# TESTING THE DOMAIN INVARIANCE HYPOTHESIS IN LOW TARGET DATA DOMAIN ADAPTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Out-of-Distribution (OOD) generalization aims to ensure robust model performance when a *target* distribution differs from the *source* distribution observed during training. Within this broader problem, Domain Generalization (DG) considers the challenging setting where no target-domain data are available during training, whereas Domain Adaptation (DA) leverages target-domain information to facilitate generalization to the target domain. In practice, the transition between DG and DA depends on the amount of target data and is particularly relevant when only a few target samples are available. We therefore revisit this dichotomy by introducing a unified formulation that treats DG and DA as *continuum* governed by the amount of available target data. To this end, we extend the experimental protocol *DomainBed* designed for DG experiments to Unsupervised Domain Adaptation (UDA) and Supervised Domain Adaptation (SDA), and empirically investigate how adaptation occurs through domain discrepancy and decision boundaries evolutions. We investigate the role of the domain-invariant hypothesis across the transition from DG to DA. If domain invariance contributes to generalization, greater domain invariance should be systematically associated with improved target-domain performance. We therefore compare statistical (MMD) and adversarial (DANN) alignment with alignment-free ERM under varying target-data availability, allowing us to examine whether alignment systematically translates into improved target performance along the proposed critical dimension. To assess the generality of our findings across modalities, we conduct experiments on the CWWS speech dataset, which includes various types of corruptions, and the popular VLCS image dataset. Across both data modalities, we observe comparable overall trends, while the supervised and unsupervised adaptation continuums exhibit distinct behaviors as target-data availability increases. Explicit domain-alignment does not outperform classical ERM in DG→SDA while they reach early performance plateaus in DG→UDA despite target error reduction in all scenarios. Quantitative analysis identifies the factors that account for these differences and assesses the empirical validity of the domain-invariant hypothesis in our experimental setting. The code will be publicly released upon publication of the paper.

## 1 INTRODUCTION

While effective, most machine learning models are vulnerable to distribution shift, often exhibiting degraded performance when the inference-time data distribution differs from the training distribution (Saenko et al., 2010; Moreno-Torres et al., 2012). The challenge of mitigating this phenomenon is known as Out-Of-Distribution (OOD) generalization (Arjovsky, 2020) and represents typical setup in real-world applications of machine learning in which a model is trained on *source* domains while deploying the algorithm on a *target* domain. The transfer of knowledge stands on an hypothesis of proximity between the *source* and the *target* distributions (Sugiyama et al., 2007) whose importance is often underestimated. In real-world applications this hypothesis may not be fulfilled.

Depending on the availability of *target* data available at training time, several OOD scenario can be defined: Domain Generalization (DG) (Aubin et al., 2021; Arjovsky, 2020; Zhang et al., 2024; Zhou et al., 2022) when only *source* training data are available and Unsupervised or Supervised Domain Adaptation (UDA or SDA) (Ben-David et al., 2006; Blitzer et al., 2007; Farahani et al., 2021) when respectively unlabeled or labeled *target* data are also available.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, we build on the domain-invariant representation hypothesis underlying much of the DG literature (Aubin et al., 2021; Arjovsky, 2020), according to which a suitable representation should elicit a predictor that remains invariant across environments, i.e., whose predictive relationship with the target is stable across environments. We further examine this hypothesis experimentally by considering all source domains as interchangeable. It motivates the choice of single-source, single-target alignment methods as a means to investigate whether stronger source-target alignment is associated with improved target performance. We compare statistical alignment using MMD (Gretton et al., 2012; Li et al., 2018) and adversarial alignment using DANN (Ganin et al., 2016) with the alignment-free ERM baseline (Vapnik, 1999). However comparing algorithms is difficult because the results are often specific to the evaluated task, the choice of domains and the selection of hyperparameters. Consequently benchmarks have been built to allow fair comparison of algorithms either in DG (Gulrajani & Lopez-Paz, 2021) or in DA (Ringwald & Stiefelhagen, 2021; Gnassounou et al., 2024; Kalluri et al., 2024).

In many practical setups, target-domain data are often available only in limited quantities. Beyond performance in pure Domain Generalization or Domain Adaptation settings, the ability to exploit increasing amounts of target data emerges as an additional dimension for characterizing algorithms: methods that perform well without target data may not be the most effective when some target data become available, and vice versa. However, how domain alignment affects generalization as target supervision increases—from domain generalization to domain adaptation—remains largely unexplored. Similarly, the transition from domain generalization to domain adaptation in the context of low-traget-data regime remains poorly understood. Particularly, recent work observe an early plateau in performance as target data becomes available, highlighting potential limitations of domain invariance (Kalluri et al., 2024). In the present work, we address this question within a unified framework spanning Domain Generalization and Domain Adaptation.

Concretely, this work aims to explore how Domain Generalization and Domain Adaptation can be framed as points along a continuum, as the availability of target domain data increases. Intuitively, a good generalization should lead to easier adaptation as more general features are learnt and we design our experiments to test this hypothesis. We use the well-known *DomainBed* (Gulrajani & Lopez-Paz, 2021) Domain Generalization framework as an experimental foundation for a controlled study of the DG–DA transition over two modalities, speech and image. By extending it to unsupervised and supervised adaptation as a function of target data availability, this work is guided by one main question:

*How does increasing target-domain information affect the invariance of learned representations and the resulting decision boundaries?*

While Gulrajani & Lopez-Paz (2021) focus on benchmarking and comparing the performance of generalization methods, our work complements this perspective by investigating how adaptation methods behave as target-domain data becomes increasingly available, with a particular focus on their structural properties and performance. Main contributions are summarized as follows:

- We **propose a unified formulation to study the continuum** (Equation 1) between DG and DA that allows a new dimension for characterizing algorithms by their ability of adaptation with few target-domain data from a DG setting.
- Through empirical observations, we **investigate the effectiveness of the domain invariance hypothesis in the low-target-data regime**, showing that increased invariance does not necessarily translate into improved target performance.

The paper is organized as follows: Section 2 introduces the key concepts and definitions underlying this work and situates our contribution within the existing literature. Section 3 introduces our unified formulation, which bridges DG and DA through varying levels of target-domain data availability, and our proposed methodology to evaluate domain invariance hypothesis. Section 4 details the experimental setup. Section 5 provides experimental results and analysis by first observing the multi-source positive impact in the low target data regime, and secondly discussing the domain invariance hypothesis through quantitative measures.

