Open Set Domain Adaptation
Supplementary material

For completeness, we report the accuracies of our method for popular domain adaptation datasets using the standard closed set protocol, where all classes are known in both domains.

Office dataset

For the *Office* dataset [11], we run experiments for the 6 domain shifts of the three provided datasets: *Amazon (A)*, *DSLR (D)* and *Webcam (W)*. We use deep features extracted from the fully connected layer-7 (fc7) from the AlexNet model [7].

Unsupervised setting In the unsupervised domain adaptation, we firstly report the classification accuracies following the protocol from [11], where we run 5 experiments for each domain shift using randomised subsamples of the source dataset. The results of the described techniques are shown in Table 1, where we compare our method with generic domain adaptation methods, *i.e*. TCA [10], gfk [6], SA [2] and CORAL [12]. At first, we see that ATI outperforms all generic domain adaptation methods in average, reporting ATI-λ a slight better improvement than ATI. However, the experiments with the lowest domain shifts between *DSLR* and *Webcam* report no improvement. The locality constrained formulation, ATI-λ-N_1, does perform worse in the unsupervised setting.

	A→D	A→W	D→A	D→W	W→A	W→D	AVG.
LSVM	62.3±3.8	55.8±3.1	42.8±1.6	90.1±0.6	41.2±0.4	92.6±1.5	64.1
TCA [10]	60.3±4.0	54.7±3.0	49.4±1.6	90.7±0.4	46.9±2.3	92.0±0.9	65.7
gfk [6]	61.3±3.7	55.7±3.0	45.6±1.6	90.6±0.4	43.1±2.3	93.4±0.9	65.0
SA [2]	60.6±3.5	55.0±3.1	47.3±1.6	90.9±0.6	44.4±1.4	93.3±0.8	65.3
CORAL [12]	64.4±3.9	58.9±3.3	52.1±1.2	**92.6±0.3**	50.0±1.0	**94.0±0.6**	68.7
ATI	**67.6±3.0**	62.3±3.1	54.8±1.3	90.3±0.8	52.4±2.1	92.6±1.7	70.0
ATI-λ	67.3±2.3	**62.6±2.5**	**55.2±2.6**	90.1±0.6	**53.4±2.5**	92.7±2.5	**70.2**
ATI-λ-N_1	64.6±2.9	60.9±1.3	51.9±1.9	90.2±0.9	48.1±1.6	93.7±2.1	68.2

Table 1: Comparison on the unsupervised Office dataset [11] with 31 shared classes and 6 domain shifts using the protocol from [11].

In order to compare our method with current state-of-the-art CNN-based domain adaptation methods [8, 9, 3], we also report the accuracies when taking all source samples in a single run as described by [5]. As shown in Table 2, our method outperforms the state-of-the-art in average. While [9] performs better when both datasets are similar, our methods outperforms [9] by +5.7% and +5.9% on the two most difficult cases D→A and W→A, respectively.

	A→D	A→W	D→A	D→W	W→A	W→D	AVG.
LSVM	65.7	60.3	43.2	94.7	44.0	98.9	67.8
DAN [8]*	66.8	68.5	50.0	96.0	49.8	99.0	71.7
RTN [9]*	**71.0**	**73.3**	50.5	**96.8**	51.0	**99.6**	73.7
BP [3]*	-	73.0	-	96.4	-	99.2	-
ATI	70.3	68.7	55.3	95.0	**56.9**	98.7	**74.2**
ATI-λ	69.0	67.0	**56.2**	95.0	**56.9**	98.7	73.8

Table 2: Comparison on the unsupervised Office dataset [11] with 31 shared classes and 6 domain shifts taking all source samples as in [5]. * Numbers are taken from [9].

	A→D	A→W	D→A	D→W	W→A	W→D	AVG.
LSVM (s)	64.6±3.8	56.4±2.7	45.8±1.5	90.5±0.4	42.3±1.4	93.6±1.2	65.5
LSVM (t)	80.1±3.0	76.4±3.8	58.5±1.4	76.4±3.8	58.5±1.4	80.1±3.0	71.7
LSVM (st)	82.6±5.5	77.0±2.5	63.4±1.6	94.0±0.8	61.8±1.1	96.3±0.8	79.2
DDC [16]*	-	84.1±0.6	-	95.4±0.4	-	96.3±0.3	-
DAN [8]*	-	**85.7±0.3**	-	**97.2±0.2**	-	96.4±0.2	-
MMC [15]*	86.1±1.2	82.7±0.8	**66.2±0.3**	95.7±0.5	65.0±0.5	**97.6±0.2**	**82.2**
ATI (labels t)	85.0±2.1	78.3±2.3	63.6±1.5	94.0±0.8	62.3±0.9	96.4±0.8	79.9
ATI	85.5±2.9	82.4±1.1	65.1±1.3	93.4±0.9	65.6±1.5	95.7±1.1	81.3
ATI-λ	85.6±2.6	82.6±0.5	65.3±1.3	93.3±1.0	65.7±1.7	95.7±1.1	81.4
ATI-λ-N_1	**88.1±1.7**	83.1±2.3	66.0±1.4	93.9±1.2	**65.9±1.5**	96.2±0.8	**82.2**
ATI-λ-N_2	87.0±3.5	84.6±3.5	65.3±1.0	93.6±1.2	**65.9±1.8**	95.8±1.3	82.0

