Inter-Semantic Domain Adversarial in Histopathological Images

Nicolas Dumas
Ummon HealthTech
Dijon, France

Valentin Derangère
Centre de Lutte contre le Cancer
Dijon, France

Laurent Arnould
Centre de Lutte contre le Cancer
Dijon, France

Sylvain Ladoire
Centre de Lutte contre le Cancer
Dijon, France

Louis-Oscar Morel
Ummon HealthTech
Dijon, France

Nathan Vinçon
Ummon HealthTech
Dijon, France
nathanvincon@ummonhealthtech.com

***Abstract*—In computer vision, data shift has proven to be a major barrier for safe and robust deep learning applications. In medical applications, histopathological images are often associated with data shift and they are hardly available. It is important to understand to what extent a model can be made robust against data shift using all available data. Here, we first show that domain adversarial methods can be very deleterious if they are wrongly used. We then use domain adversarial methods to transfer data shift invariance from one dataset to another dataset with different semantics and show that domain adversarial methods are efficient inter-semantically with similar performance than intra-semantical domain adversarial methods.**

I. Introduction

In computer vision, data shift (i.e. a shift in the data distribution) has proven to be a major barrier for safe and robust deep learning real-world applications, such as computer vision for autonomous vehicles [1] [2], pose estimation [3], medical image segmentation and classification [4] [5].

Among medical images, histopathological images are tissue sections stained and analyzed with a microscope by a pathologist to highlight features from the tissue related with diseases. These images are the gold standard for cancer diagnosis and still have a huge potential to improve clinical practices [6]. Some data shifts for histopathological images are known such as differences in the acquisition device parameters, differences in the staining and the multitude of parameters in the different steps of the histopathology slide preparation. However, risk may also come from a type of data shift that is not yet known.

Domain Adversarial (DA) [7] training has proven to be effective against data shifts notably in histopathological images [8]. However, classic DA training configuration requires a sample from the targeted source of data (e.g. data from a new laboratory). This is generally not a problem as it only requires to pick samples from the new environment and train with the DA without requiring to label these new data. However, in clinical application, it is generally not possible to fine-tune a posteriori, as it would require clinically validating the model again. But clinical applications need a proven robustness to satisfy regulatory requirements.

In light of this DA technique and the robustness requirement of medical applications, an important question is whether we can use data with different semantics (e.g. flowers images and vehicle images have different semantics, prostate cancer and lung cancer images have close but also different semantics) as DA data. For example, can we use a multi-source dataset of prostate cancer for DA training while running a lung cancer classification task. In order words, we question the transferability of the domain adversarial process across tasks and image semantics. Transferability of DA training would imply that any task could gain generalization from a large dataset not only through classical transfer learning but also through DA transfer.

In this paper, we first investigate to what extent DA methods are beneficial and whether it can be deleterious. We then investigate to what extent DA methods can be transferred across datasets of different semantics (inter-semantic domain adversarial). Our contribution is :

- We analyze the DA efficiency by describing 3 effects (Figure 4) : the **cost** (i.e. negative difference of accuracy with baseline due to DA training), the **degradation** (i.e. negative difference of accuracy with baseline due to data shift), and the **gain** (i.e. positive difference with baseline after estimation of cost and degradation due to the consistency between data shift and DA). We further combine these effects in a regression model (see Overview of our approach) and show that a misuse of artificial domain shift such as color shift can be deleterious.
- We test to what extent DA training can be effective when the main task datasets and domain adversarial datasets are of different semantics. We show that DA can be transferred inter-semantically and that a small intensity of shift is sufficient to prevent most of the performance degradation due to the data shift.

II. Background and Related Work

In this Section, we review methods that were described to increase robustness and generalization over data shift.

Among these methods, we find :

- Image augmentation [9] [10] : the technique is easy to implement but can hurt performances if augmentation method is inadapted. Image augmentation also does not increase training time (or few), and does not increase prediction time.
- Stain and brightness normalization [11] [12] [13] : the technique requires to infer a source and a target stain, then to transform the input image. It is easy to implement but can hurt performance if the stain inference step for source image is not robust enough, Ren at al. provided a solution using ensembles [14]. Stain normalization can provide precise information about the stainings that can be later used for quality controls.
- GANs such as Stain-to-Stain Translation [15], a pix2pix-based [16] method or Cycle-GAN [17]: they are efficient but complex, unstable and expensive techniques that increase prediction time.
- Domain Adversarial : a domain adaptation method where a DA branch is added after a feature extractor using a Gradient Reversal Layer, which prevents the top layer of the feature extractor from containing information about the domain considered irrelevant for the main task. DA can be seen as a disentanglement method of data shift over informative features. The DA training is directly targeting the model, therefore prediction time is not modified. Visual domain adaptation methods are reviewed in [18].

