5e$-4 for the supervised and unsupervised settings, respectively. Early stopping happens at the 5,000th iteration for the supervised experiment and the 6,000th for the unsupervised. For CascadeCNN, we set $\lambda = 2$ and the base learning rate $1e$-4 for both supervised and unsupervised settings. For the supervised case, we fine-tune the model for 8,000 iterations with the base learning rate and another 4,000 iterations with the learning rate of $1e$-5. For the unsupervised, we fine-tune the model for 10,000 iterations and divide the base learning rate by 10 at the 7,000th iteration.

4.1. Comparison results

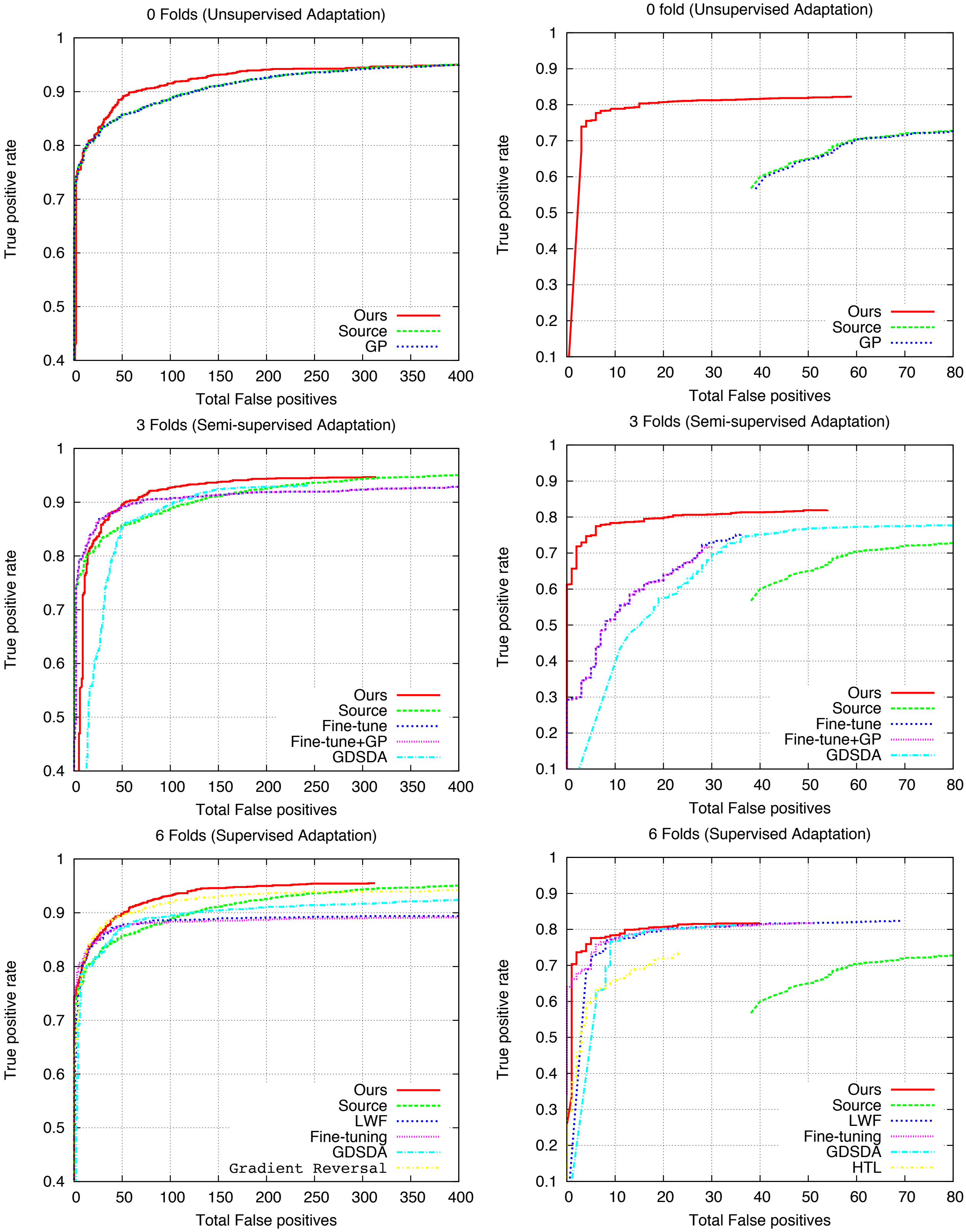
We compare our algorithm with other competing methods in this section. We evaluate the effectiveness of all the methods by varying the number of labeled data from the target domain. More specifically, all the methods have access to the 6 folds of training images for the adaptation, while only N folds out of the 6 are labeled, $N \in \{0, 1, 3, 5, 6\}$. It is a fully unsupervised setting when $N = 0$, a semi-supervised adaptation setting when $1 \leq N \leq 5$, and a supervised adaptation setting when $N = 6$. Note that not all the baseline methods can handle all the settings.