## 2 BACKGROUND AND DEFINITIONS

**Domains.** As defined by Ben-David et al. (2006), a domain $\mathcal{D}$ is a probability distribution over two random variables $(\mathcal{X} \times \mathcal{Y})$, from which pairs of samples $(x, y)$ are drawn. For classification, $x$ are features and $y$ are labels. In unlabeled case, $\mathcal{D}$ is defined as a probability distribution over $\mathcal{X}$ from which samples $x$ are drawn. In this study, we assume a multi-source learning scenario, and the single-target domain setting i.e. the source domain is the union of several domains and the target is a unique domain. We deal with both *covariate* shift, when the input distribution changes between source and target domains while the conditional distribution between inputs and ground-truths remains invariant (Sugiyama & Müller, 2005; Gretton et al., 2009), with VLCS dataset (Fang et al., 2013), and *corruption* shift, when target input differ due to noise, perturbations, or environmental conditions compared to the clean source input while the underlying semantics remain unchanged (Rajagopalan & Patel, 2026), with CWWS dataset (Derrida et al., 2026).

**The Domain Invariance Hypothesis.** One influential line of work approaches the domain shift problem through the principle of domain invariance: the model should learn a representation in which the distribution of relevant features is similar across several domains. This idea was formalized for binary classification by Ben-David et al. (2006) who derives a theoretical upper bound on the target error $\epsilon_T$. This target error is controlled by the source-domain error $\epsilon_S$, the $\mathcal{H}$-divergence between the source and target domains $d_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$, which is a measure of discrepancy between the domains in a space of representations, and $\lambda = \min_{h \in \mathcal{H}} \{\epsilon_S(h) + \epsilon_T(h)\}$ the optimal joint source-target error among a set of hypotheses. Building on this framework, Mansour et al. (2009) introduce the discrepancy distance, a distributional distance tailored to the learning problem. Their analysis further motivates minimizing source-target discrepancy as a means of improving adaptation performance. Early work by Muandet et al. (2013) formalized this idea through Domain-Invariant Component Analysis that motivated large family of approaches including statistical feature alignment approaches such as Maximum Mean Discrepancy (MMD) (Gretton et al., 2012; Li et al., 2018), adversarial alignment methods such as Domain-Adversarial Neural Networks (DANN) (Ganin et al., 2016), invariant representation learning methods such as IRM (Arjovsky et al., 2020). This idea was extended to Domain Generalization (DG). Domain alignment methods can be used to learn a representation in which the distribution of relevant features is similar across the observed source domains, with the goal of enforcing invariance with the unknown target domain (Arjovsky, 2020). The domain invariance hypothesis has lead to effective DG and DA algorithms, but the empirical evaluation of the relation between domain invariance in a representation space and target performance remains scarce. Vedantam et al. (2021) empirically evaluates the correlation of several metrics with target performance. Especially they show that the discrepancy between domains measured with *MMD-Mean-Cov* can correlate with target error.

**Benchmarks.** Several studies compare algorithms in the Domain Generalization (DG) setting (Galstyan et al., 2022; Cha et al., 2021), notably through *DomainBed* (Gulrajani & Lopez-Paz, 2021), which provides a unified evaluation protocol with standardized model selection and hyperparameter search. This framework improves reproducibility and comparability, while showing that many DG methods provide only marginal improvements over the Empirical Risk Minimization (ERM) baseline. More recently, standardized evaluation has been developed for Unsupervised Domain Adaptation (UDA): SKADA-Bench (Gnassounou et al., 2024) focuses on realistic model-selection comparisons, while UDA-Bench (Kalluri et al., 2024) investigates the influence of backbone architecture, pre-training, and, notably, the amount of unlabeled target data. Their experiments show that UDA performance often saturates rapidly, with up to 75% of target data being removed while causing less than a 1% decrease in target accuracy across the studied settings. This highlights the need to better understand how target-data availability affects adaptation performance.

**Target data availability.** Beyond benchmarking, prior work has also considered DG and DA as points along a common spectrum. Motiian et al. (2017b) proposed a unified formulation of supervised domain adaptation and domain generalization, while their subsequent work on few-shot adversarial domain adaptation Motiian et al. (2017a) explicitly considered adaptation with very limited target supervision. However, they do not systematically characterize how performance evolves along target-data availability dimension under a common evaluation protocol to understand underlying mechanisms of adaptation.

# 3 Unifying Domain Generalization and Adaptation Under Domain Invariance Hypothesis

This section defines the different learning regimes and proposes a unified point of view on the continuum between domain generalization and adaptation to study them in a structured experimental protocol. First we show how the different learning regimes can be viewed as points along the target-data availability dimension. Then we discuss the problem of hyperparameter selection along the continuum. Finally we introduce metrics for characterizing domain invariance.

## 3.1 A Unified Learning Formulation

We propose a unified formulation of the learning objective for Domain Generalization (DG), Supervised Domain Adaptation (SDA) and Unsupervised Domain Adaptation (UDA). This point of view encompasses only a subset of the domain generalization and adaptation algorithms, where alignment between domains is enforced with a regularization loss.

$$\mathcal{L} = \underbrace{\mu^S_{obj}\mathcal{L}_{obj}(\mathcal{D}_S)}_{\text{Multi-Source supervision}} + \underbrace{\mu^T_{obj}\mathcal{L}_{obj}(\mathcal{D}_T)}_{\text{Target supervision (SDA)}} + \underbrace{\mu^S_{reg}\mathcal{L}_{reg}(\mathcal{D}_S)}_{\text{Inter-Source regularization}} + \underbrace{\mu^T_{reg}\mathcal{L}_{reg}(\mathcal{D}_S, \mathcal{D}_T)}_{\text{Target-driven adaptation (SDA/UDA)}} \quad (1)$$

where $\mathcal{L}_{obj}$ denotes the objective loss (e.g., cross-entropy), while $\mathcal{L}_{reg}$ denotes a regularization loss that measures the alignment between domains. The latter is defined according to the specific alignment algorithm (MMD or DANN) and can be computed without access to class labels. Depending on data availability, each learning regime corresponds to different supervision settings, characterized by specific assumptions on the hyperparameters $\boldsymbol{\mu}$ in Equation 1:

**Domain Generalization (DG)** refers to experiments for which only source data are available during training, no target data.. Therefore, target data have no influence on the global loss defined in Equation 1 i.e. $\mu^T_{obj} = \mu^T_{reg} = 0$ (in blue in Figure 1).
**Unsupervised Domain Adaptation (UDA)** refers to experiments in which target data are available during training but without labels i.e. $\mu^T_{obj} = 0$. This regime lives in a plan (in pink in Figure 1).
**Supervised Domain Adaptation (SDA)** refers to experiments for which source and target data are seen during training and all domains contain labels associated to there data. Therefore Equation 1 keeps all degrees of freedom in terms of hyperparameters $\boldsymbol{\mu}$ (in green in Figure 1).

