

UNSUPERVISED DOMAIN ADAPTATION VIA FEATURE ALIGNMENT AND REDUNDANCY REDUCTION

Anonymous authors

Paper under double-blind review

ABSTRACT

Unsupervised domain adaptation (UDA) addresses the challenge of transferring knowledge from a labeled source domain to an unlabeled target domain for the same task under data distribution shifts. Current approaches rely on strong hypotheses about the type of domain shift, or task to perform. We propose FARR, a novel UDA method with 3 main contributions: (i) a new feature-alignment strategy based on redundancy reduction, that is task-adaptable and agnostic to the type of domain shift; (ii) a theoretical proof that our formulation effectively aligns source and target features; (iii) a comprehensive empirical evaluation across classification and segmentation tasks, using seven public and two private datasets covering diverse domain shifts. Our results show that FARR consistently outperforms existing feature-alignment methods, while remaining competitive with state-of-the-art UDA approaches across tasks and datasets. Our code is available at ¹.

1 INTRODUCTION

Modern deep learning models achieve impressive performance across a wide range of tasks, yet their success is built on the availability of large-scale labeled datasets for supervised training. However, when the deployment dataset follows a different distribution from the training data, this domain shift hampers successful generalization. Unfortunately collecting a large annotated dataset in the target domain is unfeasible in many real-world scenarios, due to the time, cost, and expertise required. Domain adaptation, a subset of transfer learning, addresses this challenge in the case where source and target tasks agree. Here, we focus on Unsupervised Domain Adaptation (UDA), where annotations are only available in the source domain. Our goal is thus to exploit the annotated source data to learn a model that performs well on the unlabeled target domain.

UDA is a well-studied and actively evolving research area, where methods can be characterized as feature-alignment, image-alignment, or self-training-based. Feature-alignment methods such as Ganin et al. (2016), Tzeng et al. (2017), Javanmardi et al. (2018) learn a feature space and align the distributions of source and target data in that space. These methods are usually *restricted* to a specific task, as in Zhang et al. (2019). Image-alignment methods such as Chen et al. (2019), Hoffman et al. (2018) aims to directly align the data distributions in the pixel space. These methods can perform very well, but they are usually *specific* to a particular kind of domain shift. For instance, Hoffman et al. (2018) assumes a similar field of view between source and target images. Both feature-alignment and image-alignment methods can be *complex* to train and *unstable* as they usually require one or several GANs Fu et al. (2024), Kamnitsas et al. (2017), Jiang et al. (2021), Liu & Tuzel (2016). Chen et al. (2019). However, image-alignment methods tend to be heavier Jiang et al. (2021). Other approaches aim to generate pseudo-labels in target domain to train a model under the supervision of these pseudo-labels Zhou et al. (2021). These methods implicitly make the hypothesis that the domain shift is limited, as they require an initialization of the network that must produce pseudo-labels of sufficient quality Lee et al. (2022).

To the best of our knowledge, there is no method that is task-adaptable, agnostic to the type of domain shift, that benefits from a rigorous theoretical background, and that achieves state-of-the-art performance.

¹<https://anonymous.4open.science/r/FARR-06B3>

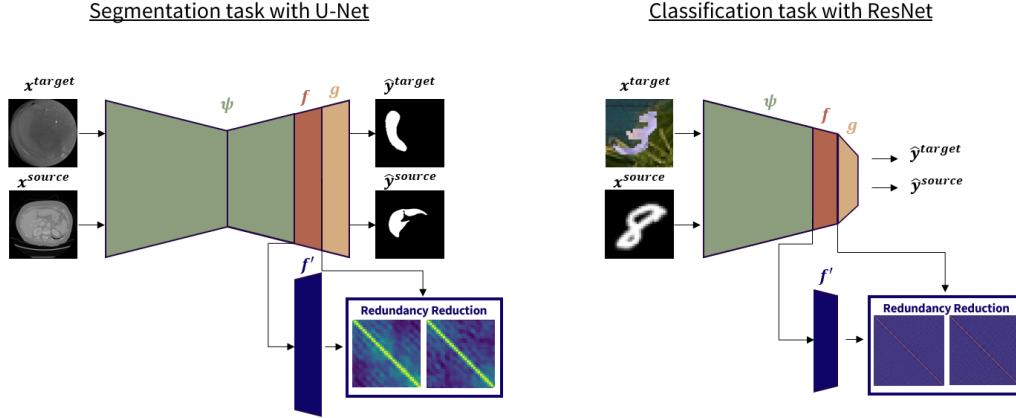


Figure 1: Our task-agnostic UDA method FARR. We decompose any architecture into a feature extractor ψ , a representation head f and its adversary f' , and a task-specific prediction head g . The features $z = \psi(x)$ are encouraged to become invariant to the domain shift between the source and target data by introducing a redundancy reduction adversarial strategy. After training, ψ extracts domain-invariant features and f' is discarded at inference.

Contributions. Our contributions are: (i) a new feature-alignment strategy based on redundancy reduction, that is task-adaptable and agnostic to the type of domain shift; (ii) a theoretical proof that our formulation effectively aligns source and target features; (iii) a comprehensive empirical evaluation across classification and segmentation tasks, using seven public and two private datasets covering diverse domain shifts. Our results show that FARR consistently outperforms existing feature-alignment methods, while remaining competitive with state-of-the-art UDA approaches across tasks and datasets.

2 RELATED WORK

Existing UDA methods can be grouped into feature-alignment, image-alignment and self-training. Feature-alignment methods learn a domain-invariant embedding space aligning source and target features. Image-alignment methods operate at the pixel level, translating target images into the source style to reduce domain shift. Self-training methods instead generate pseudo-labels in the target domain, enabling supervised training with these labels.

Feature-alignment. Early work minimizes statistical distances such as MMD Long et al. (2015) or CORAL Sun et al. (2016), using hand-crafted losses. More recent approaches rely on adversarial training: DANN Ganin et al. (2016) introduces a domain discriminator to enforce invariance, later extended to segmentation tasks with U-Net backbones Kamnitsas et al. (2017); Javanmardi et al. (2018); Panfilov et al. (2019); Fu et al. (2024). ADDA Tzeng et al. (2017) improves stability with separate source and target encoders. MDD Zhang et al. (2019) proposes a margin-based discrepancy measure with theoretical guarantees, later adapted to medical segmentation Munk et al. (2024). Beyond adversarial learning, contrastive methods have been explored, combining distributional alignment with contrastive objectives Thota et al. (2021); Avena et al. (2025), or applying slice-based contrastive pairs for OCT segmentation Gomariz et al. (2022).

Image-alignment. These methods translate images across domains to reduce pixel-level discrepancies. CoGAN Liu & Tuzel (2016) and PixelDA Bousmalis et al. (2017) use GANs to generate target-like images from source data, sometimes with feature-level adaptation. To better preserve semantic consistency during translation, cycle-consistency is often employed Hoffman et al. (2018); Chen et al. (2019; 2020); Cai et al. (2019); Huo et al. (2019); Jiang et al. (2021). Extensions such as DRANet Lee et al. (2021) further disentangle structure and style to improve transfer. While such methods can produce strong results, they usually require stronger assumptions on the type of domain shift, lack theoretical support, and involve complex architectures with multiple adversarial modules.

Self-training. These methods exploit pseudo-labels generated on target data to iteratively adapt the model Zhou et al. (2021); Zhang et al. (2021); Chen et al. (2023); Liu et al. (2024). Performance depends heavily on pseudo-label quality, often controlled by confidence thresholds. Recent improvements integrate contrastive learning and prototypes Lee et al. (2022), clustering-based consistency losses Yue et al. (2023), or additional weak labels Das et al. (2023). While robust in moderate shifts, they may fail when initial pseudo-labels are highly noisy.

3 METHOD

Driven by the previous discussion, we aim to design a method that is generic and task-adaptable, and where we make no prior assumption on the type of domain shift. Furthermore, it should be flexible, namely no constraints about the architecture, and lightweight enough to train with minimal computational resources. Eventually, to enhance feature alignment, and provide a theoretical framework to ensure domain invariance, we will take advantage of the self-supervised learning loss of Barlow Twins Zbontar et al. (2021).

We consider a source domain $\mathcal{D}_S = (\mathcal{X}^S, \mathcal{Y}, p^S(x, y))$ and a target domain $\mathcal{D}_T = (\mathcal{X}^T, \mathcal{Y}, p^T(x, y))$ with some domain shift captured by differing distributions p^S and p^T . We are given labeled samples $(x_i^S, y_i^S)_{i=1}^{N_S} \sim p^S(x, y)$ in the source domain, and unlabeled samples $(x_i^T)_{i=1}^{N_T} \sim p^T(x)$ in the target domain. The goal is to learn a map $h : \mathcal{X}^S \cup \mathcal{X}^T \rightarrow \mathcal{Y}$ that produces a consistent labeling $h(x)$ for both source and target domain samples $x \in \mathcal{X}^S \cup \mathcal{X}^T$. We proceed via a feature-alignment strategy, whereby (1) learned representations of source and target domain samples are driven to the same marginal distribution at convergence, and (2) the map h is made informative for the labeling task through explicit label supervision in the source domain. Next, we describe an adversarial strategy, that we name FARR (Feature Alignment and Redundancy Reduction), that leads to these two properties both theoretically and empirically. To derive the proposed approach, we split the predictor $h(x) \triangleq g(f(\psi(x)))$ into a feature encoder $\psi(\cdot)$, a representation head $f(\cdot)$ and a task-specific prediction head $g(\cdot)$. Example architectures for ψ, f, g are provided in Figure 1. An ablation study on splitting a ResNet-18 for classification task can be found in Appendix A.9.