III. Material and Methods

A. Datasets

We used MNIST and Fashion-MNIST datasets, both composed of 60,000 images with dimensions 28x28 (Figure 6).

We used 2 datasets of histopathological images. The first is CAMELYON [19], composed of 327,680 color images with dimensions 96x96x3 extracted from histopathological analyses of lymph node sections. Each image is annotated with a binary label indicating the presence of metastatic tissue. The second dataset, which we call TissueNet, is the dataset from the TissueNet challenge [20]. This dataset is composed of more than 5,000 images of uterine cervical tissue from 18 medical centers across France. The images are labelled on 4 levels depending on the grade of the cancer as follows :

- 0 : benign
- 1 : low malignant potential
- 2 : high malignant potential
- 3 : invasive cancer

We used the RandomResizedCrop transformation of the torchvision python library to generate 60,000 images of dimensions 96x96x3.

B. Domain Adversarial

The proposed architecture is based on the one described in Ganin *et al.* It includes a deep feature extractor and a deep label predictor, which together constitute a standard feed-forward architecture like a CNN. Domain adaptation is achieved by adding a domain classifier connected to the feature extractor via a gradient reversal layer (Figure 1). By adding a domain classifier after the feature extractor architecture, we build the domain adversarial neural network (DANN). The domain classifier is trained with a mix of datasets from different domains (i.e. with different data shifts) labeled with their domain (Figure 3).

C. Data shift

We tested both destructive shift (e.g. blur or noise), and domain shift (e.g. color shift and luminosity shift). The former aim to provide robustness against degraded input that can drastically hurt performance in a scenario where the model was trained on a curated dataset with high-quality images. The latter aim to provide invariance, therefore robustness, against input of the same quality but with natural data shift.

1) Noise: The noise fonction is denoted by N_i for i over $[0, 12]$. Let Im be an image represented by a float matrix with values between 0 and 1. N_i is defined by :

$$N_i(Im) = clip(Im + i * R, 0, 1)$$

with R a random matrix of the same dimension as Im and following a uniform distribution on $[0, 1]$

2) Blur: Blur is done by convolving an image with a normalized uniform filter. The kernel is obtained by the following equation :

$$K = 1/(k * k) * M$$

with M a matrix of 1 of dimension k by k.

3) Color shift: The color shift is based on the stain normalization method of reinhard [21]. We convert the image in the LAB format, then we fit the mean and the standard deviation of the 3 channels.

$$F(Im) = (Im - avg(Im)) * (std(T)/std(Im)) + avg(T)$$

Im is the original image, T the target image , avg is the average function and std is the standard deviation. .
More precisely, we use this normalization process as a color shift according to a reference image (Figure 2).

D. Training

The default training configuration requires 4 datasets (only 2 if we don't run DA) :

- the train and test datasets as for main classification task
- the two datasets used to train the DA (generally there is a data shift between these two DA datasets)

