Available online at www.eccomasproceedia.org
Eccomas Proceedia UNCECOMP (2025) 597-604

ECCOMAS
Proceedia

UNCECOMP 2025
6th ECCOMAS Thematic Conference on
Uncertainty Quantification in Computational Science and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou (eds.)
Rhodes Island, Greece, 15-18 June 2025

DOMAIN ADAPTATION IN STRUCTURAL HEALTH MONITORING USING LEARNING BOUNDS AND GENERATIVE NETWORKS

C.A. Lindley[1], N. Dervilis[1], and K. Worden[1]

[1]University of Sheffiel
School of MAC, Dynamics Research Group
Mappin St, Sheffiel S1 3JD
{c.a.lindley,n.dervilis,k.worden}@sheffield.ac.u

Abstract. *The use of statistical models in engineering can be limited by the availability of data. This issue is particularly prevalent in structural health monitoring, where labelled data can be expensive to obtain since it requires data from most, if not all, operational conditions that a structure might experience. As a result, there has been a surge of interest in transfer learning within the research community, aiming to use existing data to improve the learning on new structures that have yet to experience previously-unseen conditions. The approach taken in this study is based on statistical learning theory to estimate an upper bound on the expected risk of a transfer learner. An adversarial-type neural network architecture is then designed to minimise this bound, producing a latent representation that aligns different domains and, in turn, improves learning in systems where labelled data are scarce. The outcome of this approach is a novel generative domain-adversarial neural network that enables the estimation of uncertainty in predictions. A simple simulated case study is presented here to demonstrate the effectiveness of the proposed model in addressing challenges related to unsupervised domain adaptation.*

Keywords: Domain Adaptation, Generalisation Bounds, Uncertainty Quantification Multiscale Analysis, Structural Health Monitoring.

ISSN:2623-3339 © 2025 The Authors. Published by Eccomas Proceedia.
Peer-review under responsibility of the organizing committee of UNCECOMP 2025.
doi: 10.7712/120225.12380.21173

1 INTRODUCTION

The incorporation of statistical methods into Structural Health Monitoring (SHM) is a subject of increasing interest in the literature. The success of these methods partly depends on the amount of labelled data, which may, unfortunately, not often be available in SHM. This shortcoming is because structures may exhibit unprecedented operational conditions that are not covered by the existing data. In theory, if one could collect data for every possible health state a structure may experience during its lifespan, the cost would be unfeasibly high, since this scenario would mean building many copies of the structure to capture all conditions. It is thus important to account for this limitation when employing statistical methods in SHM.

Nevertheless, progress in the field has shown that machine learning and statistical methods can be successfully used for damage detection, localisation, classification and prognosis [1]. These advances offer the prospect of a framework that could work reliably in practice. One promising approach pertains to a branch of machine learning known as *transfer learning* [2]. The main idea is simple; that is, to improve the performance of an intelligent system by using existing data from external sources. In other words, by including health states already observed in other systems, it may be possibly to predict the current health state of the system in question, even when facing a condition that has not been seen before.

To illustrate the motivation for this paper, the following framework is constructed. Let a structure S be modelled by a mathematical function M which attempts to predict the behaviour of S. Furthermore, the model may be expected to account for certain aspects of the behaviour of S, referred to here as the *context* C of S. These definitions follow those given in [3], whereby M is said to be an ϵ-Mirror of S, for a given context C, if and only if,

$$d^C(M^C(t), r^C(t)) \leq \epsilon \tag{1}$$

where $d^C(t)$ denotes some metric, and $r^C(t)$ denotes the response of the system to a time series set $\{t_i\}_{i=0}^{T}$ for the context C. Suppose the model M^C has been validated and shown to be an ϵ-mirror of S^C. The question then is: *for a new structure S', behaving in a context given by small changes ΔS and ΔM, is the corresponding M' a mirror of S'^C for some value of ϵ'?*

Although the model proposed here aims to address the problem of scarce labelled datasets, it is not its main motivation. The benefits of doing so appear when trying to answer the question above. Interestingly, the learning bounds found in *statistical learning theory* [4] can help validate M' without needing to build a prototype S'. As a result, the method presented in this paper is fundamentally based on these learning bounds to develop an algorithm that addresses the original problem of limited datasets.