Source domain label supervision. The representation head $f(\cdot)$ and the prediction head $g(\cdot)$ are trained to minimize the labeling task in the source domain via a task-specific loss $\mathcal{L}_{\text{task}}(h(x^S), y^S)$.

Adversarial feature alignment. The features $z \triangleq \psi(x)$ are encouraged to become invariant to the domain shift between the source and target domains, by introducing an adversarial strategy between $\psi(\cdot)$ and an adversary $f'(\cdot)$, where f' shares the same architecture as f . The adversary $f'(\cdot)$ is trained to produce representations as dissimilar as possible from $f(\cdot)$ when using target samples x^T , and at the same time it should result in representations as similar as possible to $f(\cdot)$ when using source samples x^S . This is achieved (equation 1) via a separation loss $\mathcal{L}_{\text{sep}}$ and an alignment loss $\mathcal{L}_{\text{align}}$, respectively, described in Section 3.1. On the other hand, the feature extractor $\psi(\cdot)$ is trained (equation 2) to output features $z \triangleq \psi(x)$ that (1) are relevant for the labeling task; and (2) that make the representations of the adversary $f'(z)$ and of the representation head $f(z)$ undistinguishable both in source and target domains:

$$\arg \min_{f'} \mathcal{L}_{\text{align}}(f(z^S), f'(z^S)) + \gamma \mathcal{L}_{\text{sep}}(f(z^T), f'(z^T)) \quad (1)$$

$$\arg \min_{\psi} \mathcal{L}_{\text{task}}(h(x^S), y^S) + \alpha \mathcal{L}_{\text{align}}(f(z^S), f'(z^S)) + \gamma \mathcal{L}_{\text{align}}(f(z^T), f'(z^T)) \quad (2)$$

where $(\alpha, \gamma) \in \mathbb{R}^+ \times \mathbb{R}^+$ are user-defined hyperparameters.

Algorithm. In practice, we alternate between single optimization steps of $\mathcal{L}_{\text{task}}$, equation 1 and equation 2 using a simple stochastic gradient descent on source and target domain batches (see Algorithm 1). We will show in Section 3.3 that, at convergence, the source and target domains are aligned in feature space *i.e.*, the marginal distributions of z^S and z^T are identical.

3.1 REDUNDANCY-REDUCTION BASED REPRESENTATION ALIGNMENT

The losses $\mathcal{L}_{\text{align}}(f(z), f'(z))$ and $\mathcal{L}_{\text{sep}}(f(z), f'(z))$ should encourage alignment and separation of the latent representations $f(z), f'(z)$, respectively. To this end, we revisit Barlow Twins' redundancy reduction mechanism Zbontar et al. (2021) and encourage the correlation matrix of representations $f(z), f'(z)$ to have specific structure conducive of these properties. Consider the representations $f(\mathbf{z}) \in \mathbb{R}^{B \times D \times H \times W}$ obtained from a batch, with $H := W := 1$ for classification and arbitrary H, W for segmentation². Let $\phi_f(\mathbf{z}) \in \mathbb{R}^{D \times N}$ the reshaped, flattened, centered and L_2 -normalized (along B, H, W) version of $f(\mathbf{z})$, and similarly $\phi_{f'}(\mathbf{z})$ for $f'(\mathbf{z})$. Following Zbontar et al. (2021), we note $\mathcal{C}[i, j] \triangleq \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle$ the cross-correlation of features $i, j \in \{1, \dots, D\}$ for representations $f(\mathbf{z}), f'(\mathbf{z})$.

Alignment loss. We define $\mathcal{L}_{\text{align}}(f(z), f'(z))$ in equation 3 to encourage perfect correlation between the representation head $f(\cdot)$ and the adversary $f'(\cdot)$ along homologous feature dimensions, and to discourage redundancy across different feature dimensions:

$$\mathcal{L}_{\text{align}}(f(z), f'(z)) \triangleq \sum_{i=1}^D (1 - \mathcal{C}[i, i])^2 + \frac{1}{D} \sum_{i \neq j} \mathcal{C}[i, j]^2 \quad (3)$$

Separation loss. On the contrary, we define $\mathcal{L}_{\text{sep}}(f(z), f'(z))$ in equation 4 to encourage perfect decorrelation between the representation head $f(\cdot)$ and the adversary $f'(\cdot)$ across all feature dimensions:

$$\mathcal{L}_{\text{sep}}(f(z), f'(z)) \triangleq \sum_{i=1}^D \mathcal{C}[i, i]^2 + \frac{1}{D} \sum_{i \neq j} \mathcal{C}[i, j]^2 \quad (4)$$

Remark. Note that $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{sep}}$ are task-agnostic: they encourage representations to match or differ without direct reference to the task of interest *i.e.*, to the prediction head $g(\cdot)$. This is in contrast to previous works, such as MDD Zhang et al. (2019), where specific losses, such as cross-entropy for classification, were used. Furthermore, by using the proposed redundancy-reduction losses, we don't need an adversary task-specific prediction head $g'(\cdot)$, as in MDD, since we move the feature alignment from the space of g to the one of f .

3.2 LEARNING ALGORITHM

Using all the notations introduced above, we provide a description of the unsupervised domain adaptation algorithm.

Algorithm 1 Proposed Unsupervised Domain Adaptation Algorithm

Require: Define hyper-parameters α, γ , learning rates, batch size B , and number epochs K
 Pre-train ψ, f and g by minimizing $\mathcal{L}_{\text{task}}$ on source domain S for K epochs
 Initialize f' with the same weights as f pre-trained on source domain S
for $k \leftarrow 1$ to K **do**
 Sample batch $\{x_i^S, y_i^S\}_{i=1}^B$ from source domain and batch $\{x_j^T\}_{j=1}^B$ from target domain
 Update f and g by minimizing $\mathcal{L}_{\text{task}}$
 Update f' by minimizing equation 1
 Update ψ by minimizing equation 2
end for

Note that f' is only used during the unsupervised domain adaptation phase and discarded after training.

²in this article, we focus on classification and segmentation but the proposed method is generic and could be applied to any supervised task, such as regression or object detection.

3.3 THEORETICAL GUARANTEES

We now study the theoretical properties of the proposed method. In particular, we obtain theoretical guarantees that, at the optimum, the feature extractor $\psi(\cdot)$ realizes the alignment of the marginal distributions of features in the source and target domains, under Assumption 1.

Consider a minibatch $\mathbf{x} \in (\mathcal{X}^S \cup \mathcal{X}^T)$, and the marginal distributions $p^S(\mathbf{x}), p^T(\mathbf{x})$ according to the source and target domain distributions. Denote $p(\mathbf{z}) \triangleq \psi_{\#} p^S(\mathbf{z})$ the pushforward of p^S through ψ , and similarly $q(\mathbf{z}) \triangleq \psi_{\#} p^T(\mathbf{z})$ the pushforward of p^T through ψ .

We start by rewriting equation 1 with the losses proposed in Section 3.1, obtaining:

$$\begin{aligned} \arg \min_{f'} \mathbb{E}_{\mathbf{z} \sim p} \left[\sum_{i=1}^D (1 - \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle)^2 + \frac{1}{D} \sum_{i=1}^D \sum_{j \neq i}^D \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle^2 \right] \\ + \gamma \mathbb{E}_{\mathbf{z} \sim q} \left[\sum_{i=1}^D \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle^2 + \frac{1}{D} \sum_{i=1}^D \sum_{j \neq i}^D \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle^2 \right] \end{aligned} \quad (5)$$

Then, we reformulate the optimization problem by substituting $w(\mathbf{z})_{i,i} = \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle$ and minimizing the resulting relaxed variational problem with respect to $w(\mathbf{z})$:

$$\begin{aligned} \arg \min_{w(\mathbf{z})} \int_{\mathbf{z}} p(\mathbf{z}) \left(\sum_{i=1}^D (1 - w(\mathbf{z})_{i,i})^2 + \frac{1}{D} \sum_{j \neq i}^D w(\mathbf{z})_{i,j}^2 \right) + \gamma q(\mathbf{z}) \left(\sum_{i=1}^D w(\mathbf{z})_{i,i}^2 + \frac{1}{D} \sum_{j \neq i}^D w(\mathbf{z})_{i,j}^2 \right) d\mathbf{z} = \\ \int_{\mathbf{z}} \sum_{i=1}^D (1 - w(\mathbf{z})_{i,i})^2 p(\mathbf{z}) + \gamma q(\mathbf{z}) w(\mathbf{z})_{i,i}^2 + \frac{1}{D} \sum_{j \neq i}^D (p(\mathbf{z}) + \gamma q(\mathbf{z})) w(\mathbf{z})_{i,j}^2 d\mathbf{z} \end{aligned} \quad (6)$$

We can notice that the integrand is separable in each $w(\mathbf{z})_{i,j}$ thus resulting in an element-wise quadratic function for each $w(\mathbf{z})_{i,j}$. It is then easy to see that the optimal matrix $w(\mathbf{z})$ is diagonal with coefficients equal to:

$$w(\mathbf{z})_{i,j} = \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle = \begin{cases} \frac{p(\mathbf{z})}{p(\mathbf{z}) + \gamma q(\mathbf{z})} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad (7)$$

The solution equation 7 of the above relaxed problem is not necessarily reachable by any f' . We thus make the following assumption on the capacity of f' :

Assumption 1 For any $\psi(\cdot)$, $f(\cdot)$, there exists $f'(\cdot)$ such that its feature maps $\phi_{f'}(\mathbf{z})_i$ satisfy the constraints in equation 7 for all $i, j \in \{1, \dots, D\}$

This assumption, akin to those made in margin-based methods like MDD which rely on expressive classifiers to estimate divergence, ensures theoretical alignment of source and target features. Importantly, as shown in Appendix A.6, our method remains empirically effective even when f' has very limited capacity, suggesting that the assumption is not overly restrictive in practice.