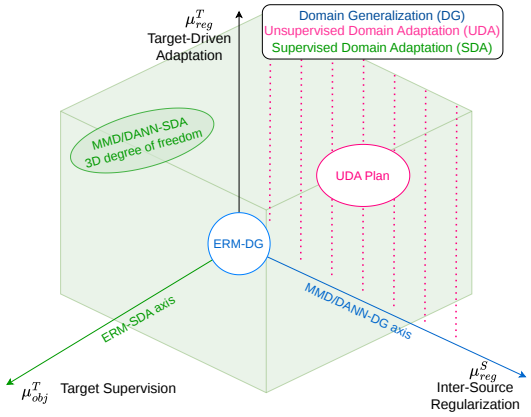



Figure 1: Unified view of the continuum between scenarios in a 3-dimensional space. Axis are non-negative loss weighting hyperparameters defined in Equation 1. The ERM, MMD and DANN algorithms are described in Section 4.2

.

## 3.2 Learning Along the Domain Continuum

As argued by Gulrajani & Lopez-Paz (2021), a fair comparison between algorithms implies a comparable budget of hyperparameters search over a reasonable set of values with a clear selection criterion (for instance source or target accuracy). We propose the following protocol.

**Hyperparameters selection.** To compare DG and DA experimentally we must have consistent experiments along the continuum. Since methods are very sensitive to the choice of hyperparameters, we place ourselves in a realistic domain adaptation scenario where several hyperparameters

are fixed along the target data availability axis. We first select hyperparameters for DG among 20 configurations and reduce the number of hyperparameters seeds from 3 to 1 (Yu et al., 2024) due to computational constraints. These hyperparameters are kept fixed in DA experiments, while only the target-specific hyperparameters are optimized through a grid search over 10 configurations. The list of hyperparameters tuned for each algorithm in DG and DA settings are detailed in Appendix A.

**Model Selection Method through DG-DA continuum.** Following Gulrajani & Lopez-Paz (2021), we consider two model selection strategies: (i) selection based on validation accuracy on source-domain data, and (ii) an *Oracle* strategy, which selects the model using validation data from the target domain. Along the proposed DG-DA continuum, for a given target-data availability $X\%$, we select the best model among all available experiments using at most $X\%$ of target data, including DG ($0\%$). Model selection is particularly challenging in DG and UDA settings, where target data are unavailable for validation. Since model selection is not the focus of this study, we provide a comparison of selection methods in Appendix B and use Oracle selection for both continuums to disentangle model selection effects in the following analysis.

### 3.3 Metrics for Quantifying Domain Invariance

Domain adaptation theoretical bounds discussed in Section 2 relate the target error to three quantities: the source error, the discrepancy between the source and target domains, and the joint error of the optimal classifier. We use these quantities as diagnostic measures to investigate whether domain invariance is associated with improved target performance along the proposed continuum. The proposed metrics are proxy measures and do not correspond to a theoretical guarantee. All metrics are evaluated on the test sets.

- **Error rate**: We report the classification error on both the source and target domains, denoted by $\hat{\epsilon}_S$ and $\hat{\epsilon}_T$, respectively.
- ***Optimal* joint-error $\hat{\lambda}$.**: We consider an empirical proxy for the joint-error term denoted by $\hat{\lambda}$. $\hat{\lambda}_{ST}$ is defined as the sum of the source and target error rates of the classifier that minimizes the joint error across the two domains. We also consider an inter-source variant, $\hat{\lambda}_{IS}$, computed analogously using the source domains only. It characterizes the intrinsic difficulty of simultaneously solving the task across the source and target domains.
- **Domain discrepancy.** We quantify the alignment between source and target domains in the learned embedding space using using the MMD-Mean-Cov discrepancy, following the experiments from Vedantam et al. (2021). We evaluate the discrepancy both between source and target domains ($ST$) and between every pairs of source domains ($IS$). See Appendix C for details).

Together, these measures provide a quantitative analysis of whether increasing target-data availability induces a progressive change in domain alignment and task performance, rather than relying solely on target accuracy to characterize the proposed continuum.

## 4 Experimental setup

### 4.1 Datasets

We evaluate on two datasets spanning vision and audio. VLCS (Fang et al., 2013) comprises images from four domains—VOC2007, LabelMe, Caltech101, and SUN09—with a five-class object classification task (bird, car, chair, dog, person). We additionally use CWWS (Derrida et al., 2026), which studies corruption shifts in speech recordings with a four languages classification task (french, spanish, german, mandarin chinese) across four conditions: *clean*, *wham*, *wind*, and *saturation*. Considering both visual and audio modalities allows us to assess whether observations hold across distinct data distributions and modalities. For each dataset, we consider four leave-one-domain-out configurations, holding one domain out as the target and using the remaining domains as sources. It results into a total of four different configurations per dataset and we report averaged measures across these four configurations. In all experiments, the quantity of data is expressed in terms of percentage of

the quantity of training data available for one domain. For VLCS this quantity is 550 images, for CWWS it is 1125 speech utterances.

### 4.2 Backbones and Algorithms

We use backbones used in previous studies : a pretrained wav2vec2 speech encoder (Baevski et al., 2020) for CWWS dataset (Derrida et al., 2026) and ResNet50 (He et al., 2016) for VLCS (Gulrajani & Lopez-Paz, 2021). We perform finetuning with the following algorithms on the two proposed datasets. Since our goal is to understand the effectiveness of the domain invariance hypothesis, we consider two alignment-based algorithms, MMD and DANN, in addition to the canonical baseline ERM allowing us to disentangle the gains arising from target data availability from those specifically attributable to alignment. We briefly describe the selected algorithms:

- *Empirical Risk Minimization (ERM)* Vapnik (1999) aims to reduce the total error by minimizing the aggregate of losses over all domains and data points. This approach supposes no regularization i.e., $\mu^S_{reg} = \mu^T_{reg} = 0$ in Equation 1.
- *Maximun Mean Discrepancy (MMD)* Sun & Saenko (2016); Gretton et al. (2012); Li et al. (2018) minimizes the distance between the means of two distributions in a high-dimension space defined by kernels. Regularization is performed by matching all pairs of domains.
- *Domain-Adversarial Neural Network (DANN)* Ganin et al. (2016) uses an adversarial objective to align feature distributions across domains via learning a discriminator between domains. Regularization is performed for all domains present in the train set at once.

## 5 Results and Discussions

This section reports results from our experiments for $DG \rightarrow SDA$ and $DG \rightarrow UDA$ continuums. In Subsection 5.1, we analyse the evolution of performance with the quantity of target data. The comparison between systems trained on several domains and a mono-source system, trained with target data, allows to define the low-target-data regime. Then Subsection 5.2 discusses the domain invariance hypothesis in the low-target-data regime through quantitative analysis. For each dataset, we focus on the domain-averaged behavior in the main text, with metrics averaged across the four configurations corresponding to the four target domains; detailed results for individual domains are reported in Appendix E.

### 5.1 Domain Generalization and Adaptation in the Low-Target-Data Regime

Figure 2 reports the evolution of classification error as a function of target data availability for which we consider experimentally five granularities in the target-data availability dimension: 0% (DG), DA-2%, DA-10%, DA-50% and DA-100%. Solid lines correspond to supervised adaptation while dashed lines to unsupervised adaptation. Colors correspond to adaptation algorithms and the black solid curve refers to a mono-source baseline system, trained only on target data with the same hyperparameters as the corresponding ERM-SDA model.