The proposed algorithm relies on a novel adversarial neural network designed to minimise an upper bound on the expected error for the target domain. This approach is achieved by combining some available (unlabelled) data with (labelled) data deriving from an external source. The neural network learns a reduced latent space representation of the data, where the different datasets are aligned, enabling accurate classification across domains.

2 THEORY

2.1 The problem of transfer learning

The transfer learning problem is formalised here by establishing the following definitions

1. *A domain $\mathcal{D}$, consisting of a marginal probability distribution p on a feature space $\mathcal{X}$. Therefore, $\mathcal{D} = \{\mathcal{X}, p(X)\}$, where $X = \{\mathrm{x}_i\}_{i=1}^{n} \in \mathcal{X}$ is a finite sample set from $\mathcal{X}$.*

2. *A task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$, for a given domain $\mathcal{D}$, defining an output space $\mathcal{Y}$ and a labelling function $f : \mathcal{X} \rightarrow \mathcal{Y}$.*

Two distinct domains are then established: (1) a *source domain* $\mathcal{D}_S$, and (2) a *target domain* $\mathcal{D}_T$, each with their corresponding tasks $\mathcal{T}_S$ and $\mathcal{T}_T$. Therefore, the idea of transfer learning is to utilise knowledge about the source $\langle \mathcal{D}_S, \mathcal{T}_S \rangle$ to improve the learning of a target predictive function $f_t(\cdot)$ for a given target $\langle \mathcal{D}_T, \mathcal{T}_T \rangle$, where $\mathcal{D}_s \neq \mathcal{D}_T$ and $\mathcal{T}_S \neq \mathcal{T}_T$. In the following sections, this problem is reduced to the case in which no labelled data are available in the target domain. The learning problem under consideration is thus one of *domain adaptation*, which may be thought of a branch of transfer learning whereby it is assumed that $\mathcal{X}_s = \mathcal{X}_t$ and $\mathcal{Y}_s = \mathcal{Y}_t$, but that $p_s(X, Y) \neq p_t(X, Y)$, with $Y = \{\mathrm{y}_i\}_{i=1}^{n} \in \mathcal{Y}$.

2.2 Derivation of generalisation bound on target expected risk

Given a finite training set $\mathrm{z} = \{(\mathbf{x}_i, \mathrm{y}_i)\}_{i=1}^{n}$, a supervisor is introduced to learn some *hypothesis* $h : \mathcal{X} \rightarrow \mathcal{Y}$, aimed at approximating $f(\cdot)$. The probability that the hypothesis disagrees with the true labelling function is defined by the *expected risk*,

$$R(\alpha) = \int \mathcal{L}(\mathrm{y}, h(\mathrm{x}, \alpha))p(\mathrm{z})\mathrm{dz}, \quad \alpha \in \Lambda \tag{2}$$

which is defined by an expectation of some *loss* functional $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$, computed with respect to the (unknown) joint distribution $p(\mathrm{z})$. In this expression, the predictive function is assumed to be a conditional distribution $p(\mathrm{y}|\mathrm{x})$. One should note that the set Λ to which α belongs can be a set of scalar quantities, vectors, or of abstract elements [5]; that is, the function space is not limited to parametric models. This expression holds under the assumption that the data are generated from a unique underlying probability distribution. In the context of transfer learning, the expected risk of this form can only hold true for one domain; therefore, when referring to the risk of a hypothesis in the source domain, equation (2) can be rewritten as,

$$R_s(\alpha) = \int \mathcal{L}(\mathrm{y}, h(\mathrm{x}, \alpha))p_s(\mathrm{z})\mathrm{dz}, \quad \alpha \in \Lambda \tag{3}$$

where the expectation is now explicitly evaluated with respect to source-domain elements. Similarly, the parallel notations R_t and p_t are used for the target domain. While the expected risk gives some insights into how a supervisor learns to generalise from data, the integral cannot be evaluated analytically without knowing the underlying generative distribution. Nonetheless, a way in which one can make progress is to upper bound expected risk by employing an *Empirical Risk Minimisation* inductive principle [4]. In this case, however, the target risk is upper bounded by using Jensen's Inequality, which is defined as follows for a concave function $g(\cdot)$,