Table 3: Comparison on the semi-supervised Office dataset [11] with 31 shared classes and 6 domain shifts, following the protocol from [11]. * Numbers are taken from [8] and [15].

Semi-supervised setting We also test the *Office* dataset in its semi-supervised setting, following the protocol from [11] in 5 runs with random subsamples of the source dataset. In this experiments, we also include ATI-λ-N_2 with locality constraints using 2 nearest neighbours and compared our methods with the state-of-the-art CNN-based methods [16, 8, 15]. The accuracies of our methods are based on the joint linear SVM training with all source samples and the provided labelled target samples. The results are reported in Table 3. We also report the accuracies when we estimate the mapping W (6) using only the labelled target samples without solving the individual assignments (1), which is denoted by ATI (labels t). This performs worse than ATI and the best result is achieved by ATI-λ-N_1. Our method achieves the same average accuracy as MMC [15].

Office+Caltech dataset

	A→C	A→D	A→W	C→A	C→D	C→W	D→A	D→C	D→W	W→A	W→C	W→D	AVG
LSVM	83.3	84.1	77.5	91.8	89.1	82.3	79.4	70.2	97.9	80.0	72.7	**100.0**	84.0
CORAL [12]	83.2	86.5	79.6	91.4	86.6	82.1	87.3	77.5	**99.3**	85.2	76.1	**100.0**	86.2
BP [3]	84.6	92.3	90.2	91.9	92.8	93.2	84.0	74.9	97.8	86.9	77.3	**100.0**	88.2
DDC [16]*	83.5	88.4	83.1	91.9	88.8	85.4	89.0	79.2	98.1	84.9	73.4	**100.0**	87.1
DAN [8]*	84.1	91.1	91.8	92.0	89.3	90.6	90.0	80.3	98.5	92.1	81.2	**100.0**	90.1
RTN[9]*	**88.1**	**95.5**	**95.2**	93.7	**94.2**	**96.9**	**93.8**	84.6	99.2	**95.5**	**86.6**	**100.0**	**93.4**
ATI	86.5	92.8	88.7	**93.8**	89.6	93.6	93.4	**85.9**	98.9	93.6	86.3	**100.0**	91.9
ATI-λ	87.1	90.6	90.7	93.4	85.4	93.4	93.6	85.8	**99.3**	93.6	86.1	**100.0**	91.8

Table 4: Classification accuracies on the unsupervised Office+Caltech dataset [6] with 10 shared classes and 12 domain shifts using deep features. We take all source samples on a single run [5]. * Numbers are taken from [9].

We also test the performance of our method with the extended version of the Office evaluation set [6], including an additional dataset: *Caltech (C)*. This setup allows for a total of 12 domain shifts, but reduces the amount of shared classes to only 10. As shown in Table 4, our method obtains very competitive results, outperforming in overall the generic domain adaptation method [12] and 3 out of 4 CNN-based methods.

Dense Testbed for Cross-Dataset Analysis

We also present an evaluation on the Dense dataset of the Testbed for Cross-Dataset Analysis [13], using the precomputed DeCaF features that they provide, for a total of 40 shared classes in 12 domain shifts from 4 popular datasets: *Bing (B)*, *Caltech (C)*, *ImageNet (I)* and *Sun (S)*. Following the protocol described in [13], we take 50 source samples per class for training and we test on 30 target images per class for all datasets, except *Sun*, where we take 20 samples per class. Reported results in Table 5 show that we outperform all generic domain adaptation methods, being ATI-λ the best reporting method.