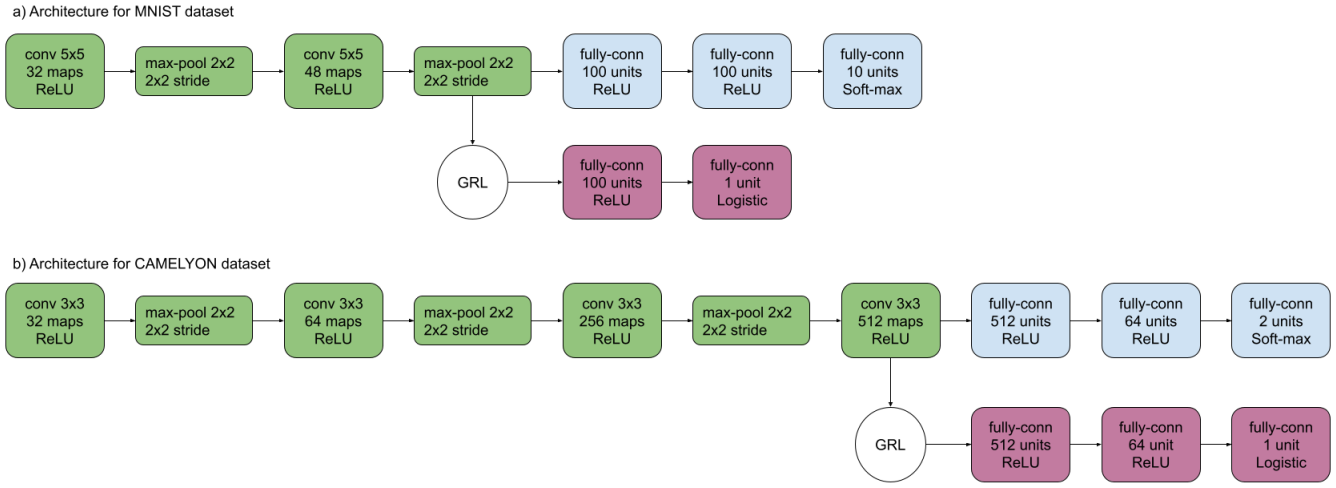

Fig. 1. Architecture of the CNN used in the experiments. Boxes correspond to the layers. Green boxes correspond to the feature extractor, blue boxes to the label predictor and red boxes to the domain classifier. GRL is the gradient reversal layer.

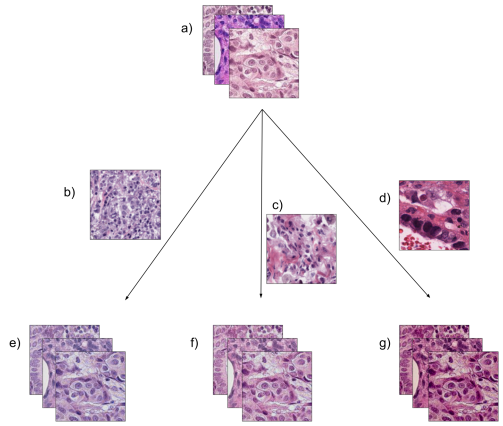

Fig. 2. Image of the CAMELYON dataset transformed by stain normalization in different domains. a) original dataset. e) f) and g) are the images of the CAMELYON dataset respectively normalized with b) c) and d) as target images.

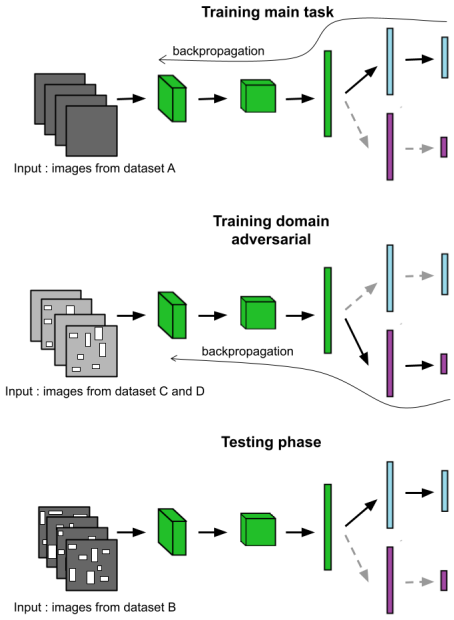

Fig. 3. Diagram of the training and testing of the model. Green boxes correspond to the feature extractor, blue boxes to the label predictor and red boxes to the domain classifier. The black arrows show the path taken by the data. The main task training and domain adversarial training are made during the same phase. Dataset A is the training dataset, dataset B is the testing dataset, datasets C and D are the DA training datasets mixed for the training.

Note that the semantic of datasets for the main classification task and for the DA learning can be different. The training of the classification task and training of the DA are simultaneous (Figure 3). Test shift refers to the data shift in the test dataset, and DA shift refers to the data shift between the two DA datasets.

E. Regression model

We used a regression model such that the data shift and DA effects are defined by 4 terms (Figure 3):

$$Accuracy = reference - degradation - cost + gain$$

- The reference is the performance of a raw CNN tested without data shift in the test set.
- The degradation term corresponds to the difference of accuracy with the reference due to a data shift in the test dataset.
- The cost term corresponds to the difference of accuracy

with the reference due to the DA training without any degradation.
- The gain term corresponds to the difference of accuracy with a model after applying degradation and cost to the reference, such that the previous equation is satisfied.