Lemma 1 Under Assumption 1, the adversary $f'(\cdot)$ that satisfies equation 7 is a global minimum of the minimization problem of equation 5.

Proof in Appendix A.1.1.

Theorem 1 Minimizers $\psi(\cdot)$ of the minimization problem of $\mathcal{L}_{\text{task}}$, equation 1, 2 align the marginal distributions of features in source and target domains i.e., $p(\mathbf{z}) = q(\mathbf{z})$ at optimum.

Proof in Appendix A.1.2

Combining previous Lemma and Theorem provides the following result. At optimum, we have:

$$\forall z, \forall (i, j) \in \{1, \dots, D\} \times \{1, \dots, D\}, \quad \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle = \begin{cases} \frac{1}{1+\gamma} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad (8)$$

In Figure 2, we will show that this theoretical result is experimentally verified.

3.4 COMPARISON WITH MDD ZHANG ET AL. (2019)

Our method shares some architectural similarities with the MDD method proposed in Zhang et al. (2019). Both methods use a feature extractor and two heads, f and f' . However, in Zhang et al. (2019), both networks f and f' have also a task-specific head and the method, originally conceived for classification only, aims at aligning the predicted probabilities of the task-specific heads using a cross-entropy loss. Contrary to that, the adversarial game we propose here concerns the correlation of the features extracted by f and f' , which is thus independent on the task to perform. As a consequence, our method can be theoretically adapted to any architecture and task.

4 EXPERIMENTS

4.1 EXPERIMENTAL DETAILS

We perform three experiments to assess the performance of our method and compare it with several SOTA approaches. All experiments were conducted on an NVIDIA V100 GPU with 32GB memory. We demonstrate our approach on two tasks: classification and segmentation. A key advantage of our method over other SOTA approaches is that we make no assumptions about the type of domain shift. Accordingly, we evaluate it on different types of shifts: intensity variations (CT $\rightarrow$ MR, MNIST $\rightarrow$ MNIST-M, MNIST $\rightarrow$ USPS, USPS $\rightarrow$ MNIST) and geometric variations (CT $\rightarrow$ CBCT, DRIVE $\rightarrow$ CHASEDB1). Examples of these shifts across source and target datasets are provided in Appendix A.2. Further implementation details for reproducibility are given in Appendix A.8. Full details on all datasets used are provided in Appendix A.7.

Digit classification. We conduct UDA experiments on three digit datasets: MNIST (70,000 images), MNIST-M (149,002 images), and USPS (9,298 images). These datasets differ in color, texture, and background style, leading to domain shifts. We evaluate adaptation performance on MNIST $\rightarrow$ MNIST-M, USPS $\rightarrow$ MNIST, and MNIST $\rightarrow$ USPS.

Retinal vasculature segmentation. We evaluate on DRIVE (40 labeled images, source) and CHASEDB1 (28 labeled images, target). The datasets differ in field of view, optic disc localization, and vessel contrast.

Liver segmentation. We further assess our method on cross-modality liver segmentation. We consider two settings: CT $\rightarrow$ CBCT using a private dataset (15,827 CT slices, 13,024 CBCT slices) and CT $\rightarrow$ MR using LiTS (5,324 CT slices) as source and CHAOS (647 MR slices) as target. These settings capture both geometric (CT $\rightarrow$ CBCT) and intensity (CT $\rightarrow$ MR) shifts. All segmentation experiments are performed on 2D axial slices resampled to a common pixel size.

4.2 RESULTS

Digit classification. As showed in Table 1, we outperform all similar adversarial feature-alignment methods such as DANN Ganin et al. (2016) or ADDA Tzeng et al. (2017) by a large margin and contrastive approaches for feature-alignment such as Thota et al. (2021) or Avena et al. (2025). When the domain shift does not rely on geometric differences between source and target domains, as in MNIST $\rightarrow$ MNIST-M experiment, image alignment methods such as pixelDA Bousmalis et al. (2017) or DRANET Lee et al. (2021) can be effective. However, this hypothesis strongly restricts the use of such methods in unsupervised domain adaptation. Furthermore, they are often complex to implement as they require to train a generator, a discriminator, and a classifier. Our method achieves 98.6% on USPS $\rightarrow$ MNIST, outperforming the image-alignment DRANet baseline

Table 1: Performance evaluation of domain adaptation methods for digit classification using MNIST (M), MNIST-M (M-M), USPS (U) datasets. We indicate the best performance in **boldface**, while the second-best is underlined.

Type	Method	Accuracy $\uparrow$			Δ parameters $\downarrow$
		M $\rightarrow$ M-M	M $\rightarrow$ U	U $\rightarrow$ M	
	Source Only	0.625	0.802	0.449	0%
Feature Alignment	DANN Ganin et al. (2016)	0.851	0.851	0.730	+4%
	MMD Long et al. (2015)	0.769	-	-	+0%
	DSN Bousmalis et al. (2016)	0.832	0.913	-	+284%
	ADDA Tzeng et al. (2017)	0.894	0.901	0.952	+116%
	CDA Thota et al. (2021)	0.602	0.942	-	+0%
	SCoDA Avena et al. (2025)	0.851	0.966	-	+0%
	FARR (ours)	<u>0.982</u>	<u>0.980</u>	0.986	+42%
Image Alignment	CoGAN Liu & Tuzel (2016)	0.620	0.912	0.891	+1819%
	pixelDA Bousmalis et al. (2017)	0.982	0.959	-	+129%
	CyCADA Hoffman et al. (2018)	0.921	0.956	0.965	+1085%
	DRANET Lee et al. (2021)	0.987	0.982	<u>0.978</u>	+95%
	Target Only	0.962 ^a	0.978	0.991	0%

^aUDA methods can occasionally outperform target-only supervised training, as already reported in Lee et al. (2021)

(97.8%) by 0.8 percentage points. This notable gain appears on a UDA setting where the source domain has lower resolution and reduced style diversity, suggesting that our feature-level alignment method can be particularly effective under such conditions. The Δ parameters metric computes the proportion of parameters that is added for domain adaptation with respect to the original classification network. Image-alignment methods, such as CyCADA Zhu et al. (2017) or CoGAN Liu & Tuzel (2016), are composed of heavy blocks for domain adaptation. In our case, we only add **42%** of the parameters, as the adversary is only composed of the last layer block of the final layer group of the ResNet (as detailed in Appendix A.8), which is much less than all other image-alignment methods. In the context of digit classification, our method achieves **SOTA performance** across diverse domain shifts, while remaining **lightweight** compared to existing SOTA approaches. In Appendix A.10, we further demonstrate that our method can be adapted to transformer architectures in the context of digit classification.

In our ablation study of hyperparameters in Appendix A.4, we evaluate the robustness of our method with respect to the weighting parameters α and γ on MNIST $\rightarrow$ MNIST-M. Although the best performance is achieved around $\alpha \approx 0.8$ and $\gamma \approx 0.05$ as described in Appendix A.8, we observe comparable results across a broad range, with $\alpha \in [0, 1]$ and $\gamma \in [0, 1]$. The α and γ parameters mentioned were determined through hyperparameter search using Optuna Akiba et al. (2019). This indicates that the method is not overly sensitive to the precise choice of these hyperparameters.

Retinal vasculature segmentation. In Table 2, we show that our method outperforms all other SOTA methods both in AUC and F1 scores. Furthermore, we only need to add +2% parameters of the original U-Net segmentation model for the domain adaptation. Our method can thus effectively align features from source and target domains, even when they have important geometric and appearance differences. These findings confirm the **adaptability** of our method, demonstrating its successful extension to a **segmentation task** and its integration with a **distinct architecture**, namely U-Net.

Liver segmentation. Finally, in Table 3, we show that our method performs well also in the context of 2D liver segmentation with two different types of domain shift. It outperforms all other SOTA UDA methods, both based on feature alignment or self-training, by a wide margin. We also compare our method against a foundation model, SAM-Med 2D Cheng et al. (2023), in a zero-shot setting. SAM-Med 2D was trained on MR and CT images but not on CBCT data. This explains why it outperforms our method in CT $\rightarrow$ MR UDA but not on the CT $\rightarrow$ CBCT UDA problem, where its performance drops.

Table 2: Performance evaluation of domain adaptation methods for segmentation. Source: DRIVE. Target: CHASEDB1. We indicate the best performance in **boldface**, while the second-best is underlined.

Type	Method	AUC $\uparrow$	F1 $\uparrow$	Δ parameters $\downarrow$
Self-Training	SGL Zhou et al. (2021)	0.978	0.718	+100 %
	ProDA Zhang et al. (2021)	0.971	0.699	+102 %
	FR-UNet Liu et al. (2022)	0.978	0.736	+0 %
	CDCL Chen et al. (2023)	0.979	0.739	-
	MILAL-PDSF Liu et al. (2024)	<u>0.981</u>	<u>0.740</u>	-
Feature Alignment	DisClusterDA Tang et al. (2022)	0.977	0.737	+0 %
	FARR (ours)	0.982	0.753	<u>+2</u> %
	Target Only	0.989	0.842	

Table 3: Performance evaluation of domain adaptation methods for segmentation on two experiments. Source: CT, Target: CBCT or MR. We indicate the best performance in **boldface**, while the second-best is underlined.