**The Low-Target-Data Regime: MS in-domain vs. Multi-Source Learning** We first consider the classification error of a Mono-Source (MS) system as a function of the amount of available target data. Figure 2 shows this evolution in black. Without any target data (DG) performance is random classification as no finetuning is possible and more target data reduces the classification error. We are interested in the regime where the quantity of target data is insufficient to reach satisfactory performance: the low-target-data regime. The performance of DG, UDA and UDA algorithms in colors show the large benefit of multi-source training. On VLCS, between 10% and 50% of target data are needed to outperform multi-source DG→DA continuums, whereas on CWWS this quantity is not reached experimentally and is supposed to be upon 100% target data availability.

**The two continuums.** Figure 2 reports target-domain classification error along the DG→UDA (dashed lines) and DG→SDA (solid lines) continuums. Global trend is similar for the two studied datasets. Both continuums improve over the target-only baseline in the low-target-data regime, demonstrating the benefit of leveraging multiple source domains when target data are scarce. Despite

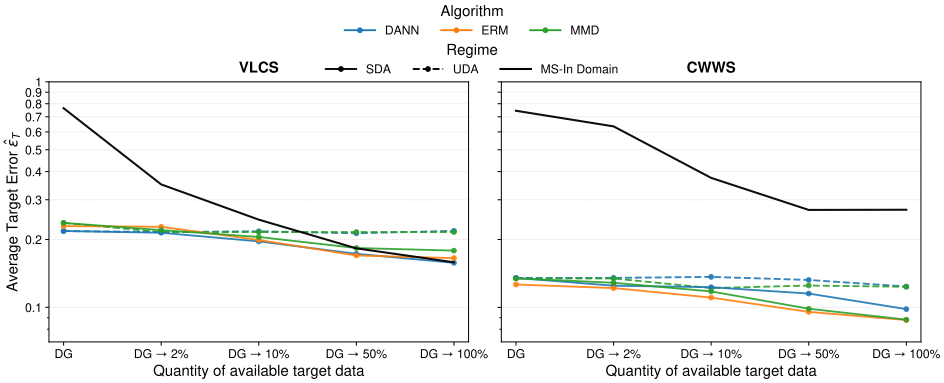



Figure 2: Average target error as the training dataset is increasingly enriched with target data. Columns compare UDA (dashed lines with dots) and SDA (solid lines with dots) continuums. Mono-Source (MS-IN) training baseline is reported in black.

global gains over DG, we observe that SDA consistently benefits more from additional target data than UDA, while UDA rapidly saturates, consistent with previous observations (Kalluri et al., 2024). Across both continuums, ERM, MMD, and DANN exhibit similar target performance, despite their different objectives which extends the findings of Gulrajani & Lopez-Paz (2021) in our setting.

**Summary –** We observe empirically the low-target-data regime as a range where multi-source training outperforms mono-source in-domain training under a given target data budget. An overall trend is shared between datasets, despite the quantity of data corresponding to the boundary varies across them. A natural hypothesis is that this multi-source efficiency arises from learning a representation that is globally invariant across source domains. However, the similar behavior observed across ERM, DANN, and MMD leaves two possibilities open: (i) global invariance may emerge implicitly even without an explicit alignment objective, or (ii) global representation invariance may not be the primary mechanism underlying the efficiency of multi-source training. Distinguishing these explanations motivates the subsequent quantitative analysis and investigate the domain invariance hypothesis.

### 5.2 Is global invariance actually responsible for low target data efficiency ?

Figure 3 compares the UDA (left column) and SDA (right column) continuums from the point of view of domain invariance. We examine how the embedding space evolves as the amount of target data increases, with the metrics introduced in Section 3.3: error rates, discrepancies and joint-error measurements. Since similar behavior can be observed between the two studied datasets, we report results for VLCS dataset in Figure 3 and provide results for both datasets in Appendix E. Through quantitative measurements, we discuss the domain invariance hypothesis along the target data availability dimension by answering a series of questions.

**Does multi-domain training alone produce domain invariance?** If multi-domain training implicitly encourages domain-invariant representations, ERM trained on multiple source domains should reduce inter-source discrepancies relative to the mono-source (MS-IN) setting. Such a reduction should also translate into greater source–target compatibility and, ultimately, improve target performance. Figure 3 provides mixed evidence for this hypothesis. Compared with MS-IN, the multi-source setting (ERM-DG and ERM-SDA) generally exhibits lower source–target discrepancies, together with lower joint-error terms ($\hat{\lambda}_{ST}$, $\hat{\lambda}_{IS}$). Multi-source training produces representations that are more compatible across domains. However the inter-source discrepancy is not systematically smaller for ERM-DG than MS-IN. Our results therefore do not support domain invariance as a systematic consequence of multi-domain ERM across the datasets and target-data regimes considered. Either multi-source ERM implicitly learns a form of domain invariance that is not captured uniformly by the discrepancy measures, or global domain invariance is not primarily responsible for the multi-source advantage.

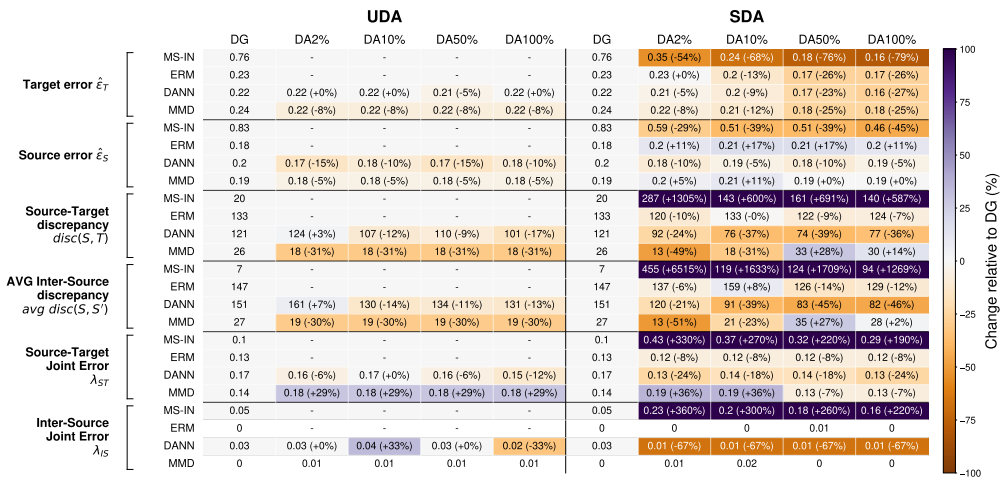



Figure 3: Quantitative metrics for invariance analysis for VLCS dataset Domain-averaged target and source errors, source-target optimal joint-error $\hat{\lambda}_{ST}$, inter-source (IS) optimal joint error $\hat{\lambda}_{IS}$. Colors and values in parenthesis highlight a reduction (in read) or an increase (in blue) of the corresponding metric relatively to DG initial value. We report algorithms and MS-IN baseline defined in Section 5.1. Both datasets are reported in Figure 6. Appendix E reports per domain results.