$$\log \mathbb{E}_{p(z)}\{g(z)\} \geq \mathbb{E}_{p(z)}\{\log g(z)\} \tag{4}$$

The upper bound on the target risk can then be derived as follows,

$$\begin{aligned} \log R_s &\geq \int p_t(\mathrm{z}) \log \frac{\mathcal{L}(\mathrm{y}, h(\mathrm{x}))p_s(\mathrm{z})}{p_t(\mathrm{z})}\mathrm{dz} \\ &= \int p_t(\mathrm{z}) \log \mathcal{L}(\mathrm{y}, h(\mathrm{x}))\mathrm{dz} + \int p_t(\mathrm{z}) \log \frac{p_s(\mathrm{z})}{p_t(\mathrm{z})}\mathrm{dz} \\ &= \mathbb{E}_{p_t(\mathrm{z})}\{\log \mathcal{L}(\mathrm{y}, h(\mathrm{x}))\} - \mathrm{KL}[p_t(\mathrm{z})||p_s(\mathrm{z})] \end{aligned} \tag{5}$$

The second term in the right-hand side is the negative of the Kullback-Leibler (KL) divergence between the target joint distribution and source joint distribution. Rearranging the expression yields an upper bound on the target risk,

$$\mathbb{E}_{p_t(\mathrm{z})}\{\log \mathcal{L}(\mathrm{y}, h(\mathrm{x}))\} \leq \log \mathbb{E}_{p_s(\mathrm{z})}\{\mathcal{L}(\mathrm{y}, h(\mathrm{x}))\} + \mathrm{KL}[p_t(\mathrm{z})||p_s(\mathrm{z})] \tag{6}$$

The full derivation of the upper bound is defined as,

$$\begin{aligned}\mathbb{E}_{p_t(\mathrm{x},\mathrm{y})}\{\log \mathcal{L}(h(\mathrm{x}), \mathrm{y})\} \leq \log \mathbb{E}_{p_s(\mathrm{x},\mathrm{y})}\{\mathcal{L}(h(\mathrm{x}), \mathrm{y})\} + \mathbb{E}_{p_t(\mathrm{x})}\mathbb{E}_{p_t(\mathrm{y}|\mathrm{x})}\{\log p_t(\mathrm{y}|\mathrm{x}) - \log p_s(\mathrm{y}|\mathrm{x})\} \\ + \mathrm{KL}[p_t(\mathrm{x})||p_s(\mathrm{x})]\end{aligned} \tag{7}$$

or equivalently, in an unsupervised framework,

$$R_t(\hat{\mathcal{L}}) \leq \log R_s(\mathcal{L}) + \mathrm{KL}[p_t(\mathrm{x})||p_s(\mathrm{x})] \tag{8}$$

where $\hat{\mathcal{L}} = \log \mathcal{L}$. The second term in (7) is ignored when the set of labelled target data is unavailable. Removing this term from the expression holds under the assumption that conditionals are equal across domains, and reduces to zero when $\mathrm{KL}[p_t(\mathrm{x})||p_s(\mathrm{x})]$ approaches zero. Therefore, by this assumption, minimising the right-hand side of (8) implies minimising $\mathbb{E}_{p_t(\mathrm{x})}\mathbb{E}_{p_t(\mathrm{y}|\mathrm{x})}\{\log p_t(\mathrm{y}|\mathrm{x}) - \log p_s(\mathrm{y}|\mathrm{x})\}$. It must be noted, however, that such an allowance can promote the likelihood of *negative transfer*.

2.3 Interpretation

The terms in equation (7) are readily interpretable. The first term in the upper bound represents the log of the source expected risk. The second term measures the difference between the tasks across domains. Similarly, the third term calculates the divergence between the input marginals of the domain. In other words, for a given hypothesis h, this expression estimates the expected risk on the target domain by including elements from the source domain. One can thus leverage on this upper bound to determine how effective an algorithm is at improving the transfer of knowledge across domains. For example, if the aim is to use data from the source domain to improve the generalisation performance in the target domain, some algorithm could be devised with the objective to minimise the terms in equation (7).

Unfortunately, these terms cannot be (rigorously) computed, for two main reasons:

1. The target conditional $p_t(\mathrm{y}|\mathrm{x})$ is unlikely to be known.

2. The marginal distributions $p_s(\mathrm{x})$ and $p_t(\mathrm{x})$ are only possible to estimate either by having an infinite amount of data or knowledge about the distributions.