Figure 2 and Figure 3 together show the ROC curves of the discrete scores on FDDB for the (a) CascadeCNN detector and (b) Faster-RCNN detector; the curves of the continuous scores are included in the supplementary materials.

When $N = 0$ (unsupervised adaptation), most of the above-mentioned competing methods are not applicable any more. As shown in Figure 2, in this challenging setting, we observe **GP** cannot improve the pre-trained high-quality face detectors while our method still brings extra gains.

When $N = 6$, all the training images of the target domain are labeled (supervised adaptation), we outperform

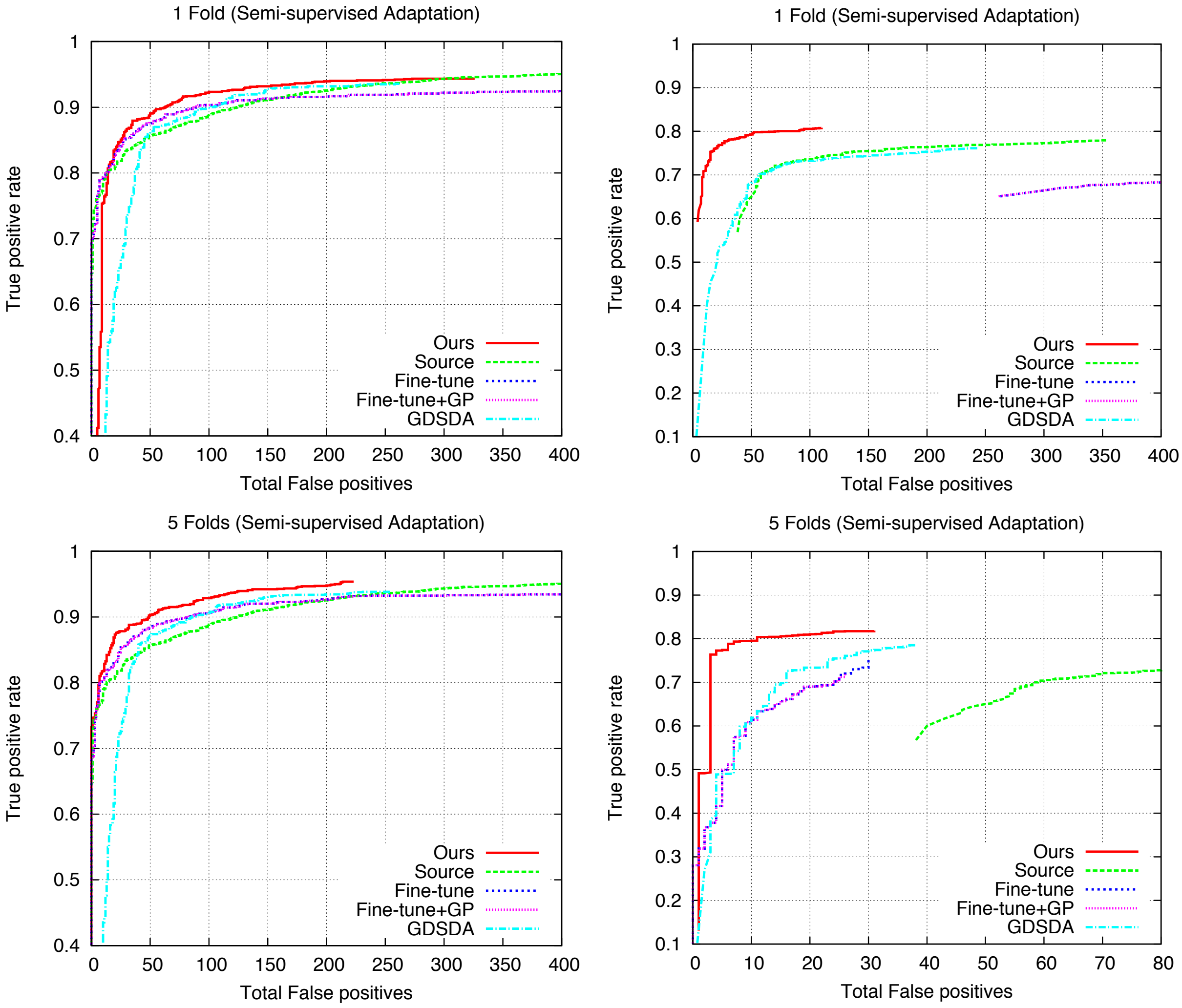
[1] Please refer to our supplementary materials for the training details of the competing methods.