**Do explicit domain-alignment objectives actually align domains?** The comparison of DG algorithms shows that MMD makes the source domains closer to each other, compared to ERM. In addition, MMD regularization between source domains (DG) also reduces the discrepancy between source and target domains. This behavior validates the underlying hypothesis of domain invariance for DG. However it is accompanied by a degradation of the joint source-target error, suggesting that domains become less compatible. The DG→UDA continuum shows that for both datasets MMD algorithms reduce significantly the source–target and inter–source discrepancies while DANN provides limited reductions from ERM setting. The SDA continuum provides contrary observations since MMD globally increases discrepancies measures along the target-data dimension while DANN reduces widely suggesting a large impact of labels compared to data distributions during learning.

**Does global source-target alignment translate into better target performance?** Empirical results show that this relationship is not systematic. For instance, VLCS-UDA with MMD leads to a reduction from 130 (ERM-DG) to 18 (MMD-UDA 100%) in the source–target discrepancy while the target error $\hat{\epsilon}_T$ reduces from 0.23 (ERM-DG) to 0.22 (MMD-UDA 100%). Interestingly the joint-error term $\hat{\lambda}_{ST}$ increases. For several systems, we observe that an improvement of domain invariance, in terms of discrepancy, can incur a degradation of the embedding space for joint classification, as measured by $\lambda$. This possible behavior is theoretically analyzed by Zhao et al. (2019) and our work constitutes an empirical validation.

**Is inter-source invariance efficient in the low-target-data regime?** The three algorithms, ERM, MMD and DANN, lead to broadly similar trends across both UDA and SDA continuums. Since modest improvements are observable in UDA setting with source invariance and explicit alignment, it suggests that source invariance does not, by itself, consistently make the integration of a target domain more effective. Thus, the multi-source advantage in the low-target-data regime cannot be attributed straightforwardly to global inter-source invariance.

**Summary –** The generalization and adaptation algorithms inspired from the domain invariance hypothesis can substantially change domain representations but do not systematically explain target performance evolution. In particular, large reductions in source–target discrepancy can coexist with an increase in $\hat{\lambda}_{ST}$. This motivates examining source domains individually rather than treating them globally: considering the aggregation of source domains may obscure substantial differences between individual source domains and their relationships with the target.

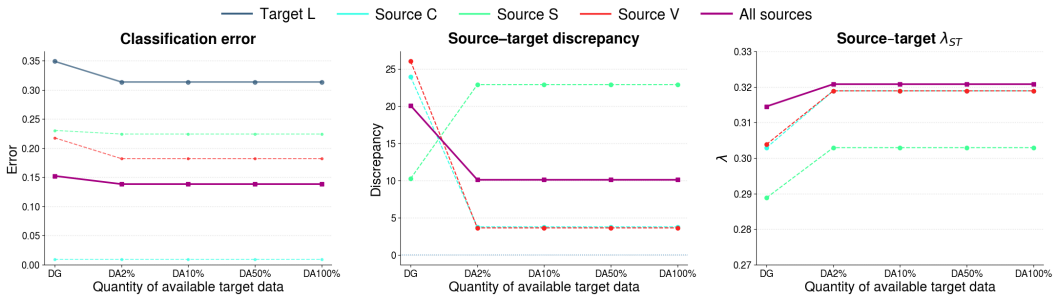



Figure 4: Analysis of the MMD algorithm for UDA on VLCS dataset with $L$ target domain. Detailed metrics evolution through target data availability.

**Should all source domains be considered equally?** Figure 4 provides a detailed example from VLCS, using domain $L$ as the target and MMD as the alignment method. We report the individual source–target metrics across target-data availability, in addition to the aggregate quantities obtained when considering all source domains together. We observe heterogeneous relations between source and target domains (Source S in green behaves differently compared to source C), which is largely obscured by the aggregate source–target measure (pink curve). This does not contradict the domain-invariance hypothesis, but suggests that global invariance objective may provide an incomplete description. This motivates investigating whether source–target proximity or structured relationships among source domains better characterize the low-target-data regime and can be leveraged by structured formulations (Turrisi et al., 2020; Yang et al., 2026).

**Limitations and awareness.** While our results provide insights into the role of domain invariance in the low-target-data regime, several considerations should be taken into account when interpreting these findings. This study is conducted on two datasets CWWS (audio) and VLCS (vision). It represents only a single dataset per modality, which may limit the direct generalization of quantitative conclusions across all possible domain shifts. Our goal is to analyze trends in a controlled setting.

## 6 Conclusion

In this paper, we focused on domain adaptation with very few target data availability through an experimental protocol involving two modalities with four domains each and fair hyperparameter selection. Using three various algorithmic approaches, we investigated the effectiveness of the domain invariance hypothesis in this critical transition. We introduced a unified formulation bridging Domain Generalization (DG) and Domain Adaptation (DA) to study this transitional regime. This perspective extends the traditional DG–DA dichotomy by providing a more nuanced view of the underlying dynamics of adaptation across the low-target-data regime, and opens a promising direction for future work: quantitatively characterizing this regime target information measurements.

Through quantitative invariance metrics measurements, our experimental protocol showed limitations of the invariance hypothesis to continually integrate target data in the supposed domain-invariant representation (DG). Interestingly, we observed similarities between global behaviors across the two considered modalities, speech and image. The three considered algorithms exhibit comparable behavior across the target data continuums, suggesting that the learned representations do not systematically capture the domain structure required for more effective target-domain integration. We experimentally verified relative invariance efficiency in the multi-source setting compared to a Mono-Source baseline in the low target data regime. We tested the domain invariance hypothesis effectiveness in the low target data regime since intuitively an invariant representation should easily lead to quick integration of a new domain. In unsupervised continuum, we observed a moderate gain upon generalization not observed when target labels are available (supervised continuum).

Further improvements may require more structured approaches to the relationship between source and target domains, rather than relying solely on global domain-invariant representations. We hope our extension of the domainbed framework to SDA and UDA will facilitate further research on adaptation with limited target data.

### AI USE STATEMENT

We used large language models (LLMs) as writing assistance tools to improve clarity, conciseness and grammatical correctness of the manuscript. All generated text was reviewed and verified by authors, who take full responsibility for the proposed content.

### REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method and experimental setup to facilitate reproducibility. The complete source code, configuration files, and instructions required to reproduce our experiments will be publicly released upon publication of the paper. We use publicly available benchmark datasets and evaluate our methods for research purposes.