Nevertheless, an attempt to approximate the upper bound with finite samples may offer solutions that could be useful from a more pragmatic point-of-view. In fact, it is often the case in which statistical algorithms are employed without explicitly acknowledging converging guarantees. A paradigmatic example that may be of relevance here relates to *Variational Auto-Encoders* (VAE) [6]. In fact, the terms in (8) resemble those used in the objective function of a VAE, with the main difference being that VAEs use the KL divergence to regularise the inference of some posterior distribution with respect to a corresponding prior distribution in a Bayesian framework. While it is emphasised that a non-rigorous handling of this problem will require some form of empirical validation, if sufficient data are available, the result of minimising (8) may still provide a valuable outcome in an engineering application.

2.4 Towards a neural network approach

Looking closely into the parallels such an approach would have with a VAE, it is similarly proposed here to have a deep neural network minimise the RHS of (8). There is, however, a prevalent limitation preventing a conventional VAE from taking over such an objective. In particular, the objective-function formulation of VAEs is often derived by assuming an isotropic Gaussian prior distribution, which simplifies the KL divergence metric to a tractable closed-form solution (as long as the likelihood is also assumed Gaussian). It is desired to relax these assumptions for the domain adaptation problem, since the shape of both the source marginal $p_s(\mathrm{x})$, and target marginal $p_t(\mathrm{x})$, are unlikely to be known in advance.

This limitation has been acknowledged in the literature, and adversarial networks have been proposed as an alternative to improve the flexibility of VAEs when inferring complex latent distributions. Of particular interest here is the approach proposed by Meschder et al. [7]. Their model preserves the essence of a VAE while incorporating a discriminative network trained to distinguish samples drawn either from the inferred posterior or from the prior distribution. The encoder is then tuned to produce samples aimed at fooling the discriminator without compromising the likelihood maximisation, and thus produce samples that resemble those from the prior distribution, regardless of their shape.

By adopting the network framework used in [7], the terms in (8) can be similarly minismised. The difference here, however, is that the marginals are transformed to be aligned in a reduced latent space, rather than inferring a posterior distribution that assimilates to a given prior distribution.

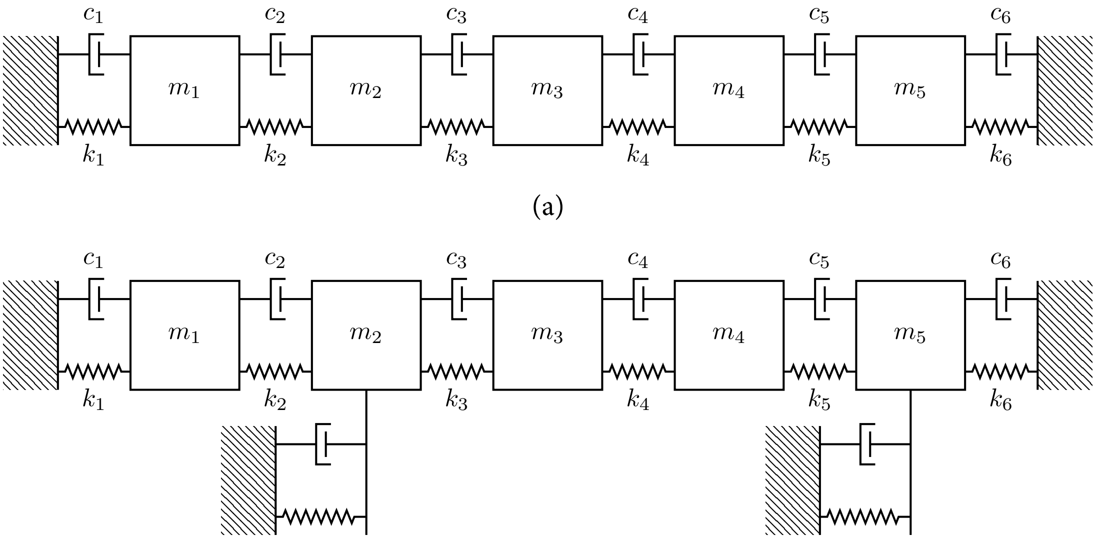
3 CASE STUDY

The current demonstration involved simulating an impulse response of some structure S, modelled as a lumped-mass model consisting of five masses $\{m_i\}_{i=1}^{5} = 1$, with spring coefficient $\{k_i\}_{i=1}^{5} = 10^5$, and damping coefficient $\{c_i\}_{i=1}^{5} = 20$. Three different states were considered in the simulation; corresponding to one healthy (normal) operating condition, and two different damaged conditions. In this case, one of the damaged states was simulated by degrading the stiffness of spring k_3 to 75% of its original value. Similarly, the remaining damaged state was simulated by degrading the stiffness of spring k_5 by the same amount. The features of interest were the damped natural frequencies at each health state, which were extracted from the system parameters via eigendecomposition. Finally, different sequences of n i.i.d. Gaussian samples were added to each of the natural frequencies to introduce noise corresponding to 1% RMS of the damped-natural frequencies. These quantities thus comprise the source features used for analysis, $X_s \in \mathbb{R}^{n \times 5}$.