IV. Experiments and Analyses

A. Characterization of the effect of domain adversarial training over model performance against noise perturbation

We first characterized the influence of data shift and DA training over the model performance. We used a regression model to link model performance to the data shift and DA training. This model is composed of 4 terms, the reference, the degradation due to data shift, the cost due to DA and the gain due to DA after degradation and cost (see Materials and Methods).

Using MNIST dataset and noise as data shift, the reference model reached 0.98 accuracy. We found a sigmoidal relation between noise intensity and performance degradation and the accuracy falls to 0.20 (degradation of 0.78) with a noise intensity of 1.2. We found that DA training had no deleterious impact over the model accuracy, therefore DA has no cost. We model the bivariate DA gain function as the product of two univariate functions: a gaussian function dependent on the noise intensity, and a quadratic function dependent on the DA intensity (Figure 5).

Interestingly, gain depends heavily on the data shift intensity but is almost invariant to DA intensity, suggesting that most of the DA benefits can be reached at very low DA intensity, this is further discussed in the Discussion. The maximum gain of 0.50 accuracy is reached with an intermediate noise intensity (1.1) and a low DA intensity (0.5).

B. Inter-semantic transferability of the domain adversarial training in the MNIST dataset against noise perturbation

Similar tests were performed using Fashion-MNIST instead of MNIST for the DA training, noise has been applied as previously. In this configuration, the DA training runs inter-semantically.

Training using DA training on Fashion-MNIST shows results that are close to the previous experiment (Figure 7). We find a reference accuracy of 0.98 and a sigmoidal relation between noise intensity and performance degradation, and a maximum degradation of 0.78 with a maximum noise of 1.2. We find no cost associated with DA training. The gain is also the same as the configuration with MNIST in the DA datasets, modeled by the product of a gaussian function and a quadratic function, and mainly sensitive to noise intensity. The maximum gain of 0.45 accuracy is reached with an intermediate noise intensity as test shift (0.9) and an intermediate noise intensity as DA shift (0.9).

This experiment shows that using DA inter-semantically shows similar DA gain compared to intra-semantical DA. It suggests that the DA training for noise is entirely transferable between MNIST and Fashion-MNIST, such that Fashion-MNIST can be used equivalently to MNIST for the DA training in a MNIST classification task.

We performed similar tests using black images (Figure 6) instead of the Fashion-MNIST dataset as DA datasets and showed it has significant DA gain of 0.27. However, the gain is smaller in this experiment compared to previous experiments. Therefore the DA training is partially transferable from black images to MNIST.

C. Characterization of the effect of domain adversarial against blur perturbation

In a second step we replaced noise by blur as data shift. Blur is a perturbation that often occurs in digitized histopathological samples. Here, we use MNIST for the train and test datasets, and also MNIST for the DA. The reference accuracy is 0.98 and there is no DA cost. There is a sigmoidal relation between blur intensity and performance degradation and accuracy falls to a minimum accuracy of 0.23 (degradation of 0.75) for a kernel size of 9. Once more, the gain is mainly sensitive to the blur intensity as a Gaussian function. The maximum gain is 0.25 for a kernel size of 7 in the dataset test and of 3 for the DA (Figure 7).

We then replaced the MNIST dataset by a Fashion-MNIST dataset as DA datasets. We still have a degradation that can be modeled by a sigmoid function, we find a constant cost equal to zero and an important gain of 0.26. With similar results using Fashion-MNIST in the DA rather than MNIST, we showed that the inter-semantical DA transferability is applicable also with another data shift than noise, here blur.

D. Characterization of the effect of domain adversarial training on histopathological datasets against blur perturbation and color shift

We next used the CAMELYON dataset for both the main task and the DA and used blur as data shift as it is often found as natural perturbation in histopathological images. The reference accuracy is 0.85 and DA has a significant cost of 0.05 while there was no significant gain. This shows that DA can be deleterious depending on the processed dataset, and that the gain of DA is not ubiquitous. A possible explanation could be that blur erases important information from the histopathological data, therefore the model cannot recover from this kind of data degradation. Maximum blur with kernel size of 9x9 showed a degradation of 0.18 making accuracy fall to 0.67 (Figure 8).