## REFERENCES

Martin Arjovsky. *Out of distribution generalization in machine learning*. PhD thesis, New York University, 2020.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. URL https://arxiv.org/abs/1907.02893.

Benjamin Aubin, Agnieszka Słowik, Martin Arjovsky, Leon Bottou, and David Lopez-Paz. Linear unit-tests for invariance discovery. *arXiv preprint arXiv:2102.10867*, 2021.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20, 2007.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Francois Derrida, Raphaël Duroselle, Thomas Courtat, and Jean-François Bonastre. Acoustic domain shift in spoken language identification: From systematic domain generalization to real-world application. working paper or preprint, September 2026. URL https://hal.science/hal-05759404.

Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.

Tigran Galstyan, Hrayr Harutyunyan, Hrant Khachatrian, Greg Ver Steeg, and Aram Galstyan. Failure modes of domain generalization algorithms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19077–19086, 2022.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Théo Gnassounou, Oleksii Kachaiev, Rémi Flamary, Antoine Collas, Yanis Lalou, Antoine de Mathelin, Alexandre Gramfort, Ruben Bueno, Florent Michel, Apolline Mellot, Virginie Loison, Ambroise Odonnat, and Thomas Moreau. Skada : Scikit adaptation, 7 2024. URL https://scikit-adaptation.github.io/.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, Bernhard Schölkopf, et al. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3 (4):5, 2009.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Tarun Kalluri, Sreyas Ravichandran, and Manmohan Chandraker. Uda-bench: Revisiting common assumptions in unsupervised domain adaptation using a standardized framework. In *European Conference on Computer Vision*, pp. 199–220. Springer, 2024.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5400–5409, 2018.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *Conference on Learning Theory (COLT)*, 2009.

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. *Advances in neural information processing systems*, 30, 2017a.

Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5716–5726, 2017b. doi: 10.1109/ICCV.2017.609.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 10–18, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/muandet13.html.

Sudarshan Rajagopalan and Vishal M Patel. Your pre-trained diffusion model secretly knows restoration. *arXiv preprint arXiv:2604.04924*, 2026.

Tobias Ringwald and Rainer Stiefelhagen. Adaptiope: A modern benchmark for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 101–110, 2021.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010.

Masashi Sugiyama and Klaus-Robert Müller. Model selection under covariate shift. In *International Conference on Artificial Neural Networks*, pp. 235–240. Springer, 2005.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Rosanna Turrisi, Rémi Flamary, Alain Rakotomamonjy, and Massimiliano Pontil. Multi-source domain adaptation via weighted joint distributions optimal transport. *CoRR*, abs/2006.12938, 2020. URL https://arxiv.org/abs/2006.12938.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Ramakrishna Vedantam, David Lopez-Paz, and David Schwab. An empirical investigation of domain generalization with empirical risk minimizers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28131–28143. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/ecf9902e0f61677c8de25ae60b654669-Paper.pdf.

Chenxiao Yang, Zhiyuan Li, Shai Ben-David, and Nathan Srebro. Hierarchical domain generalization, 2026. URL https://arxiv.org/abs/2607.16528.

Han Yu, Xingxuan Zhang, Renzhe Xu, Jiashuo Liu, Yue He, and Peng Cui. Rethinking the evaluation protocol of domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21897–21908, 2024.

Wangyou Zhang, Robin Scheibler, Kohei Saijo, Samuele Cornell, Chenda Li, Zhaoheng Ni, Anurag Kumar, Jan Pirklbauer, Marvin Sach, Shinji Watanabe, et al. Urgent challenge: Universality, robustness, and generalizability for speech enhancement. *Proc. Interspeech 2024*, 2024.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International conference on machine learning*, pp. 7523–7532. PMLR, 2019.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

# A HYPERPARAMETERS DESCRIPTION

| Algorithm | Generalization-tuned Hyperparameters | Adaptation-tuned Hyperparameters |
|---|---|---|
| ERM | batch_size | erm_adapt_coeff |
| | lr | – |
| | wav2vec2_dropout | – |
| | weight_decay | – |
| MMD | batch_size | mmd_adapt_coeff_obj |
| | lr | mmd_adapt_coeff_pen |
| | wav2vec2_dropout | – |
| | weight_decay | – |
| | mmd_gamma | |
| DANN | batch_size | dann_adapt_coeff_obj |
| | lr | dann_adapt_coeff_disc |
| | wav2vec2_dropout | dann_adapt_coeff_gen |
| | weight_decay | – |
| | lr_g | – |
| | lr_d | – |
| | weight_decay_g | – |
| | weight_decay_d | – |
| | d_steps_per_g_step | – |
| | grad_penalty | – |
| | beta1 | – |
| | mlp_width | – |
| | mlp_depth | – |
| | mlp_dropout | – |

Table 1: Generalization-tuned and adaptation-tuned hyperparameters for ERM, MMD, and DANN. Revers to hyperparameter description in Section 3.2

## B SELECTION METHOD COMPARISON IN UDA CONTINUUM

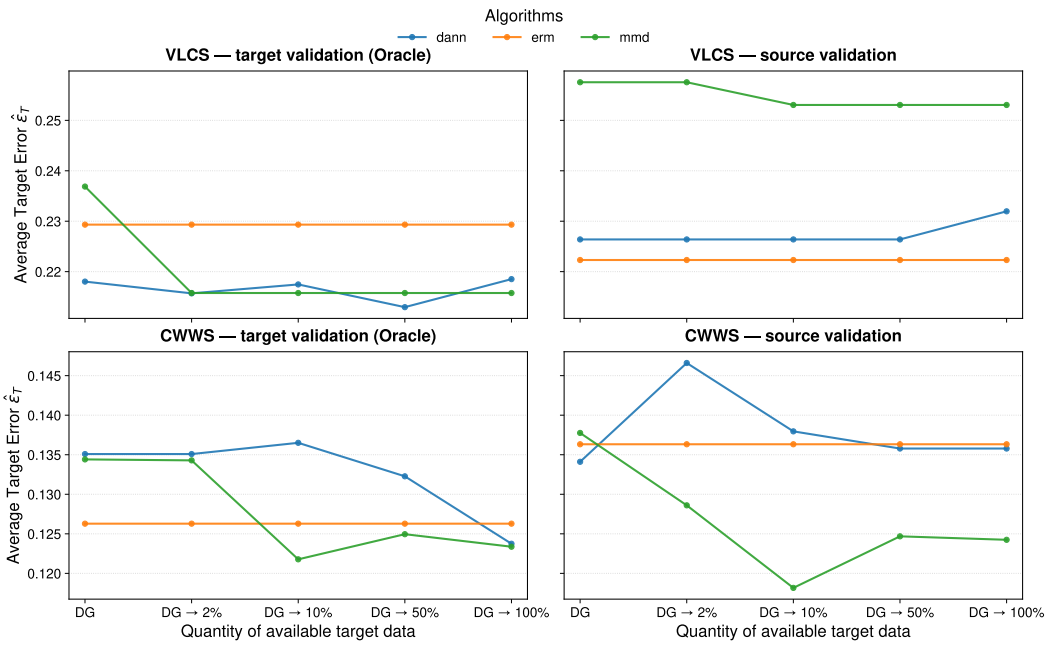



Figure 5: Comparison for all dataset between validation drawn from target (Oracle on the left) distributions or source (on the right) in UDA continuum. Related comments are in Section 3.2