In order to generate the target data, a modified structure S' was also constructed by including additional connections to the original lumped-mass model of S. Concretely, the connections here were included to masses m_2 and m_5. The models are shown in Figure 1a and Figure 1b, respectively. Following the same steps as before, a healthy state and two damaged states were then simulated for S'.

The premise of the transfer task being pursued here is to make use of the data gathered from S, to improve the predictive ability of the learner when presented with data from S'. To illustrate the benefit of employing the algorithm outlined above, a classifier was first trained only using source domain data; or similarly, by training the neural network without the adversarial discriminator.

Therefore, for the first part of this demonstration, the main architecture of the neural network



(a)

(b)

Figure 1: Lumped mass-damper-spring models: (a) Structure S and (b) Structure S'

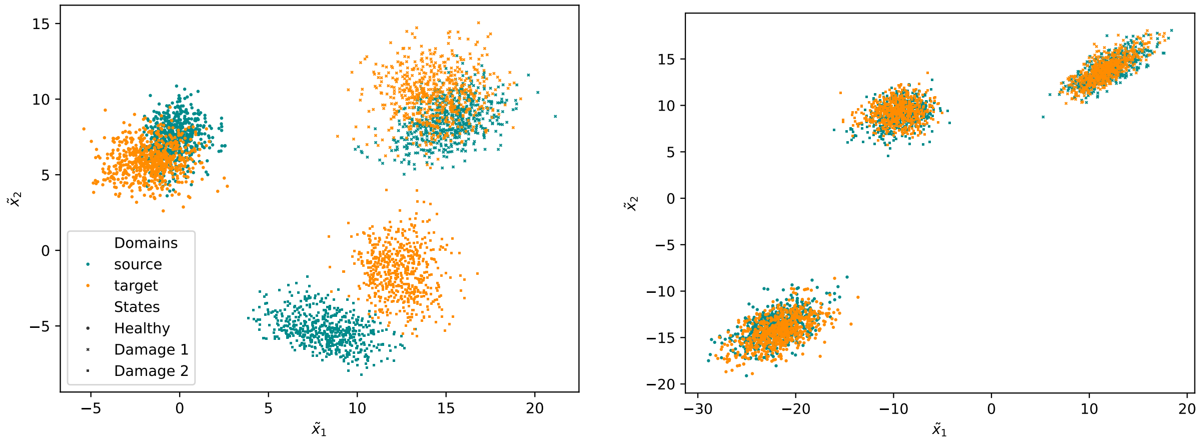
consisted of multi-layer perceptrons used to model the feature extractor g_ϕ and classifier h_θ. The feature extractor was designed with three hidden layers, each containing 128 nodes with *elu* activation functions, followed by a two-node linear output layer. The outputs from the feature extractor were then fed into the classifier, which also had three hidden layers of 128 nodes, each using *elu* activations. Since this task involved classification a *softmax* output layer with three nodes was used-one for the healthy class and two for the damage classes, labelled as damage class 1 and damage class 2.

The source data deriving from structure S were split into a training set and a validation set. A simple preprocessing step was conducted in which the empirical mean was subtracted from the data. Additionally, the set of weights $\{\phi\}$ and $\{\theta\}$ were initialised with a *He standard* initialiser [8], and $L2$-regularised to mitigate the network from overfitting the training data. An *early-stopper* was also enabled for this purpose.

Although this exercise may seem trivial, the problem took a more interesting turn when feeding in data from the modified structure S'. On can note in Figure 2a the discrepancy that exists between source and target representations, which it is not all that surprising given that the feature extractor learnt from source data alone.