Type	Method	F1 score $\uparrow$	
		CT $\rightarrow$ CBCT	CT $\rightarrow$ MR
	Source Only	0.541	0.548
Self-Training	BDCL Lee et al. (2022)	0.600	-
Feature Alignment	DANN Fu et al. (2024)	0.683	0.661
	MDD Munk et al. (2024)	<u>0.700</u>	0.681
	FARR (ours)	0.751	<u>0.728</u>
Zero-shot	SAM-MED 2D Cheng et al. (2023)	0.461	0.812^a
	Target Only	0.903	0.868

^aThis model is trained fully supervised with MR labels

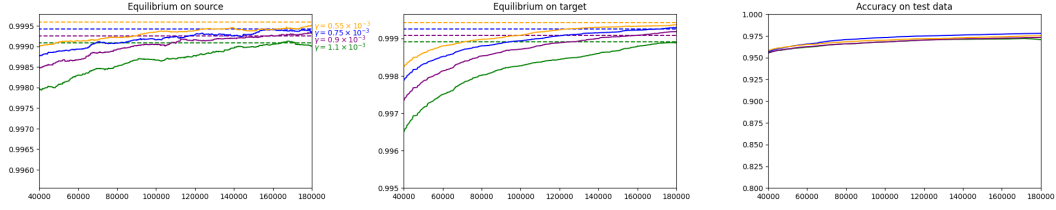


Figure 2: Convergence with respect to γ parameter on MNIST $\rightarrow$ MNIST-M. The two plots on the left show the mean correlation values of $\phi_f(\mathbf{z})_i$ and $\phi_{f'}(\mathbf{z})_i$ (y axis) over i with respect to the training iteration number (x axis). The plot on the right shows the accuracy on test data from target domain (y axis) with respect to the training iteration number (x axis).

Experimental confirmation of theoretical results. As proved in equation 9, at convergence, the correlation of embeddings $\phi_f(\mathbf{z})_i$ and $\phi_{f'}(\mathbf{z})_j$, when $i = j$, should converge to $\frac{1}{1+\gamma}$. In Figure 2, we show that this theoretical result is verified experimentally. We trained 4 models using different values of γ (0.55×10^{-3} , 0.75×10^{-3} , 0.9×10^{-3} , 1.1×10^{-3}) while fixing the other hyper-parameters (α , batch size, learning rates) to the same values. As showed in Figure 2, we reach the equilibrium expected from equation 9 for both source and target domains. Indeed, equilibrium on source (resp. target) shows the average value of the correlation of $\phi_f(\mathbf{z})_i$ and $\phi_{f'}(\mathbf{z})_i$ over i , for features extracted from source (resp. target) images. These values are computed for each training iteration, with a batch size of 32. For each γ , these mean correlation values actually converge to $\frac{1}{1+\gamma}$. Further experiments about the cross-correlation matrix $\mathcal{C}[i, j]$ and representations spaces can be found in Appendix A.5 and Appendix A.3.

5 CONCLUSION

We proposed a novel feature-alignment approach for UDA based on redundancy reduction, and demonstrated its performance through qualitative and quantitative experiments on various datasets, with various types of domain shift. Our method is lightweight, and can be theoretically adapted to any architecture. It outperforms other state-of-the-art UDA approaches belonging to the same category on different datasets and tasks. Furthermore, the method is generic, makes no assumptions on the supervised task or on the type of domain shift, and it has reasonable theoretical guarantees validated experimentally.

Limitations. The results we provided on classification and segmentation tasks are appealing, but they were only obtained using 2D datasets. In medical image segmentation, using the entire 3D volumes usually improves the results. Nevertheless, we believe that our method could be easily adapted to 3D scans, using a 3D UNet for the segmentation tasks, for instance. Another limitation of this article is the lack of experiment on regression task. While our method is task-agnostic by design, current experiments are limited to classification and segmentation. Evaluating on regression tasks would further validate its generality.

Future directions. As our method makes no hypothesis about the task, we will investigate its performance in other tasks such as regression or object detection. Another research path might be the extension of this UDA method to the semi-supervised setting. Indeed, in some context, it is possible to benefit from a few number of labeled data in target domain. Eventually, we also plan to extend the proposed method to a Domain Generalization setting, leveraging multiple source domains.

Reproducibility statement. We ensure reproducibility by releasing the full codebase for training and evaluation, accompanied by detailed instructions³. All theoretical results and proofs are provided in the paper Section 3.3 and Appendix A.1. Comprehensive training procedures, hyperparameter configurations, and optimization details for all experiments are reported in Appendix A.8. The majority of datasets used are publicly available and described in Appendix A.7; two datasets are private but are documented in Appendix A.8.3 with all pre-processing and augmentation details. In addition, we demonstrate in Appendix A.4 that our method is robust to hyperparameter variations, which further facilitates reproducibility in practice.

³<https://anonymous.4open.science/r/FARR-06B3>

REFERENCES

- Ganin, Y. & Ustinova, E. & Ajakan, H. & Germain, P. & Larochelle, H. Laviolette, F. & March, M. & Lempitsky, V. (2016) Domain-Adversarial Training of Neural Networks, *Journal of Machine Learning Research* 17, pp. 1-35
- Kamnitsas, K. & Baumgartner, C. & Ledig, C. & Newcombe, V. & Simpson, J. & Kane, A. & Menon, D. & Nori, A. & Criminisi, A. & Rueckert, D. & Glocker, B. (2017) Unsupervised Domain Adaptation in Brain Lesion Segmentation with Adversarial Networks, *Information Processing in Medical Imaging. IPMI 2017. Lecture Notes in Computer Science*, vol 10265. Springer, pp. 597–609
- Javanmardi, M. & Tasdizen T. (2018) Domain Adaptation for biomedical image segmentation using adversarial training, *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 2018* pp. 554-558
- Panfilov, E. & Tiulpin, A. & Klein, S. & Nieminen, M. T. & Saarakkala, S. (2019) Improving Robustness of Deep Learning Based Knee MRI Segmentation: Mixup and Adversarial Domain Adaptation, *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019* pp. 450-459
- Fu, J. & Bendazzoli, S. & Smedby, Ö. & Moreno, R. (2024) Unsupervised Domain Adaptation for Pediatric Brain Tumor Segmentation, *MICCAI Workshop on Advancing Data Solutions in Medical Imaging AI*
- Isensee, F. & Jaeger, P. F. & Kohl, S. A. & Petersen, J. & Maier-Hein, K. H (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, *Nat Methods* 18 pp. 203–211
- Long, M. & Cao, Y. & Wang, J. & Jordan, M. I. (2015) Learning Transferable Features with Deep Adaptation Networks, *Proceedings of the 32nd International Conference on Machine Learning, PMLR* 37 pp. 97-105
- Saito, K. & Watanabe, K. & Ushiku, Y. & Harada, T. (2018) Maximum Classifier Discrepancy for Unsupervised Domain Adaptation, *IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018* pp. 3723-3732
- Zhang, Y. & Liu, T. & Long, M. & Jordan, M. I. (2019) Bridging Theory and Algorithm for Domain Adaptation, *Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR* 97, 2019
- Munk, A. & Ma, A. & Nielsen, M. (2024) MDD-UNet: Domain Adaptation for Medical Image Segmentation with Theoretical Guarantees, a Proof of Concept, *Proceedings of the 5th Northern Lights Deep Learning Conference (NLDL), PMLR* 233, 2024
- Hoffman, J. & Tzeng, E. & Park, T. & Zhu, J.-Y. & Isola, P. & Saenko, K. & Efros, A. A. & Darrell, T. (2018) CyCADA: Cycle-consistent adversarial domain adaptation, *Proceedings of the 35th International Conference on Machine Learning, PMLR* 80, pp. 1989–1998.
- Chen, C. & Dou, Q. & Chen, H. & Qin, J. & Heng, P.-A. (2019) Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation, *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pp. 865–872.
- Chen, C. & Dou Q. & Chen, H. & Qin, J. & Heng, P.-A. (2020). Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4066–4075.
- Cai, J. & Zhang, Z. & Cui, L. & Zheng, Y. & Yang, L. (2019). Towards cross-modal organ translation and segmentation: A cycle- and shape-consistent generative adversarial network, *Medical Image Analysis*, 58,
- Huo, Y. & Xu, Z. & Moon, H. & Bao, S. & Assad, A. & Moyo, T. K. & Savona, M. R. & Abramson, R. G. & Landman, B. A. (2019) SynSeg-Net: Synthetic segmentation without target modality ground truth, *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*
- Jiang, J. & Hu, Y. C. & Tyagi, N. & Rimner, A. & Lee, N. & Deasy, J. O. & Berry, S. & Veeraraghavan, H. (2021) PSIGAN: Joint probabilistic segmentation and image distribution matching for unpaired cross-modality adaptation based MRI segmentation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1485–1494
- Lee, G. & Eom, C. & Lee, W. & Park, H. & Ham, B. (2022) Bi-directional contrastive learning for domain adaptive semantic segmentation, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 7238–7257.

540 Zbontar, J. & Jing, L. & Misra, I. & LeCun, Y. & Deny, S. (2021) Barlow twins: Self-supervised learning via
541 redundancy reduction. *Proceedings of the 38th International Conference on Machine Learning, PMLR 139*,
542 pp. 12310–12320.

543 Mansour, Y. & Mohri, M. & Rostamizadeh, A. (2009) Domain adaptation: Learning bounds and algorithms,
544 *Proceedings of the 22nd Annual Conference on Learning Theory*, pp. 1–16.

545 Bilic, P. & Christ, P. & Li, H. B. & Vorontsov, E. (2023). The Liver Tumor Segmentation Benchmark (LiTS),
546 *Medical Image Analysis*, 84, 102680.