	B→C	B→I	B→S	C→B	C→I	C→S	I→B	I→C	I→S	S→B	S→C	S→I	AVG
LSVM	63.8±2.2	57.4±0.7	20.2±1.0	38.3±0.8	62.9±0.9	21.7±1.6	39.3±1.4	70.8±1.5	24.6±1.8	16.6±1.0	26.1±2.0	26.3±0.7	39.0
TCA [10]	53.8±1.3	49.1±1.1	17.1±1.1	35.6±1.8	59.2±0.8	18.9±1.2	36.4±1.2	66.3±2.3	22.2±1.4	13.8±1.4	23.2±1.5	23.2±1.5	34.9
gfk [6]	63.4±1.8	57.2±1.1	20.6±1.3	38.3±0.9	62.9±1.2	21.7±1.4	38.8±1.3	70.9±1.1	24.4±1.4	16.3±0.9	26.7±1.8	26.1±1.0	38.9
SA [2]	63.0±1.9	57.1±1.4	20.2±1.4	38.3±0.9	62.8±1.0	21.5±1.2	39.0±1.3	71.1±1.3	24.2±1.4	16.0±0.9	26.8±1.9	26.4±1.1	38.9
CORAL [12]	63.9±2.1	57.8±0.8	20.4±2.0	38.3±0.8	63.4±0.9	22.5±1.2	39.0±1.2	71.2±1.3	24.9±1.6	16.8±1.0	27.4±2.2	27.7±0.5	39.4
ATI	69.1±1.3	62.4±1.9	23.4±1.1	**39.0±1.4**	**66.9±1.2**	25.2±0.9	39.7±1.8	74.4±1.6	**25.9±2.1**	18.3±1.1	37.1±3.2	**35.0±1.0**	42.8
ATI-λ	**69.4±1.4**	**62.9±1.3**	**23.6±1.0**	**39.0±1.4**	**66.9±1.1**	**25.3±0.9**	**39.8±1.8**	**74.8±1.5**	25.8±2.0	**18.7±0.7**	**37.4±2.9**	34.8±0.8	**43.2**

Table 5: *Testbed* dataset [14] with 40 common classes for a total of 12 domain shifts.

Sentiment Analysis

	B→E	D→B	E→K	K→D	AVG.
LSVM	75.5±1.6	78.2±2.5	83.1±1.8	73.3±1.8	77.5
TCA [10]	76.6±2.2	78.5±1.6	**83.8±1.5**	75.0±1.4	78.5
gfk [6]	77.0±2.0	**79.2±1.8**	83.7±1.7	73.7±1.9	78.4
SA [2]	75.9±1.9	78.4±2.1	83.0±1.7	72.1±1.9	77.4
CORAL [12]	76.2±1.7	78.4±2.0	83.1±2.0	74.2±3.0	78.0
ATI	**79.9±2.0**	**79.2±1.9**	83.7±2.1	**75.6±1.9**	**79.6**
ATI-λ	79.6±1.4	79.0±1.8	83.6±2.1	74.4±1.7	79.2

Table 6: Accuracies of 4 domain shifts on the Sentiment dataset [1] using the bag-of-words features and the protocol from [4].

To show the behaviour of our method with a different type of feature descriptor, we also present an evaluation on the *Sentiment analysis* dataset [1]. This dataset gathers reviews from Amazon for four products: *books (B)*, *DVDs (D)*, *electronics (E)* and *kitchen appliances (K)*. Each domain contains 1000 reviews labelled as *positive* and another set of 1000 reviews as *negative*. We use the data provided by [4], which extracts bag-of-words features from the 400 words with the largest mutual information across domains. We report the mean accuracy of experiments on 20 random splits, where at each run the training set contains 1600 samples and the test set 400 samples. Table 6 shows how our method obtains the best overall results. In this case, ATI-λ gets slightly worse accuracies than the standard ATI without outlier rejection, probably due to the nature of the dataset with only two class labels.

References

[1] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Association for Computational Linguistics*, 2007.

[2] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *IEEE International Conference on Computer Vision*, pages 2960–2967. IEEE, 2013.

[3] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.

[4] B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 222–230, 2013.

[5] B. Gong, K. Grauman, and F. Sha. Reshaping visual datasets for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1286–1294, 2013.

[6] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073, 2012.

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

[8] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015.

[9] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016.

[10] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. In *International Jont Conference on Artifical Intelligence*, IJCAI'09, pages 1187–1192, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[11] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *IEEE European Conference on Computer Vision*, pages 213–226, 2010.

[12] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *AAAI Conference on Artificial Intelligence*, 2015.

[13] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars. A deeper look at dataset bias. *CoRR*, abs/1505.01257, 2015.

[14] T. Tommasi and T. Tuytelaars. A testbed for cross-dataset analysis. In *IEEE European Conference on Computer Vision Task-CV Workshop*, pages 18–31, 2014.

[15] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.

[16] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.