We next studied color shift, another common shift from histopathological images, as data shift with the CAMELYON dataset. As color shift depends on more than one parameter, it is difficult to introduce a consistent intensity for a color shift.

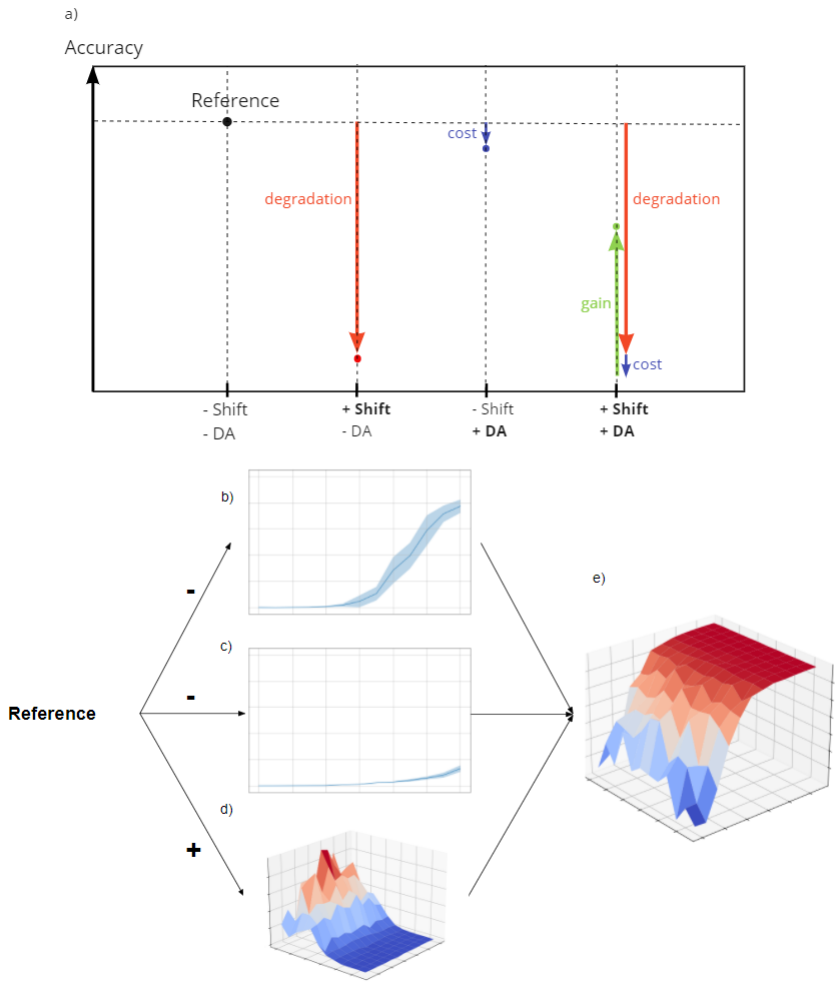

Fig. 4. a) Diagram of the 4 pipeline configurations for calculating the reference, degradation, cost and gain. In the configuration with test shift and DA shift, the gain is obtained by the equation

$$gain = accuracy - (reference - (degradation + cost))$$

b) Example plot of the degradation according to the test shift intensity c) Example plot of the cost according to the DA shift intensity d) Example plot of the gain according to both test shift intensity and DA shift intensity. The x-axis is the test shift intensity, the y-axis is the score, the z-axis is the DA shift intensity e) Example plot of the accuracy according to both test shift intensity and DA shift intensity. The x-axis is the test shift intensity, the y-axis is the score, the z-axis is the DA shift intensity.

Thus, we create datasets of different domains of color shift that show different degradation by normalizing the dataset according to different reference images (Figure 2).

The reference has an accuracy of 0.85. The degradation varies from 0.03 to 0.26 depending on the domain shift. The cost varies from 0.0 to 0.10 with a small correlation with the amplitude of the degradation. As the cost is not null, it is important to carefully design the DA architecture and training process in order to prevent a loss of accuracy for the main task. The gain varies from 0.04 to 0.28, the gain is well correlated with the degradation. This is intuitive because the more performance degrades, the more DA training can be helpful.