(a) FasterRCNN Detector Adaptation

(b) CascadeCNN Detector Adaptation

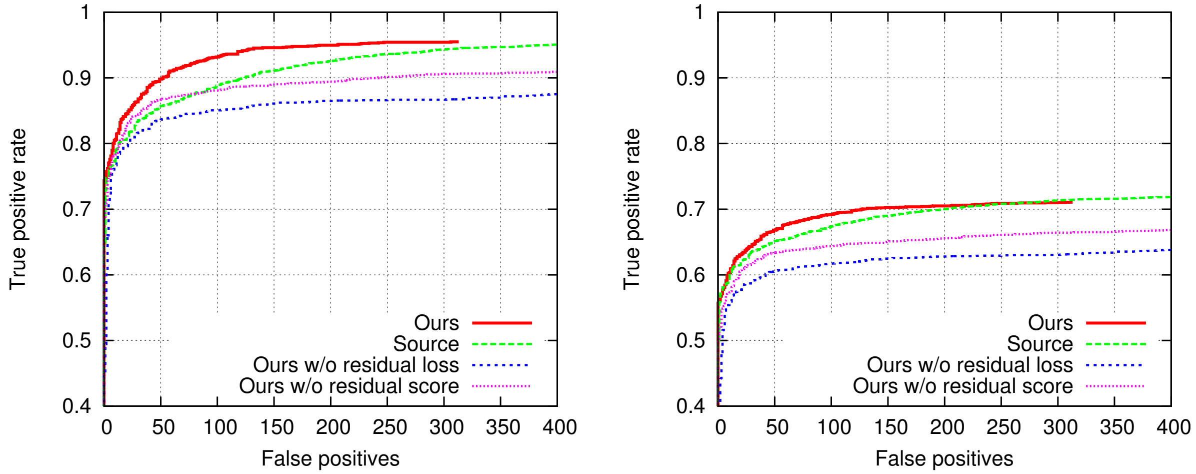
Figure 2. Detection results comparison on FDDB under unsupervised (0 out of 6 folds labeled), semi-supervised (3 out of 6 folds labeled), and supervised settings: our method generally outperforms all competing methods and does not suffer from negative transfer.