547 Kavur, A. E. & Gezer, N. S. & Barış, M. & Aslan, S. & Conze, P.-H. & Groza, V. & Pham, D. D. & Chatterjee,
548 S. & Ernst, P. & Özkan, S. & Baydar, B. & Lachinov, D. & Han, S. & Pauli, J. & Isensee, F. & Perkonig, M.
549 & Sathish, R. & Rajan, R. & Sheet, D. & Dovletov, G. & Speck, O. & Nürnberger, A. & Maier-Hein, K. H. &
550 Bozdağı Akar, G. & Ünal, G. & Dicle, O. & Selver, M. A. (2021) CHAOS challenge—Combined (CT-MR)
551 healthy abdominal organ segmentation, *Medical Image Analysis*, 69

552 Wasserthal, J. & Breit, H.-C. & Meyer, M. T. & Pradella, M. & Hinck, D. & Sauter, A. W. & Heye, T. & Boll, D.
553 & Cyriac, J. & Yang, S. & Bach, M. & Segeroth, M. (2023) TotalSegmentator: Robust segmentation of 104
554 anatomical structures in CT images, *Radiology: Artificial Intelligence*, 5

555 Cheng, J. & Jin, Y. & Dong, Z. & Chen, J. & Liu, T. & Wang, H. & Shen, H. & Han, J. & Zhang, Z. & He, Z. &
556 Zhang, M. & Zhang, Y. & Yu, Q. (2023) SAM-Med2D

557 Barbano, C. A. & Dufumier, B. & Tartaglione, E. & Grangetto, M. & Gori, P. (2023) Unbiased supervised
558 contrastive learning, *The Eleventh International Conference on Learning Representations*

559 Sun., B. & Feng, J. & Saenko, K. (2016) Return of Frustratingly Easy Domain Adaptation, *Proceedings of the*
560 *AAAI Conference on Artificial Intelligence (Vol. 30, No. 1)*

561 Zhou, Y. Q. & Yu, H.C. et al. (2021) Study group learning: improving retinal vessel segmentation trained
562 with noisy labels, *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th*
563 *International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*,
564 *Springer (2021)* pp. 57–67

565 Zhang, P. & Zhang, B. & Zhang, T. & Chen, D. & Wang, Y. & Wen, F. (2021) Prototypical pseudo label denoising
566 and target structure learning for domain adaptive semantic segmentation, *Proceedings of the IEEE/CVF*
567 *Conference on Computer Vision and Pattern Recognition (CVPR)*.

568 Liu, W.T. & Yang, H.H. & Tian, T. et al. (2022) Full-resolution network and dual-threshold iteration for retinal
569 vessel and coronary angiograph segmentation, *IEEE Journal of Biomedical and Health Informatics*, 26 (2022)
570 pp. 4623–4634

571 Tang, H. & Wang, Y.W. et al. (2022) Unsupervised domain adaptation via distilled discriminative clustering,
572 *Pattern Recogn.*, 127 (2023), Article 108638.

573 Chen, S. & Liu, Z. & Zhang, Y. & Yang, Y. (2023) Cross-domain contrastive learning for unsupervised domain
574 adaptation, *IEEE Transactions on Medical Imaging* pp. 1665–1673

575 Liu, J. & Zhao, J. & Xiao, J. & Zhao, G. & Xu, P. & Yang, Y. & Gong, S. (2024) Unsupervised domain adaptation
576 multi-level adversarial learning-based crossing-domain retinal vessel segmentation, *Computers in Biology*
577 *and Medicine*

578 Staal, J. & Abràmoff, M.D. & Niemeijer, M. et al. (2004) Ridge-based vessel segmentation in color images of
579 the retina; *IEEE Trans. Med. Imag.*, 23 (4) (2004) pp. 501–509

580 Fraz, M.M. & Remagnino, P. & Hoppe, A. et al. (2004) An ensemble classification-based approach applied to
581 retinal blood vessel segmentation, *IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng.*, 59 (9) (2012) pp.
582 2538–2548

583 Bousmalis, K. & Trigeorgis, G. & Silberman, N. & Krishnan, D. & Erhan, D. (2016) Domain separation
584 networks, *Advances in Neural Information Processing Systems (NIPS)*, pp. 343–351

585 Tzeng, E. & Hoffman, J. & Saenko, K. & Darrell, T. (2017) Adversarial discriminative domain adaptation, *Proc.*
586 *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 7167–7176

587 Liu, M.-Y. & Tuzel, O. (2016) Coupled generative adversarial networks, *Advances in Neural Information*
588 *Processing Systems (NIPS)*, pp. 469–477

594 Bousmalis, K. & Silberman, N. & Dohan, D. & Erhan, D. & Krishnan, D. (2017) Unsupervised pixel-level
595 domain adaptation with generative adversarial networks, *Proceedings of IEEE Conference on Computer
596 Vision and Pattern Recognition (CVPR)*, pp. 3722–3731.

597
598 Lee, S. & Cho, S. & Im, S. (2021) DRANet: Disentangling Representation and Adaptation Networks for
599 Unsupervised Cross-Domain Adaptation, *IEEE/CVF Conference on Computer Vision and Pattern Recognition
600 (CVPR), Nashville, TN, USA, 2021* pp. 15247–15256

601
602 Ye, S. & Wu, S. & Zhou, S. & Yang, S. & Tan, S. H. & Xu, S. & Song, S. & Bao, S. & Ma, S. (2020) Light-weight
603 calibrator: a separable component for unsupervised domain adaptation, *Proceedings of IEEE Conference on
604 Computer Vision and Pattern Recognition (CVPR)* pp. 13736–13745

605
606 Zhu, J.-Y. & Park, T. & Isola, P. & Efros, A. A. (2017) Unpaired image-to-image translation using cycleconsis-
607 tent adversarial networks, *Proceedings of IEEE International Conference on Computer Vision (ICCV)* pp.
608 2223–2232

609
610 Thota, M. & Leontidis, G. (2021) Contrastive Domain Adaptation, *Proceedings of the IEEE/CVF Conference on
611 computer vision and pattern recognition* pp. 2209–2218

612
613 Avena, Vinicius & Couto, Rodrigo S. & Costa, Luís Henrique M. K. & Da Silva, Eduardo A. B. (2025)
614 Bridging Domain Shifts Through Self-Contrastive Learning And Distribution Alignment, *IEEE International
615 Conference on Image Processing (ICIP)* pp. 2181–2186

616
617 Gomariz, A. & Lu, H. & Li, Y. Y. & Albrecht, T. & Maunz, A. & Benmansour, F. & Valcarcel, A. M. &
618 Luu, J. & Ferrara, D., & Goksel, O. (2022) Unsupervised domain adaptation with contrastive learning
619 for OCT segmentation, *Proceedings of the International Conference on Medical Image Computing and
620 Computer-Assisted Intervention (MICCAI)* pp. 34

621
622 Das, A. & Xian, Y. & Dai, D. & Schiele, B. (2023) Weakly-Supervised Domain Adaptive Semantic Segmentation
623 With Prototypical Contrastive Learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and
624 Pattern Recognition (CVPR)* pp. 15434–15443

625
626 Yue, Z. & Sun, Q. & Zhang, H. (2023) Make the U in UDA Matter: Invariant Consistency Learning for
627 Unsupervised Domain Adaptation, *Proceedings of the 37th Conference on Neural Information Processing
628 Systems (NeurIPS)* pp. 1–20

629
630 Akiba, T. & Sano, S. & Yanase, T. & Ohta, T. & Koyama, M. (2019) Optuna: A Next-generation Hyperparameter
631 Optimization Framework, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge
632 Discovery and Data Mining* pp. 2623–2631

633
634 LeCun, Y. & Bottou, L. & Bengio, Y. & Haffner, P. (1998) Gradient-based learning applied to document
635 recognition, *Proceedings of the IEEE*, 86(11), pp. 2278–2324

636
637 Denker, J. & Gardner, W. & Graf, H. & Henderson, D. & Howard, R. (1989) Neural network recognizer for
638 handwritten zip code digits, *Proceedings of the IEEE International Conference on Neural Networks*, 1, pp.
639 347–352

640
641 Staal, J. & Abramoff, M. D. & Niemeijer, M. & Viergever, M. A. & van Ginneken, B. (2004) Ridge-based vessel
642 segmentation in color images of the retina, *IEEE Transactions on Medical Imaging*, 23(4), pp. 501–509

643
644 Hunter, A. & Steel, D. (2013) The CHASE-DB1 dataset: Retinal images for vessel segmentation, *The British
645 Journal of Ophthalmology*, 97(5), pp. 632–637

646
647 Kavur, A. E. & et al. (2021) CHAOS challenge – Combined (CT-MR) healthy abdominal organ segmentation,
arXiv preprint

Bilic, P. & Christ, P. F. & et al. (2019) The Liver Tumor Segmentation Benchmark (LiTS), *arXiv preprint*

Oquab, M. & Darcet, T. & Moutakanni, T. & Vo, H. & Szafraniec, M. & Khalidov, V. & Fernandez, P. & Haziza,
D. & Massa, F. & El-Nouby, A. & Assran, M. & Ballas, N. & Howes, R. & Huang, P.-Y. & Li, S.-W. & Misra,
I. & Rabbat, M. & Sharma, V. & Synnaeve, G. & Xu, H. & Jegou, H. & Mairal, J. & Labatut, P. & Joulin, A.
& Bojanowski, P. (2023) DINOv2: Learning Robust Visual Features without Supervision, *arXiv preprint*

A APPENDIX

A.1 PROOF OF THE THEORETICAL GUARANTEES

A.1.1 PROOF OF LEMMA 1

In this subsection, we provide the proof of the Lemma 1 of Section 3.3. First, let's recall equation 5 and equation 7 of the paper.

$$\begin{aligned} \arg \min_{f'} \mathbb{E}_{\mathbf{z} \sim p} & \left[\sum_{i=1}^D (1 - \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle)^2 + \frac{1}{D} \sum_{i=1}^D \sum_{j \neq i} \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle^2 \right] \\ & + \gamma \mathbb{E}_{\mathbf{z} \sim q} \left[\sum_{i=1}^D \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle^2 + \frac{1}{D} \sum_{i=1}^D \sum_{j \neq i} \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle^2 \right] \end{aligned} \quad (5)$$

$$\langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle = \begin{cases} \frac{p(\mathbf{z})}{p(\mathbf{z}) + \gamma q(\mathbf{z})} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad (7)$$

Assumption 1 For any $\psi(\cdot)$, $f(\cdot)$, there exists $f'(\cdot)$ such that its feature maps $\phi_{f'}(\mathbf{z})_i$ satisfy the constraints in equation 7 for all $i, j \in \{1, \dots, D\}$

Lemma 1 Under Assumption 1, the adversary $f'(\cdot)$ that satisfies equation 7 is a global minimum of the minimization problem of equation 5.