E. Characterization and inter-semantic transferability of the domain adversarial training on histopathological datasets against color shift

We next used the TissueNet dataset for the DA training instead of the CAMELYON dataset, while keeping CAMELYON dataset for the main classification task and using color shift as data shift. This configuration is similar to the previous experiment using MNIST for the main task and Fashion-MNIST for the DA (Figure 9). However, there is no notion of data shift intensity for color shift because color shift is multidimensional.

In this configuration, the reference accuracy is 0.82. The degradation varies from 0.03 to 0.26 depending on the domain shift. The cost varies from 0.0 to 0.16 with a

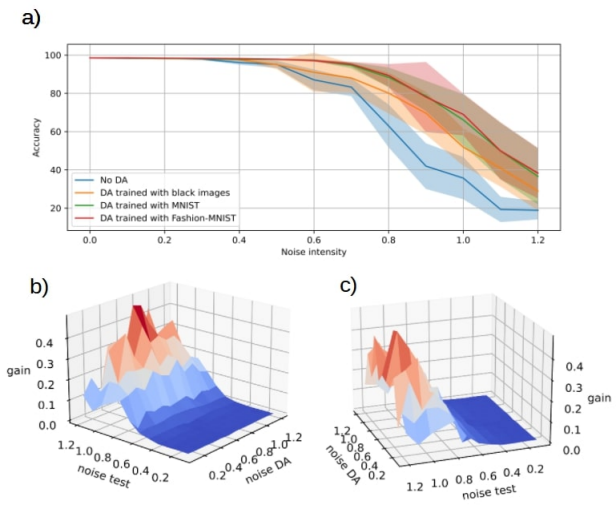

Fig. 5. a) Plot of the accuracy for image classification of MNIST dataset as a function of noise intensity. For each curve, DA is trained with a different dataset. The noise intensity in DA is the same as in the test dataset. b) Representation of the gain with DA trained with MNIST dataset and noise as data shift. c) Same image as b) with different orientation.

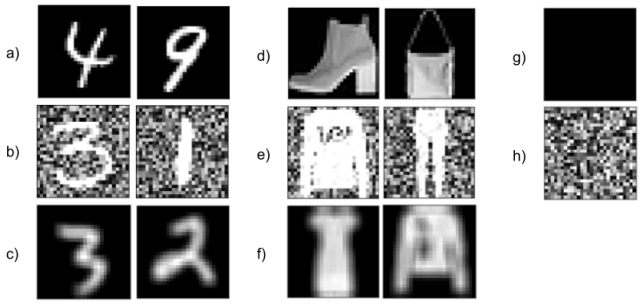

Fig. 6. Images extracted from the different datasets used a) MNIST b) Noised MNIST c) Blurred MNIST d) Fashion-MNIST e) Noised Fashion-MNIST f) Blurred Fashion-MNIST g) Black image h) Noised black image

small correlation with the amplitude of the degradation. The gain varies from 0.02 to 0.28, the gain is well correlated with the degradation. Values of cost, degradation and gain are very close to the previous intra-semantic configuration, showing that DA training can be efficiently transferred inter-semantically in real histopathological images. Further investigations will be needed in order to understand if DA training can be applied to histopathological data with a gain greater than a cost universally, therefore increasing robustness of the model.

V. Discussion and Conclusion

Histopathological data is highly heterogeneous due to the diversity of acquisition devices and the lack of standard, while histopathological data is hardly available because many regulatory requirements are necessary to get access to clinical data. Together, lack of availability and heterogeneity are a major barrier for the development of safe and robust

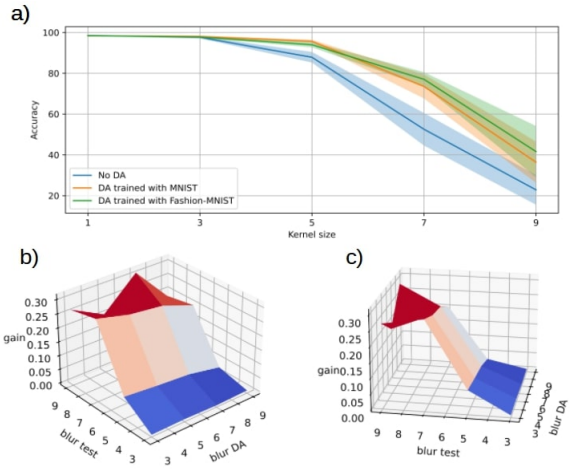

Fig. 7. a) Representation of accuracy for image classification of MNIST dataset based on blur intensity. For each curve DA is trained with a different dataset. The blur intensity in DA is the same as in the test dataset. b) Representation of the gain with DA trained with Fashion-MNIST and blur as data shift. c) Same image as b) with different orientation.