(a) FasterRCNN Detector Adaptation

(b) CascadeCNN Detector Adaptation

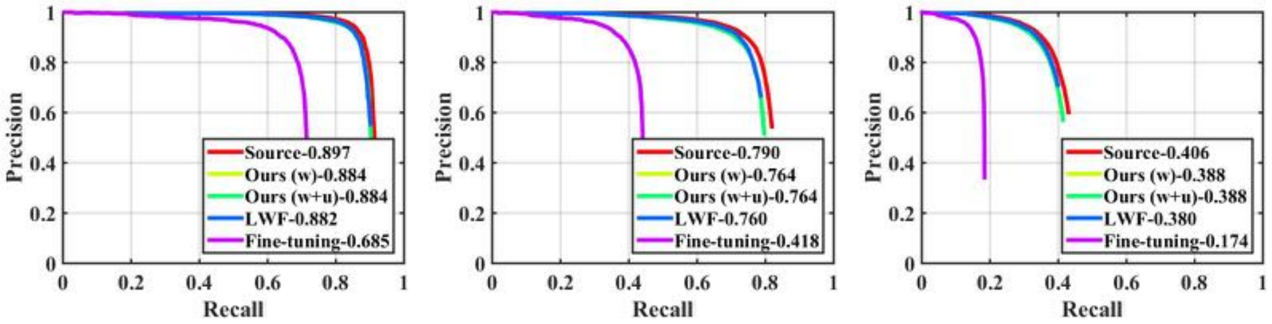
Figure 3. More detection results under semi-supervised settings with $N = \{1, 5\}$ out of 6 folds training images annotated. Combined with Figure 2, our method can generally bring additional performance gains from additional annotated data.



(a) Discrete Scores

(b) Continuous Scores

Figure 4. Ablation Studies about our approach on FDDB (supervised adaptation)



(a) Easy Set (b) Medium Set (c) Hard Set

Figure 5. Evaluation of catastrophic forgetting on source domain after supervised adaptation to target domain: detection results on validation set of WIDER FACE (Easy, Medium and Hard sets).

all the competing methods when adapting the CascadeCNN detector. Even for the high-quality FasterRCNN detector, our method gives rise to the largest improvement among all the methods, including *Gradient Reversal* which takes advantage of the extra training data in the source domain.

Under the semi-supervised setting, which is more realistic, our method achieves significant and consistent improvement for both face detectors over the original **Source** detectors. With the additional results shown in Figure 3[2], varying N from 0 to 6, our method generally performs better and better as more annotated data become available.

Overall, compared with **Source** models, our method does not cause negative transfer, while all the other competing methods suffer from negative transfer to some extent excluding *Gradient Reversal*.

4.2. Ablation study

We investigate our proposed method by examining its ablated versions. Recall that our approach is two-pronged. On the one hand, it uses the residuals in the cost function to explicitly prevent negative transfer in terms of the cross-entropy loss. On the other hand, it re-parameterizes the classifier of the target detector by $\widetilde{\mathbf{w}} = \mathbf{w} + \mathbf{u}$, where $\mathbf{w}$ is the classifier weights of the source detector. Figure 4 shows that both components contribute to the performance improvement in our method. The ROC curve of the source detector is included for reference. Clearly, we observe that the two components mutually complement. Besides, removing the residual loss (Ours w/o residual loss) hurts our method more than directly optimizing the classifier weights $\widetilde{\mathbf{w}}$ without reparameterization (Ours w/o residual score).

4.3. The effect of no catastrophic forgetting

Finally, we evaluate the catastrophic forgetting in the domain adaptation context. After adapting all competing methods to the target domain (FDDB), we evaluate their performance back to the source domain (WIDER Face). We test on the validation set of the WIDER Face in our experiment. **Source** refers to the one without adaptation and is thus with no forgetting at all.

