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. Our approach is applicable whenever the target domain supplies many representative images, no matter they are labeled or not. 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...
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 \cite{20, 21} it using both la-
beled 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 tar-
get domain. While the expected performance gain of a net-
work is mainly estimated by labeled data, we also augment
it by deriving a closed form of the network’s worst possi-
ble performance degradation that can be estimated by the
unlabeled images of the target domain.

Our approach shares some spirits with AdaBoost \cite{22}
and residual learning \cite{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 as-
suming smoothness of the detections and that the detected
regions of high scores are more likely correct than the oth-
ers \cite{7}. Wang et al. \cite{8} and Li et al. \cite{9} make similar as-
umptions and yet use the regions of high detection scores
to re-train a new detector for the target domain using vo-
cabulary trees and probabilistic elastic part models, respec-
tively. When the target domain comprises video sequences,
the motion and tracking cues are usually very effective for
adapting the detectors \cite{24, 25, 26, 27, 28}.

Domain adaptation. There has been a rich line of works on
domain adaptation for generic visual recognition \cite{13, 29},
such as object recognition \cite{14}, action recognition \cite{30},
Webly-supervised learning \cite{31, 32, 33}, attribute detec-
tion \cite{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 \cite{35, 36} is the most rele-
vant to ours. Such methods use the source models either
for regularization \cite{36} or to augment the features of the tar-
get data \cite{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, 46] 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 \( \theta \) 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 \( \sigma(\mathbf{w}^T F(\mathbf{x}; \bar{\theta})) \) 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 pre-trained in the source domain, our goal 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

\[
\text{RES}_t(\mathbf{w}, \bar{\theta}) := C(y_t, \sigma(\mathbf{w}^T F(\mathbf{x}_t; \bar{\theta}))) - C(y_t, \sigma(\mathbf{w}^T F(\mathbf{x}_t; \theta))),
\]

where \( C(y, \hat{y}) \) 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 \( C(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}) \) 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,

\[
\min_{\mathbf{u}, \bar{\theta}} \frac{\lambda}{2} \left\| \mathbf{u} \right\|^2 + \mathbb{E}_t \max_{y_t \in \{0, 1\}} \text{RES}_t(\mathbf{w} + \mathbf{u}, \bar{\theta}),
\]

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 \( \mathbf{w} = \mathbf{w} + \mathbf{u} \). Second, notice that the \( \ell_2 \) regularization over the offset weights \( \mathbf{u} \) effectively constrains the classifier \( (\mathbf{w}) \) 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 \( \bar{\theta} \) for the overall target face detector to generate right predictions. Accordingly, the resultant representations \( F(\mathbf{x}, \bar{\theta}) \) 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.

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 \( \sigma(\mathbf{w}^T F(\mathbf{x}; \bar{\theta})) \) 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
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 \( a_t = \sigma((w + u)^T F(x_t; \tilde{\theta})) \), \( \tilde{a}_t = 1 - a_t \), \( b_t = \sigma(w^T F(x_t; \tilde{\theta})) \), \( \tilde{b}_t = 1 - b_t \). We have the following,

\[
\max_{y_t \in \{0,1\}} \text{RES}_t(w + u, \tilde{\theta}), \quad \forall t \tag{3}
\]

\[
\Leftrightarrow \max_{y_t \in \{0,1\}} -y_t \log a_t - (1 - y_t) \log \tilde{a}_t \\
+ y_t \log b_t + (1 - y_t) \log \tilde{b}_t \tag{4}
\]

\[
\Rightarrow y_t = 1 \text{ if } \log a_t + \log \tilde{b}_t - \log \tilde{a}_t - \log b_t < 0 \\
\text{and } y_t = 0 \text{ otherwise.} \tag{5}
\]

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 \((x_t, y_t)\) — 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 \((\tilde{\theta}, w, u)\) by the source detector \((\tilde{\theta}, w, 0)\). 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 \(\{x_t\}\).

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 \(\{(x_t, y_t)\}_{t=1}^T\) 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,

\[
\min_{u, \tilde{\theta}} \frac{\lambda}{2} \|u\|^2 + \mathbb{E}_{t} \text{RES}_t(w + u, \tilde{\theta}). \tag{6}
\]

Note that the second cross-entropy term of Eq. (1) has no actual effect in the problem (6) — the minima of \((u, \tilde{\theta})\) 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
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.
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.

Figure 4. Ablation Studies about our approach on FDDB (supervised adaptation)
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, 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 \( \mathbf{w} = \mathbf{w} + \mathbf{u} \), 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 \( \mathbf{w} \) without re-parameterization (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 (FDDDB), 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 \( \ell_2 \) 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.
References


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