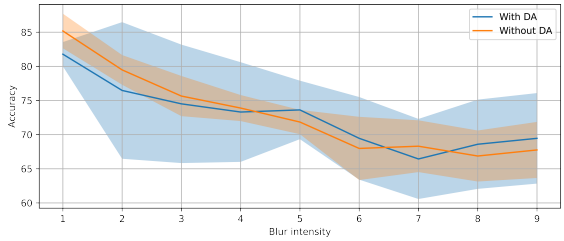

Fig. 8. Comparison of the accuracy with and without DA training according to blur intensity as data shift on CAMELYON dataset. DA datasets are also CAMELYON datasets.

models. Therefore, we developed here a strategy to increase robustness using all available data diversity using domain adversarial methods.

By systematic analysis of DA effect over the model performance, we found that when DA is efficient, a low intensity in DA shift is sufficient to provide most of the possible gain from DA. But DA is not always efficient, blur degradation on histopathological datasets could not be retrieved using DA methods, this could be explained by two reasons: first, blur is already present in the original CAMELYON dataset therefore DA has no effect, and second, blur is a destructive noise which quickly makes classification impossible because relevant information may be found in high resolution patterns. Finally, inter-semantic DA transferability is an efficient strategy as it works using different dataset and with non-real data as the DA with black images showed an effective performance improvement.

In conclusion, DA training is transferable inter-semantically and the robustness of clinical algorithms can be increased by taking advantage of the heterogeneity of available datasets,

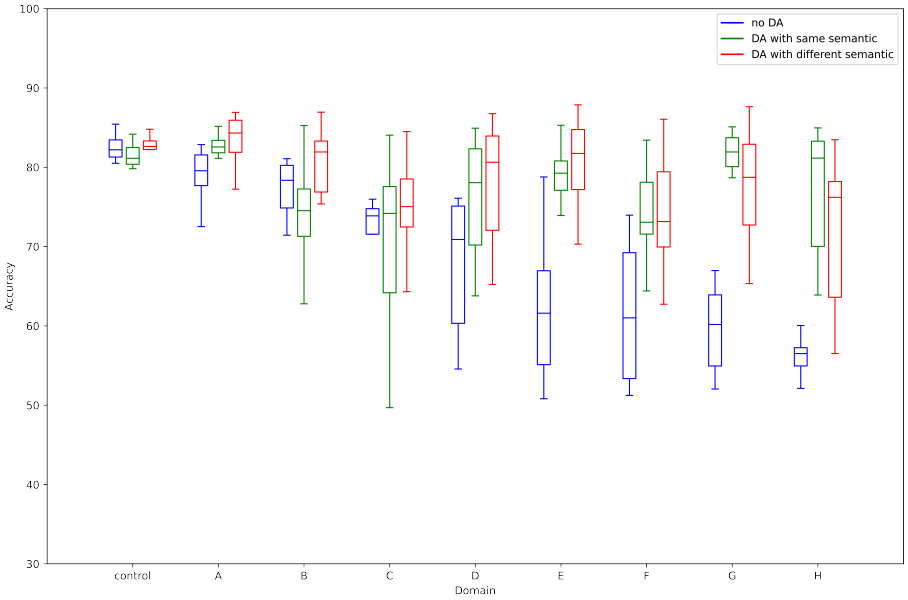

Fig. 9. Box plot of the accuracy in different domains, the blue boxes are the accuracy without using the DA, the green boxes are the accuracy using a DA with the same semantics and the red boxes are the accuracy using a DA with different semantics. The datasets used for the DA training are normalized in the same domains as the test datasets.

whatever their semantic content is. Further investigation will be needed to understand when DA training is beneficial and when it is deleterious. Another remaining question is whether DA training can be done with inner inter-semantic datasets (data of different domains are also of different semantic). In this configuration, DA might erase features that are relevant for the main task. However, use of DA should be careful as it can significantly and negatively affect model performance.