As shown in Figure 5, it is not surprising to see that fine-tuning leads to severe forgetting about the source domain. This observation is well-aligned with prior arts. After all, domain adaptation can be seen as a special case of the sequential multi-task learning, under which previous studies have shown that fine-tuning causes catastrophic forgetting [16, 48]. Both **LWF** and our methods maintain a reasonably good performance in the source domain compared with the **Source** detector. **LWF** prevents forgetting about the source domain using a knowledge distillation loss, while we do so by the residual loss coupled with the residual detection score. Thanks to the ℓ_2 regularization over the offset vector $\mathbf{u}$ in the classifier of the adapted detector, there is no noticeable difference between the new classifier $(\mathbf{w} + \mathbf{u})$ and that $(\mathbf{w})$ of the source face detector. We test both classifiers stacked over the network of the adapted detector and find that their corresponding curves almost overlap, as shown in Figure 5.

5. Conclusion

In this paper, we revisit the face detector adaptation problem under the new context of deep learning based face detectors. The approach we proposed offers three key properties which we contend are missing or not explicitly discussed in the existing face detector adaptation works. In short, the adaptation of face detectors is supposed to be executed in the absence of the source domain's data, with little negative transfer, and incurring no catastrophic forgetting about the source domain. Our approach explicitly accounts for all the requirements by two residuals: a residual loss to avoid negative transfer and a residual classifier to alleviate catastrophic forgetting. We demonstrated the effectiveness of our approach by adapting two face detectors from two large-scale source datasets to two smaller target datasets.

Acknowledgements. The authors wish to thank the support of NSF awards IIS-1741431 and IIS-1566511 and two gifts respectively from Adobe and NVIDIA.

[2] The scale of the horizontal axis of the top-right panel differs from the other panels of CascadeCNN. If we used the same scale as the others instead, the fine-tuning results would be left out.

References

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2, 4

[2] Huaizu Jiang and Erik G. Learned-Miller. Face detection with the faster R-CNN. *CoRR*, abs/1606.03473, 2016. 2, 3, 4

[3] Yu Liu, Hongyang Li, Junjie Yan, Fangyin Wei, Xiaogang Wang, and Xiaoou Tang. Recurrent scale approximation for object detection in cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[4] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S. Davis. Ssh: Single stage headless face detector. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[5] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. S3fd: Single shot scale-invariant face detector. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[6] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 4

[7] Vidit Jain and Erik G. Learned-Miller. Online domain adaptation of a pre-trained cascade of classifiers. 2011. 1, 2, 4, 5

[8] Xiaoyu Wang, Gang Hua, and T.X. Han. Detection by detections: Non-parametric detector adaptation for a video. 2012. 1, 2, 4

[9] Haoxiang Li, Gang Hua, Zhe Lin, Jonathan Brandt, and Jianchao Yang. Probabilistic elastic part model for unsupervised face detector adaptation. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013. 1, 2, 4

[10] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International Conference on Algorithmic Learning Theory*, pages 38–53. Springer, 2008. 1

[11] Boqing Gong, Kristen Grauman, and Fei Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, pages 222–230, 2013. 1

[12] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015. 1

[13] Raghuraman Gopalan, Ruonan Li, Vishal M Patel, Rama Chellappa, et al. Domain adaptation for visual recognition. *Foundations and Trends® in Computer Graphics and Vision*, 8(4):285–378, 2015. 1, 2

[14] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. *Computer Vision–ECCV 2010*, pages 213–226, 2010. 1, 2

[15] Boqing Gong, Kristen Grauman, and Fei Sha. Learning kernels for unsupervised domain adaptation with applications to visual object recognition. *International Journal of Computer Vision*, 109(1-2):3–27, 2014. 1

[16] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. 1, 3, 8

[17] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation*, 24:109–165, 1989. 1, 3

[18] JL McClelland. A connectionist perspective on knowledge and development. 1995. 1, 3

[19] Daniel Jurafsky and James H. Martin. *Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2017. 1

[20] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *Artificial Intelligence and Statistics*, pages 448–455, 2009. 2, 5

[21] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014. 2

[22] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156, 1996. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[24] Kevin Tang, Vignesh Ramanathan, Li Fei-Fei, and Daphne Koller. Shifting weights: Adapting object detectors from image to video. In *Advances in Neural Information Processing Systems*, pages 638–646, 2012. 2

[25] Pramod Sharma and Ram Nevatia. Efficient detector adaptation for object detection in a video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 2