This approach required the addition of a discriminator to the main architecture of the neural network. Here, the discriminative network also consisted of a multi-layer perceptron, which was designed with five hidden layers, each containing 256 nodes and *elu* activations. A *sigmoid* output layer with a single node was used to determine the probability of a sample originating from the target domain. As before, the same preprocessing and regularisation measures were enforced to mitigate the possibility of overfitting

Having trained the complete network, the entire validation set was fed to the feature extractor. Visually, as shown in Figure 2b, the latent-space representations of the domain marginals appear to have aligned nicely. With an accurate class alignment, the classifying network is able to correctly assign labels to both the source data and the target data. Therefore, the model can be said to have successfully transferred its ability to make predictions across domains.



(a)

(b)

Figure 2: Latent-space representation: (a) pre-transfer and (b) post-transfer

4 DISCUSSION AND CONCLUSIONS

Despite the promising results shown in the previous section, there are a few considerations worth noting. The first and foremost, is the fact that the upper bound on the target expected risk defined in (8) is not rigorously computed by the neural network. Instead, the conducted strategy computes some empirical estimation of the upper bound in disregard of the model complexity when presented with finite amounts of data. In other words, the risk estimate is inaccurate, and therefore, no real guarantee can be made in terms of how well the model can generalise to new data. Nevertheless, this consideration is commonly overlooked in practice, particularly when adopting deep learning models to solve learning problems. The generalisation problem is rather addressed implicitly by assuming that enough data are available to ensure some reasonable degree of generalisation, in addition to regularisation measures that are often used in a more heuristic manner.

Another important consideration worth highlighting is the challenging validation process involved in tuning the hyperparameters of the network. Designing the architecture required deciding on the number of hidden layers, number of nodes and activation functions among other relevant factors. In this case, these were manually tuned to yield good-enough results for the sake of the demonstration. The appropriate model validation will be pursued in future work with more stringent strategies for optimal hyperparameter selection.

The benefits of this model over other methods for unsupervised domain adaptation will also be explored via appropriate benchmarking tests. Current state-of-the-art models should be examined to determine the conditions under which it is worth preferring the current model. For example, the unsupervised framework presented above resembles that of the *domain-adversarial neural network* (DANN) [9], developed from the upper bound estimates on the target expected risk proposed by Ben David et al. [10]. An immediate distinction that can be made at this stage, however, is that the current version can be somewhat interpreted as a generative version of the DANN, and thus account for uncertainty in the predictions. Having a generative model in this context may benefit regression-type problems whereby the prediction is of some dynamic response across domains. This exercise might also be worthy of further investigation.

Finally, further developments will be investigated by adopting a semi-supervised setting via the incorporation of the second term in (7).

5 ACKNOWLEDGMENTS

The authors of this paper gratefully acknowledge the support of the UK Alan Turing Institute via funding of the Turing Research and Innovation Cluster on Digital Twins. For the purpose of open access, the authors has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

REFERENCES

[1] C. Farrar and K. Worden, *Structural Health Monitoring: A Machine Learning perspective*. Wiley, 2013.

[2] K. Weiss and T.M. Khoshgoftaar and D. Wang, A survey of transfer learning, *J Big Data*, 3(1):1–40, 2016.

[3] K. Worden and E.J. Cross and R.J. Barthorpe and D.J. Wagg and P. Gardner, On digital twins, mirrors, and virtualizations: Frameworks for model verificatio and validations, *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, 6(3): 030902, 2020.

[4] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer New York, 1999.

[5] V. Vapnik and S. Kotz, *Estimation of Dependences Based on Empirical Data*. Springer New York, 2006.

[6] D.P. Kingma, Auto-encoding variational Bayes, *arXiv*, 2013.

[7] L. Mescheder and S. Nowozin and A. Geiger, Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks, *International Conference on Machine Learning*, 2391–2400, 2017.

[8] K. He and X. Zhang and S. Ren and J. Sun, Delving deep into rectifiers Surpassing human-level performance on ImageNet classification *Proceedings of the 2015 IEEE International Conference on Computer Vision*, 1026–1034, 2015.

[9] Y. Ganin and E. Ustinova and H. Ajakan and P. Germain and H. Larochelle and F. Laviolette and M. March and V. Lempitsky, Domain-adversarial training of neural networks, *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[10] S. Ben-David and J. Blitzer and K. Crammer and A. Kulesza and F. Pereira and J.W. Vaughan, A theory of learning from different domains, *Machine Learning*, 79:151–175, 2010.