Proof.

$$\begin{aligned} \forall f', \quad \mathbb{E}_{\mathbf{z} \sim p} & \left[\sum_{i=1}^D (1 - \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle)^2 + \frac{1}{D} \sum_{i=1}^D \sum_{j \neq i} \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle^2 \right] \\ & + \gamma \mathbb{E}_{\mathbf{z} \sim q} \left[\sum_{i=1}^D \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle^2 + \frac{1}{D} \sum_{i=1}^D \sum_{j \neq i} \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle^2 \right] \\ & \geq \sum_{i=1}^D \int_{\mathbf{z}} p(\mathbf{z}) (1 - \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle)^2 + \gamma q(\mathbf{z}) \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle^2 d\mathbf{z} \\ & = \sum_{i=1}^D \int_{\mathbf{z}} (p(\mathbf{z}) + \gamma q(\mathbf{z})) \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle^2 - 2p(\mathbf{z}) \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle + p(\mathbf{z}) d\mathbf{z} \\ & \geq \sum_{i=1}^D \int_{\mathbf{z}} p(\mathbf{z}) - \frac{(-2p(\mathbf{z}))^2}{4(p(\mathbf{z}) + \gamma q(\mathbf{z}))} d\mathbf{z} \\ & = \sum_{i=1}^D \int_{\mathbf{z}} \frac{\gamma p(\mathbf{z}) q(\mathbf{z})}{p(\mathbf{z}) + \gamma q(\mathbf{z})} d\mathbf{z} \end{aligned} \quad (9)$$

Indeed, the minimum of a quadratic polynomial $aX^2 + bX + c$ with $a > 0$ is $c - \frac{b^2}{4a}$, and we actually have: $\forall \mathbf{z}, \forall \gamma > 0, p(\mathbf{z}) + \gamma q(\mathbf{z}) > 0$.

If this lower bound is reached by some $f'(\cdot)$, then it is a global minima of the minimization problem of equation 5.

Under Assumption 1, let's consider an adversary $f'(\cdot)$ that satisfies equation 7. We have:

$$\begin{aligned}
& \mathbb{E}_{\mathbf{z} \sim p} \left[\sum_{i=1}^D (1 - \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle)^2 + \frac{1}{D} \sum_{i=1}^D \sum_{j \neq i}^D \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle^2 \right] \\
& + \gamma \mathbb{E}_{\mathbf{z} \sim q} \left[\sum_{i=1}^D \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle^2 + \frac{1}{D} \sum_{i=1}^D \sum_{j \neq i}^D \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle^2 \right] \\
& = \sum_{i=1}^D \int_{\mathbf{z}} p(\mathbf{z}) \left(1 - \frac{p(\mathbf{z})}{p(\mathbf{z}) + \gamma q(\mathbf{z})}\right)^2 + \gamma q(\mathbf{z}) \left(\frac{p(\mathbf{z})}{p(\mathbf{z}) + \gamma q(\mathbf{z})}\right)^2 d\mathbf{z} \\
& = \sum_{i=1}^D \int_{\mathbf{z}} \frac{\gamma p(\mathbf{z}) q(\mathbf{z})}{p(\mathbf{z}) + \gamma q(\mathbf{z})} d\mathbf{z}
\end{aligned} \tag{10}$$

From equation 9 and equation 10, we can conclude that the adversary $f'(\cdot)$ that satisfies equation 7 is a global minimizer of the minimization problem of equation 5.

A.1.2 PROOF OF THEOREM 1

In this subsection, we provide a proof of the Theorem 1 of Section 3.3. First, let's recall 1, 2.

$$\arg \min_{f'} \mathcal{L}_{\text{align}}(f(z^S), f'(z^S)) + \gamma \mathcal{L}_{\text{sep}}(f(z^T), f'(z^T)) \tag{1}$$

$$\arg \min_{\psi} \mathcal{L}_{\text{task}}(h(x^S), y^S) + \alpha \mathcal{L}_{\text{align}}(f(z^S), f'(z^S)) + \gamma \mathcal{L}_{\text{align}}(f(z^T), f'(z^T)) \tag{2}$$

Theorem 1 Minimizers $\psi(\cdot)$ of the minimization problem of $\mathcal{L}_{\text{task}}$, equation 1, 2 align the marginal distributions of features in source and target domains i.e., $p(\mathbf{z}) = q(\mathbf{z})$ at optimum.

Proof.

We suppose that the term $\mathcal{L}_{\text{task}}(h(x^S), y^S)$ is negligible. The task error on source domain is supposed to remain low during UDA with no restriction for the choice of ψ .

We are interested in the following minimization problem:

$$\begin{aligned}
& \arg \min_{\psi} \alpha \mathbb{E}_{\mathbf{z} \sim p} \left[\sum_{i=1}^D (1 - \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle)^2 + \frac{1}{D} \sum_{i=1}^D \sum_{j \neq i}^D \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle^2 \right] \\
& + \gamma \mathbb{E}_{\mathbf{z} \sim q} \left[\sum_{i=1}^D (1 - \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_i \rangle)^2 + \frac{1}{D} \sum_{i=1}^D \sum_{j \neq i}^D \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle^2 \right]
\end{aligned} \tag{11}$$

As $f'(\cdot)$ minimizes the optimization problem of equation 1, it satisfies equation 7. Adding the constraint $\int_{\mathbf{z}} q(\mathbf{z}) d\mathbf{z} = 1$ to ensure that q is a density, the optimization problem becomes

$$\arg \min_{\psi} \sum_{i=1}^D \int_{\mathbf{z}} \left[(\alpha p(\mathbf{z}) + \gamma q(\mathbf{z})) \left(1 - \frac{p(\mathbf{z})}{p(\mathbf{z}) + \gamma q(\mathbf{z})}\right)^2 \right] - \lambda \left(\int_{\mathbf{z}} q(\mathbf{z}) d\mathbf{z} - 1 \right) \tag{12}$$

At optimum, the following conditions must be verified:

$$\begin{aligned}
& \begin{cases} \gamma \left(1 - \frac{p(\mathbf{z})}{p(\mathbf{z}) + \gamma q(\mathbf{z})}\right)^2 + 2(\alpha p(\mathbf{z}) + \gamma q(\mathbf{z})) \frac{\gamma p(\mathbf{z})}{(p(\mathbf{z}) + \gamma q(\mathbf{z}))^2} \left(1 - \frac{p(\mathbf{z})}{p(\mathbf{z}) + \gamma q(\mathbf{z})}\right) - \lambda = 0 \\ \int_{\mathbf{z}} q(\mathbf{z}) d\mathbf{z} = 1 \end{cases} \\
& \Leftrightarrow \begin{cases} -\lambda w^3 + (2\alpha\gamma - 3\lambda)w^2 + 3(\gamma - \lambda)w + (\gamma - \lambda) = 0 \\ \int_{\mathbf{z}} q(\mathbf{z}) d\mathbf{z} = 1 \end{cases} \quad \text{with } w = \frac{p(\mathbf{z})}{\gamma q(\mathbf{z})}
\end{aligned} \tag{13}$$

The polynomial of the first condition admits a solution. There exists λ, w such that all the conditions are verified. In this case, we have two densities of probability q and p such that: $\forall z, p(z) = \gamma w q(z)$. This implies $p(z) = q(z)$. Therefore, there exists λ such that the conditions of equation 13 are fulfilled, leading to $p(z) = q(z)$.

This proves that the minimizers of $\mathcal{L}_{\text{task}}$, equation 1, 2 align the marginal distributions of features in source and target domains *i.e.*, $p(z) = q(z)$ at optimum.

A.2 TYPES OF DOMAIN SHIFTS

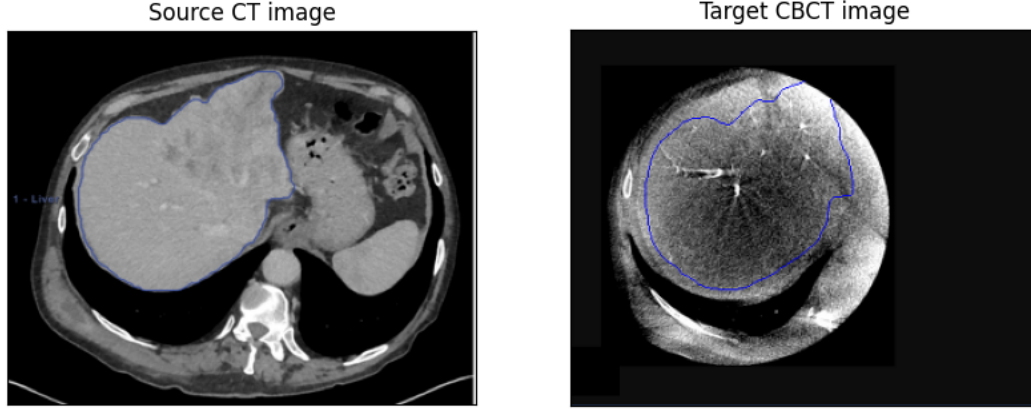


Figure 3: Left: Axial slice from abdominal CT (source domain). Right: Axial slice from abdominal CBCT (target domain) with intra-arterial injection of the same patient, with the same visualization window. Both come a from private dataset.