[26] Meng Wang and Xiaogang Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3401–3408. IEEE, 2011. 2

[27] Enver Sangineto. Statistical and spatial consensus collection for detector adaptation. In *European Conference on Computer Vision*, pages 456–471. Springer, 2014. 2

[28] Peter M. Roth, Sabine Sternig, Helmut Grabner, and Horst Bischof. Classifier grids for robust adaptive object detection. In *cvpr*, 2009. 2

[29] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017. 2

[30] Ruonan Li and Todd Zickler. Discriminative virtual views for cross-view action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2855–2862. IEEE, 2012. 2

[31] Alessandro Bergamo and Lorenzo Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *Advances in Neural Information Processing Systems*, pages 181–189, 2010. 2

[32] Lixin Duan, Dong Xu, Ivor Wai-Hung Tsang, and Jiebo Luo. Visual event recognition in videos by learning from web data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1667–1680, 2012. 2

[33] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1431–1439, 2015. 2

[34] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 87–97, 2016. 2

[35] Boris Chidlovskii, Stéphane Clinchant, and Gabriela Csurka. Domain adaptation in the absence of source domain data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 451–460. ACM, 2016. 2

[36] Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *ICML (3)*, pages 942–950, 2013. 2, 5

[37] Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich. To transfer or not to transfer. In *NIPS 2005 Workshop on Transfer Learning*, volume 898, 2005. 3

[38] Liang Ge, Jing Gao, Hung Ngo, Kang Li, and Aidong Zhang. On handling negative transfer and imbalanced distributions in multiple source transfer learning. *Statistical Analysis and Data Mining*, 7(4):254–271, 2014. 3

[39] Chun-Wei Seah, Yew-Soon Ong, and Ivor W Tsang. Combating negative transfer from predictive distribution differences. *IEEE transactions on cybernetics*, 43(4):1153–1165, 2013. 3

[40] Hao Shao, Bin Tong, and Einoshin Suzuki. Compact coding for hyperplane classifiers in heterogeneous environment. *Machine Learning and Knowledge Discovery in Databases*, pages 207–222, 2011. 3

[41] Yu-Feng Li and Zhi-Hua Zhou. Towards making unlabeled data never hurt. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 1081–1088, New York, NY, USA, June 2011. ACM. 3

[42] Yu-Feng Li, James T Kwok, and Zhi-Hua Zhou. Towards safe semi-supervised learning for multivariate performance measures. In *AAAI*, pages 1816–1822, 2016. 3

[43] Marco Loog. Contrastive pessimistic likelihood estimation for semi-supervised classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):462–475, 2016. 3

[44] Roger Ratcliff. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285–308, 1990. 3

[45] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. 3

[46] Rupesh K Srivastava, Jonathan Masci, Sohrob Kazerounian, Faustino Gomez, and Jürgen Schmidhuber. Compete to compute. In *Advances in neural information processing systems*, pages 2310–2318, 2013. 3

[47] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 3

[48] Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, pages 614–629. Springer, 2016. 3, 8

[49] Matthew Riemer, Elham Khabiri, and Richard Goodwin. Representation stability as a regularizer for improved text analytics transfer learning. *arXiv preprint arXiv:1704.03617*, 2017. 3

[50] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3, 5

[51] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017. 3

[52] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 3

[53] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *CVPR*, pages 5325–5334. IEEE Computer Society, 2015. 4

[54] Peter M. Roth Martin Koestinger, Paul Wohlhart and Horst Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011. 4

[55] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, 2010. 5

[56] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, ICCV '13, pages 1513–1520, Washington, DC, USA, 2013. IEEE Computer Society. 5

[57] Zhizhong Li and Derek Hoiem. Learning without forgetting. *CoRR*, abs/1606.09282, 2016. 5

[58] Shuang Ao, Xiang Li, and Charles X Ling. Fast generalized distillation for semi-supervised domain adaptation. In *AAAI*, 2017. 5

[59] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015. 5

[60] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015. 5