REFERENCES

[1] Angelos Filos, Panagiotis Tigkas, Rowan Mcallister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. Can Autonomous Vehicles Identify, Recover From, and Adapt to Distribution Shifts? In *Proceedings of the 37th International Conference on Machine Learning*, pages 3145–3153. PMLR, November 2020. ISSN: 2640-3498.

[2] Emanuele Alberti, Antonio Tavera, Carlo Masone, and Barbara Caputo. IDDA: a large-scale multi-domain dataset for autonomous driving. *IEEE Robotics and Automation Letters*, 5(4):5526–5533, October 2020. arXiv: 2004.08298.

[3] Shaobo Zhang, Wanqing Zhao, Ziyu Guan, Xianlin Peng, and Jinye Peng. Keypoint-Graph-Driven Learning Framework for Object Pose Estimation. pages 1065–1073, 2021.

[4] Qinming Zhang, Luyan Liu, Kai Ma, Cheng Zhuo, and Yefeng Zheng. Cross-denoising Network against Corrupted Labels in Medical Image Segmentation with Domain Shift. June 2020.

[5] Eduardo H. P. Pooch, Pedro L. Ballester, and Rodrigo C. Barros. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. September 2019.

[6] Elodie Pronier, Benoît Schmauch, Alberto Romagnoni, Charlie Saillard, Pascale Maillé, Julien Calderaro, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, Thomas Clozel, Matahi Moarii, Pierre Courtiol, and Gilles Wainrib. Abstract 2105: HE2RNA: A deep learning model for transcriptomic learning from digital pathology. *Cancer Research*, 80(16 Supplement):2105–2105, August 2020. Publisher: American Association for Cancer Research Section: Clinical Research (Excluding Clinical Trials).

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *arXiv:1505.07818 [cs, stat]*, May 2016. arXiv: 1505.07818.

[8] Maxime W. Lafarge, Josien P. W. Pluim, Koen A. J. Eppenhof, Pim Moeskops, and Mitko Veta. Domain-adversarial neural networks to address the appearance variability of histopathology images. *arXiv:1707.06183 [cs]*, 10553, 2017. arXiv: 1707.06183.

[9] Khrystyna Faryna. Tailoring automated data augmentation to H&E-stained histopathology. page 11.

[10] David Tellez, Geert Litjens, Peter Bandi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in

convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, December 2019. arXiv: 1902.06543.

[11] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110, Boston, MA, USA, June 2009. IEEE.

[12] Adnan Mujahid Khan, Nasir Rajpoot, Darren Treanor, and Derek Magee. A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Color Deconvolution. *IEEE Transactions on Biomedical Engineering*, 61(6):1729–1738, June 2014. Conference Name: IEEE Transactions on Biomedical Engineering.

[13] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. *IEEE Transactions on Medical Imaging*, 35(8):1962–1971, August 2016. Conference Name: IEEE Transactions on Medical Imaging.

[14] Jian Ren, Ilker Hacihaliloglu, Eric A. Singer, David J. Foran, and Xin Qi. Unsupervised Domain Adaptation for Classification of Histopathology Whole-Slide Images. *Frontiers in Bioengineering and Biotechnology*, 7:102, 2019.

[15] Pegah Salehi and Abdolah Chalechale. Pix2Pix-based Stain-to-Stain Translation: A Solution for Robust Stain Normalization in Histopathology Images Analysis. In *2020 International Conference on Machine Vision and Image Processing (MVIP)*, pages 1–7, Iran, February 2020. IEEE.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, Honolulu, HI, July 2017. IEEE.

[17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, Venice, October 2017. IEEE.

[18] Mei Wang and Weihong Deng. Deep Visual Domain Adaptation: A Survey. *Neurocomputing*, 312:135–153, October 2018. arXiv: 1802.03601.

[19] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, Quirine F Manson, Nikolas Stathonikos, Alexi Baidoshvili, Paul van Diest, Carla Wauters, Marcory van Dijk, and Jeroen van der Laak. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*, 7(6), June 2018.

[20] DrivenData. TissueNet: Detect Lesions in Cervical Biopsies.

[21] Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. Color Transfer between Images. *IEEE Computer Graphics and Applications*, page 8, 2001.