In Figure 3, we display CT (source domain) and CBCT (target domain) slices that were used in the UDA experiment for liver segmentation task. The images of this Figure are aligned for illustration purpose, but we used only unpaired slices in our UDA experiment. The blue border in each image demarcates the liver area. The corresponding quantitative results of this experiment are presented in the Table 3 of our paper. CBCT is different from the traditional CT as it entails a limited reconstructed field of view, specific artifacts and an intra-arterial injection of contrast medium. The domain shift between these two modalities is consequently composed of intensity and geometry differences.

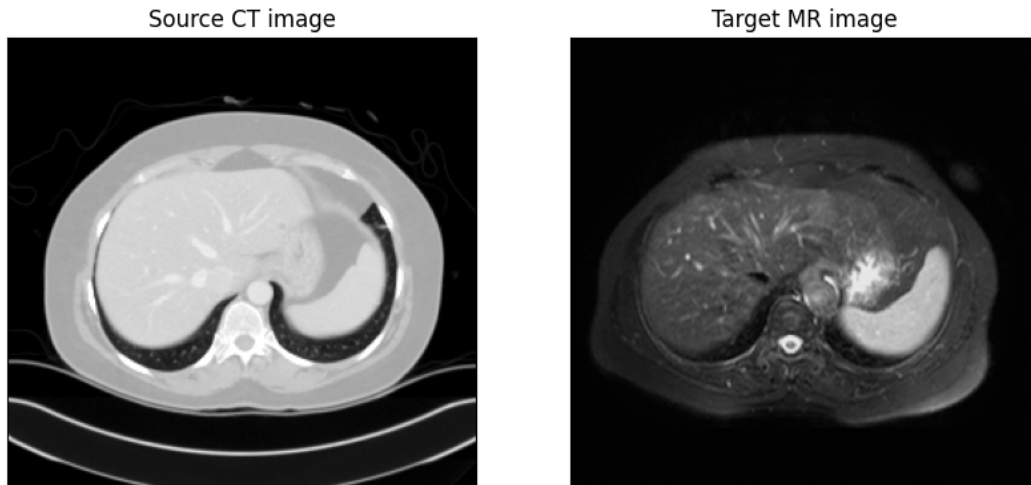


Figure 4: Left: Axial slice from abdominal CT from CHAOS + LiTS datasets (source domain). Right: Axial slice from abdominal MR from CHAOS dataset (target domain).

In Figure 4, we display CT (source domain) and MR (target domain) slices that were used in our UDA experiment for liver segmentation task. The corresponding quantitative results of this experiment are presented in the Table 3 of our paper. The domain shift between CT and MR images differs from the one observed between CT and CBCT, as previously discussed. While CT and MR modalities share comparable fields of view, they exhibit substantial differences in intensity distributions.

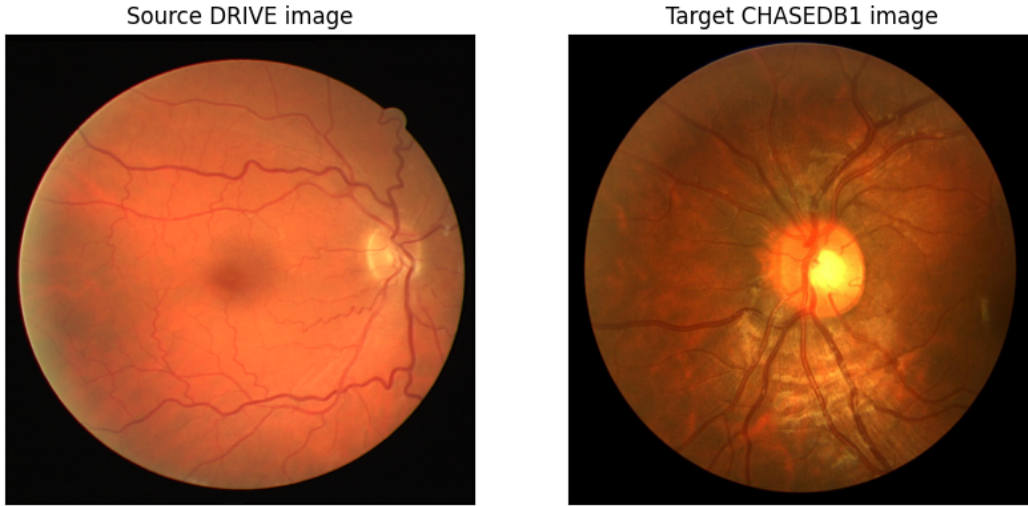


Figure 5: Left: Digital Retinal Image from DRIVE dataset (source domain). Right: Digital Retinal Image from CHASEDB1 dataset (target domain).

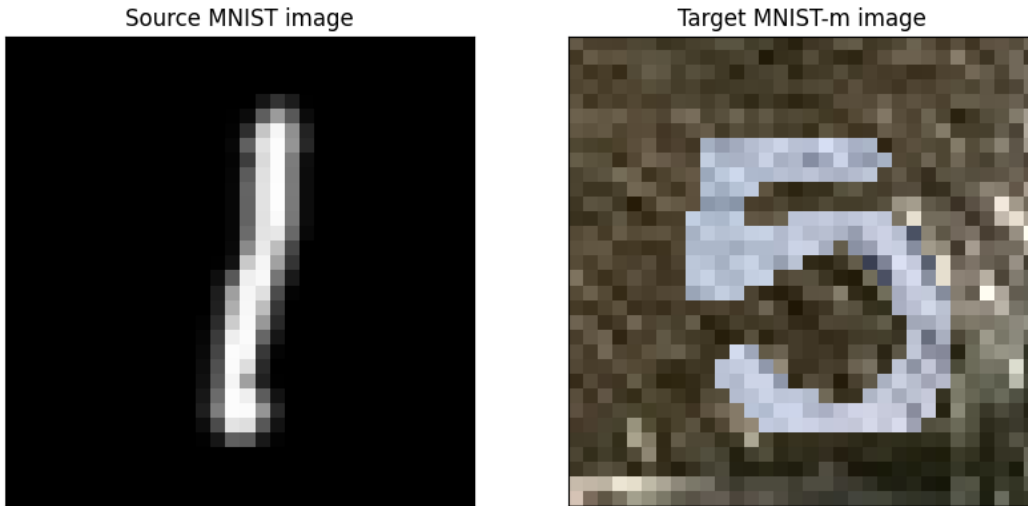


Figure 6: Left: Grayscale handwritten digits from MNIST dataset (source domain). Right: RGB handwritten digits from MNIST-M dataset (target domain).

In Figure 5, we display digital retinal images from DRIVE (source domain) and from CHASEDB1 (target domain) datasets that were used in our UDA experiment for retinal vasculature segmentation. The corresponding quantitative results of this experiment are presented in the Table 2 of our paper. As in CT to CBCT UDA, the domain shift is also composed with intensity and geometry differences. DRIVE images were captured with a wider field of view than CHASEDB1 images. Furthermore, the optic disk position is different between these source and target images. In DRIVE images, it appears on the left or on the right of the fundus region, whereas it is located on its center in CHASEDB1

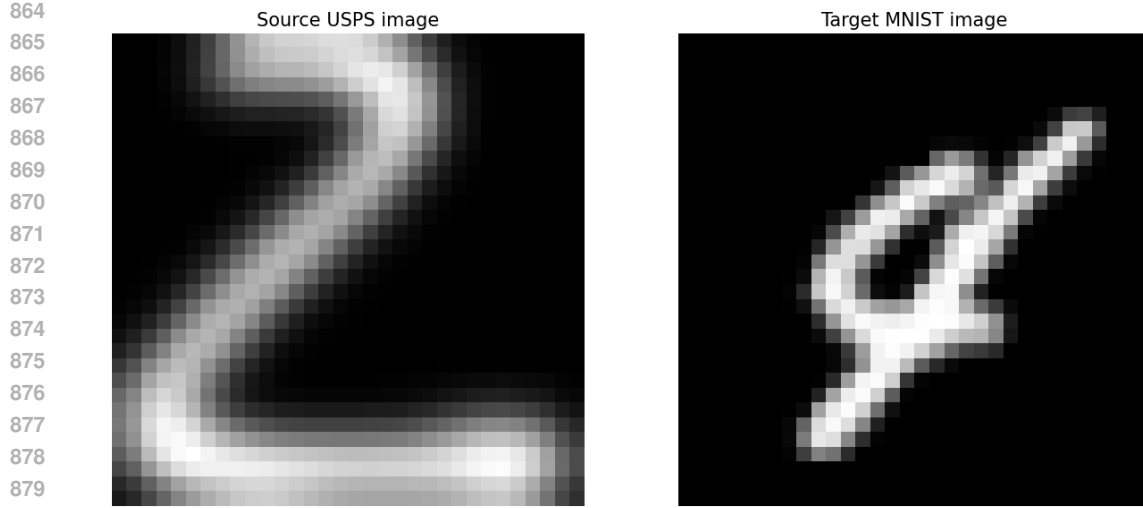


Figure 7: Left: Grayscale handwritten digits from USPS dataset (source domain). Right: Grayscale handwritten digits from MNIST dataset (target domain).

images. Finally, there is more intensity variations inside blood vessels regarding CHASEDB1 images comparing to DRIVE images.

In Figure 6, we display handwritten digits from MNIST (source domain) and MNIST-M (target domain) datasets that were used in our UDA experiment for digit classification. The corresponding quantitative results of this experiment are presented in the Table 1 of our paper. The MNIST and MNIST-M datasets exhibit notable differences in both color composition and intensity distribution. MNIST-M was derived by superimposing the original grayscale MNIST handwritten digits onto complex, natural image backgrounds sampled from RGB datasets. As a result, MNIST-M images are represented in three color channels (RGB), in contrast to the single-channel grayscale format of the original MNIST dataset. Furthermore, the backgrounds of MNIST-M images introduce visual noise and texture, making the classification task more challenging.