**Model Selection through the continuum.** When unlabeled target data are available an important question is: *uppon which criterion can we select models ?* This question is central in many DG and UDA research but it is not the core of our approach. Figure 5 compares on the left Oracle (validation drawn from target distribution) and on the right source validation selection methods for CWWS dataset through the DG→UDA continuum. It should be noted that the source validation reported on the right more closely reflects a real-world setting, as unsupervised learning is typically employed when no target data are available. We observe no major differences, apart from the expected slight performance improvement with Oracle and the lack of improvement observed with DANN along the continuum. In the following study, we consider the Oracle selection method in order to compare with SDA continuum and avoid model selection bias.

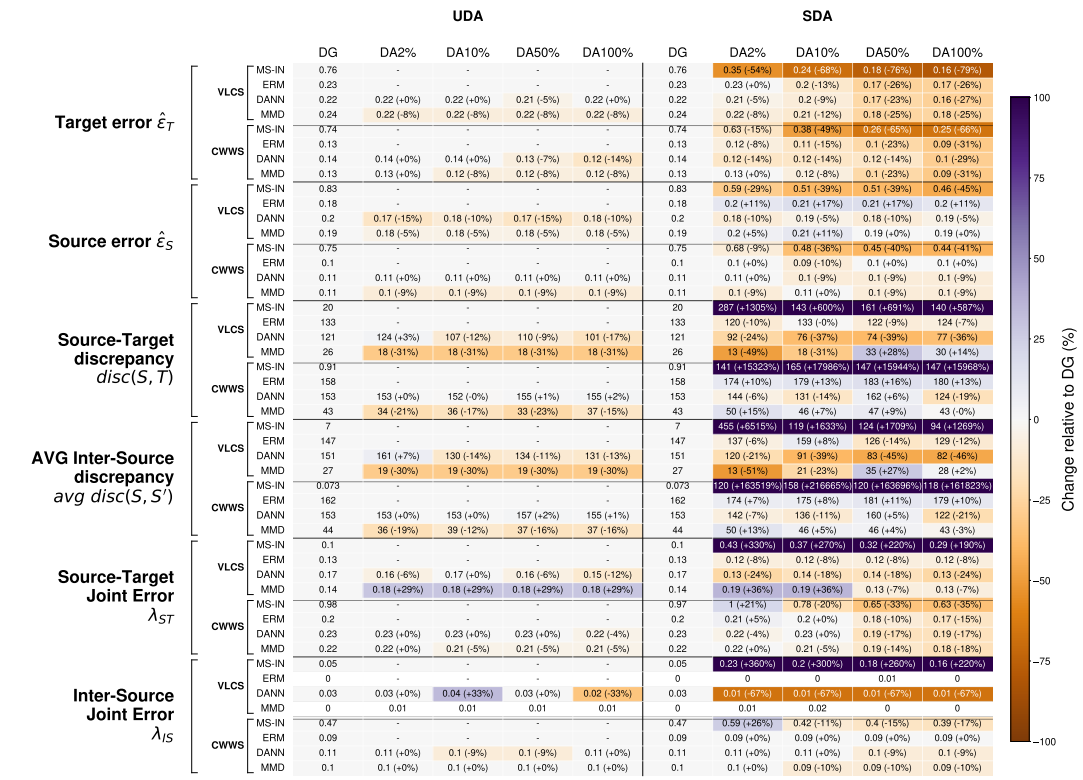



Figure 6: Analysis of the embedding space along the target data availability dimension for CWWS and VLCS datasets. Domain-averaged target and source errors, source-target optimal joint-error $\hat{\lambda}_{ST}$, nter- source optimal joint error $\hat{\lambda}_{IS}$. Colors and values in parenthesis highlight a reduction (in read) or an increase (in bluue) of the corresponding metric relatively to DG initial value. We report the three considered algorithms (ERM, MMD and DANN) and the baseline MS-IN defined in Section 5.1. The only-VLCS results are reported in Figure 3.

## C Discrepancy measures computation details

Following Vedantam et al. (2021), we measure the discrepancy between domains with *MMD-Mean-Cov*. This computation is performed in the embedding space, before the final classification layer. Inter-source discrepancy is computed as the average of the discrepancy between all pairs of source domains. Source-target discrepancy is computed as the average of the discrepancy between the target domain and each source domain.

## D Analysis metrics for both datasets

## E DETAILS PER TARGET DOMAIN PER ALGORITHM

In Figures 7, 8, 9, we present the metrics for each target domain. Each line correspond to an experiment with a specific target domain. The three other domains are the source domains. The metrics discussed in Section 5 are averaged across the four target domains for each dataset.

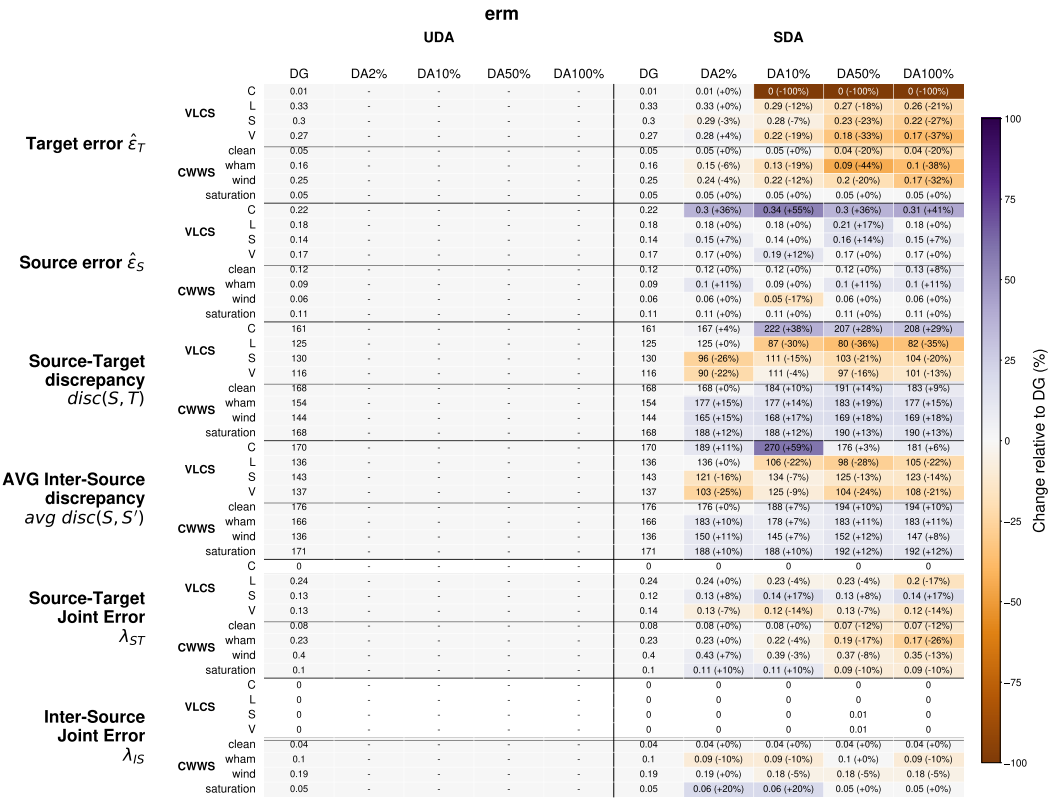



Figure 7: All metrics per domain for ERM algorithm. This plot details per domain behavior of results presented in Section 5.

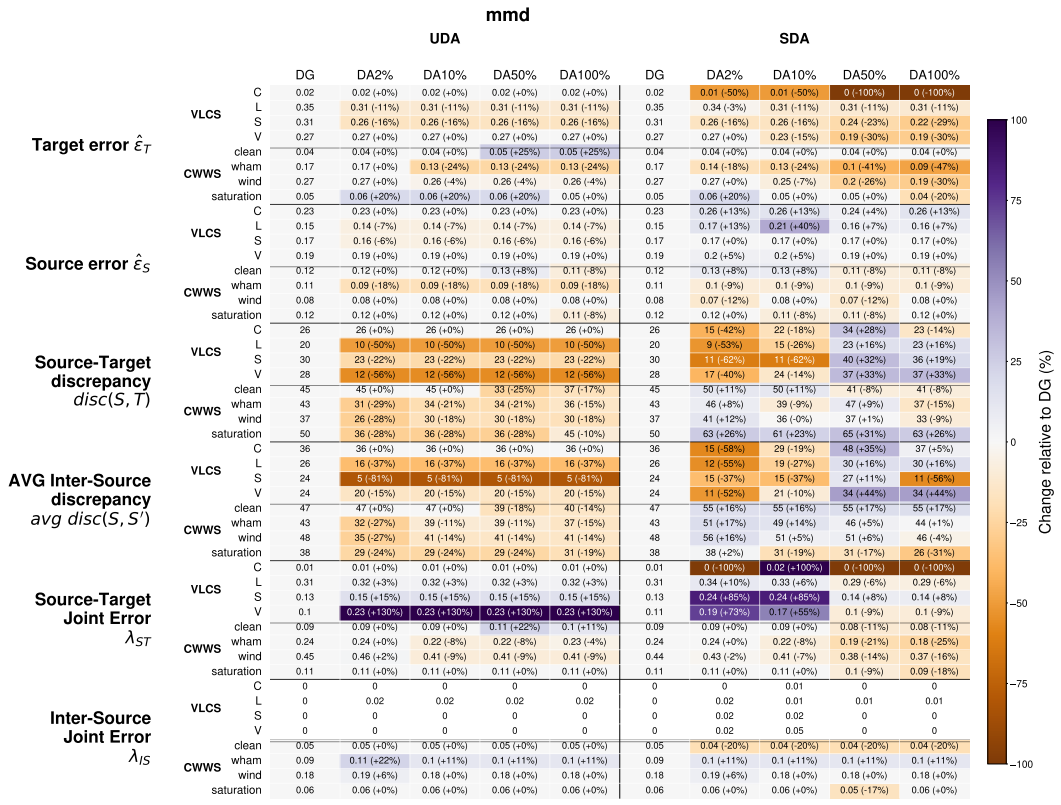



Figure 8: All metrics per domain for MMD algorithm. This plot details per domain behavior of results presented in Section 5.

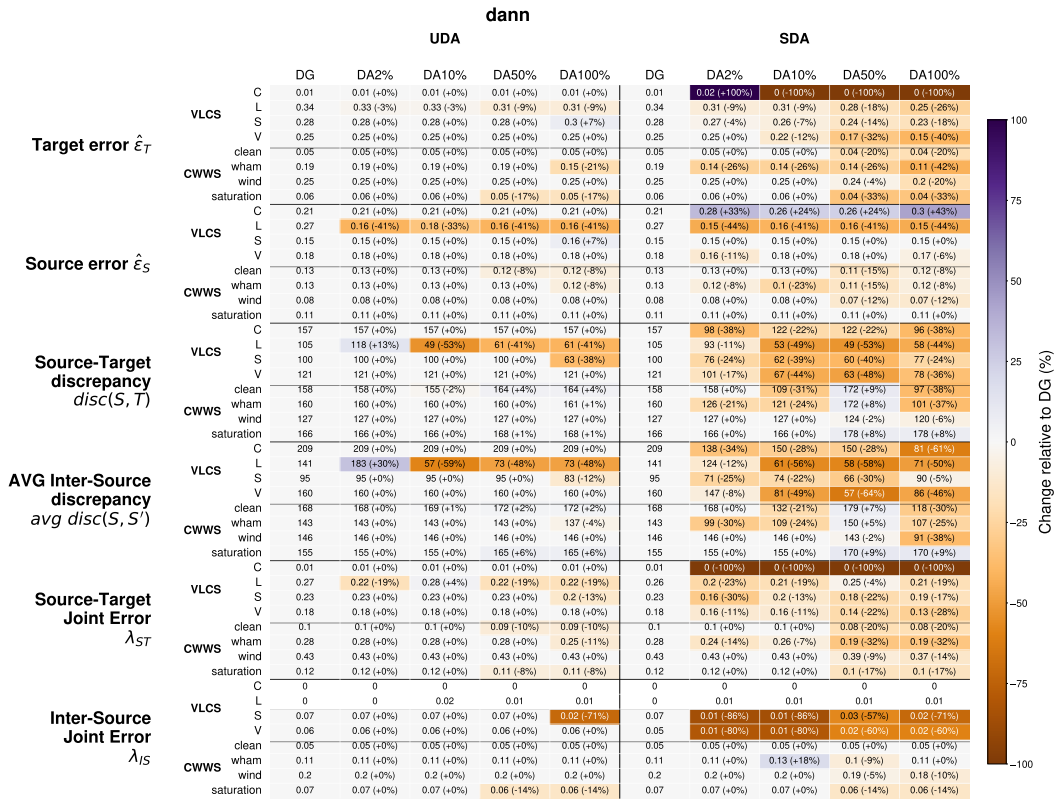



Figure 9: All metrics per domain for DANN algorithm. This plot details per domain behavior of results presented in Section 5.