As shown in Figure 7, there is a clear domain shift between USPS and MNIST. USPS digits are collected at a lower resolution (16×16 compared to 28×28 in MNIST), tend to be written more neatly, and exhibit smoother, less varied handwriting styles. In contrast, MNIST digits are larger, noisier, and include a much wider diversity of handwriting.

In addition, USPS has far fewer images (around 9,000 samples compared to MNIST’s 70,000), and the digit styles are less complex and less diverse. This means a model trained on USPS captures only a limited range of digit variations. When transferred to MNIST, the model encounters digit appearances it has never seen before (e.g., thicker strokes, unusual slants, or noisier backgrounds), leading to poor generalization.

Because of these differences, models trained on USPS (source) tend to struggle when applied to MNIST (target), as they fail to generalize to the broader variations present in MNIST. Conversely, the MNIST $\rightarrow$ USPS transfer direction is generally more successful, since MNIST’s larger sample size and richer variability cover much of the USPS distribution. This asymmetry highlights the importance of both dataset size and distributional diversity in cross-domain generalization.

A.3 FEATURE ALIGNMENT

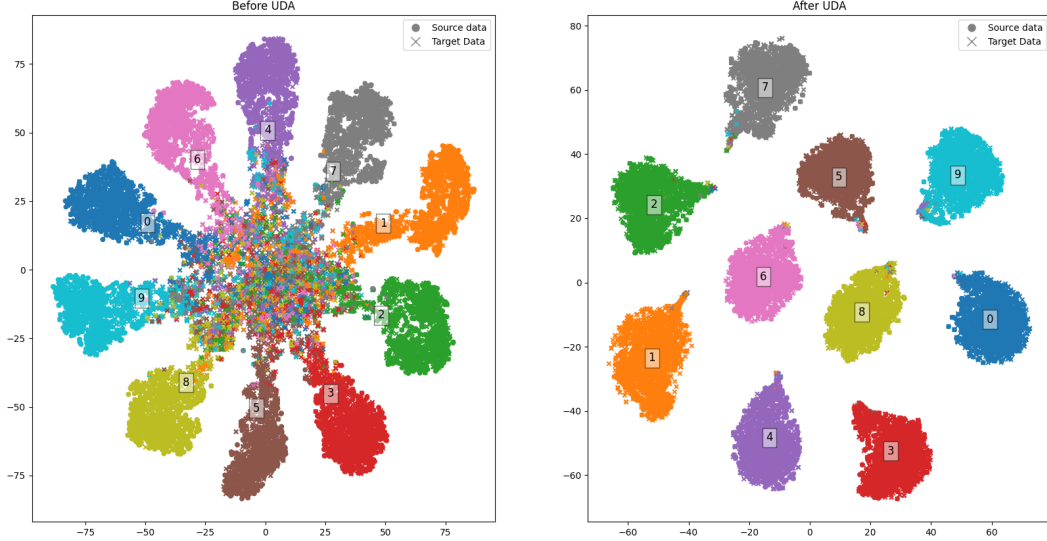


Figure 8: T-SNE showing source (MNIST) and target (MNIST-M) embeddings before (left) and after (right) applying our FARR UDA.

In Figure 8, we propose a qualitative evaluation of our FARR method on MNIST (source domain) to MNIST-M (target domain) UDA for classification task. We use t-distributed Stochastic Neighbor Embedding (t-SNE) to visualize the high-dimensional feature representations learned by the model. In this Figure, we can visualize the 2D projections of features extracted by ψ from source and target images before (left) and after (right) UDA.

In Figure 8, the circles correspond to representations from source images and the crosses correspond to representations from target images. The class labels are represented by colors and by Figures in each plot.

This visualization reveals that the learned features are well-aligned between source and target data after applying our FARR UDA. Before UDA, the representations from source images are closed when they belong to the same class, and separated when they do not. However, the representations from target images are mixed between different classes. This qualitatively shows that the model trained on source domain only does not extract domain-invariant features. On the other hand, after UDA, samples from the target domain are mixed with their corresponding source counterparts. This suggests that the model has learned domain-invariant features. This alignment is critical for achieving good performance on the unlabeled target domain and indicates successful adaptation.

These results indicate that FARR successfully aligns features from source and target domains on the MNIST to MNIST-M UDA task.

A.4 ABLATION STUDY ON THE SENSITIVITY OF OUR METHOD TO HYPERPARAMETERS α , γ

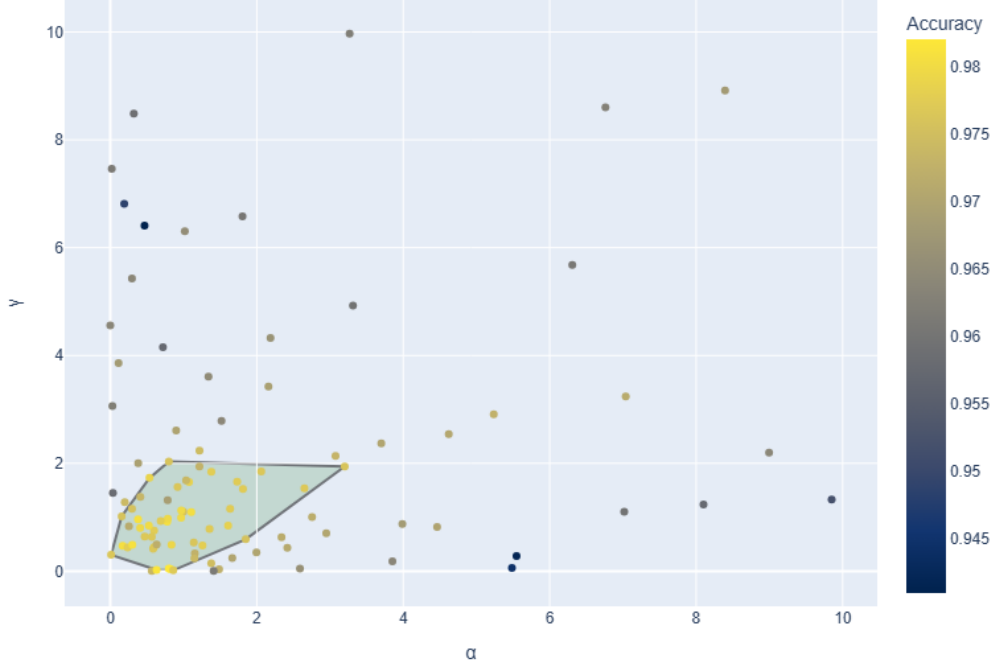


Figure 9: Ablation study on hyperparameters α and γ . Test accuracy on the MNIST-M target domain was evaluated following unsupervised domain adaptation from the MNIST source domain using the proposed method FARR. A heatmap was generated to illustrate the impact of varying the hyperparameters α (x-axis) and γ (y-axis) on performance in the target domain. The results demonstrate that the method exhibits robustness to changes in these hyperparameters, with the highest accuracies observed when both α and γ fall within the range $[0, 1]$. This region of stability is highlighted by a polygon that encloses the set of hyperparameter configurations yielding a target domain accuracy exceeding 0.975.

In Figure 9, we present a systematic study of the stability of our proposed method FARR with respect to the hyperparameters α and γ . To ensure comparability, we adopt the exact same experimental setup as described in the main paper, including both the model architecture and the optimization procedure. We then perform UDA training while varying α and β over a broad interval, $[0, 10]$. The resulting performance on the target domain MNIST-M reveals two key observations. First, within the polygon-shaped region of the grid, the method consistently achieves accuracy above 0.975, indicating that small to moderate variations of α and γ have negligible impact on performance. Second, even for extreme values at the boundaries of the search range, the lowest accuracy remains around 0.945. This is still higher than all feature-alignment baselines reported in the main article (Table 1). Taken together, these results demonstrate that the proposed method exhibits strong robustness to hyperparameter selection, making it practical and reliable in real-world scenarios where extensive tuning is often infeasible.

A.5 REDUNDANCY REDUCTION

In this section, we provide experimental results related to the cross-correlation matrix $\mathcal{C}$ defined as $\mathcal{C}[i, j] \triangleq \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle$.

	Source Domain	Target Domain
Average on-diagonal	0.9997	0.9998
Average off-diagonal	0.0025	0.0027

Table 4: Average on-diagonal and off-diagonal values of the cross-correlation matrix $\mathcal{C}$ obtained from 8.960 MNIST images (source domain) and 8.960 MNIST-M images (test domain) after UDA.

As presented in our paper Section 3.3, combining previous Lemma 1 and Theorem 1 provides the following result. At optimum, we have:

$$\forall z, \forall (i, j) \in \{1, \dots, D\} \times \{1, \dots, D\}, \quad \mathcal{C}[i, j] = \langle \phi_f(\mathbf{z})_i, \phi_{f'}(\mathbf{z})_j \rangle = \begin{cases} \frac{1}{1+\gamma} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad (9)$$

In Table 4, we provide a table with average on-diagonal and off-diagonal values of the cross-correlation matrix $\mathcal{C}$, computed using 8.960 source images and 8.960 target images from the test datasets. The features were extracted using ψ, f, f' after performing our FARR UDA training with hyperparameter $\gamma = 3.0 \times 10^{-4}$. From equation 9, we would expect to have an average on-diagonal value equal to 0.9997 and an average off-diagonal value close to 0 for both source and target domains. Our results in Table 4 confirm this theoretical result. In Figure. 10 and Figure 11, we provide examples of such cross-correlation matrices obtained for mnist $\rightarrow$ MNIST-M and USPS $\rightarrow$ MNIST UDA.

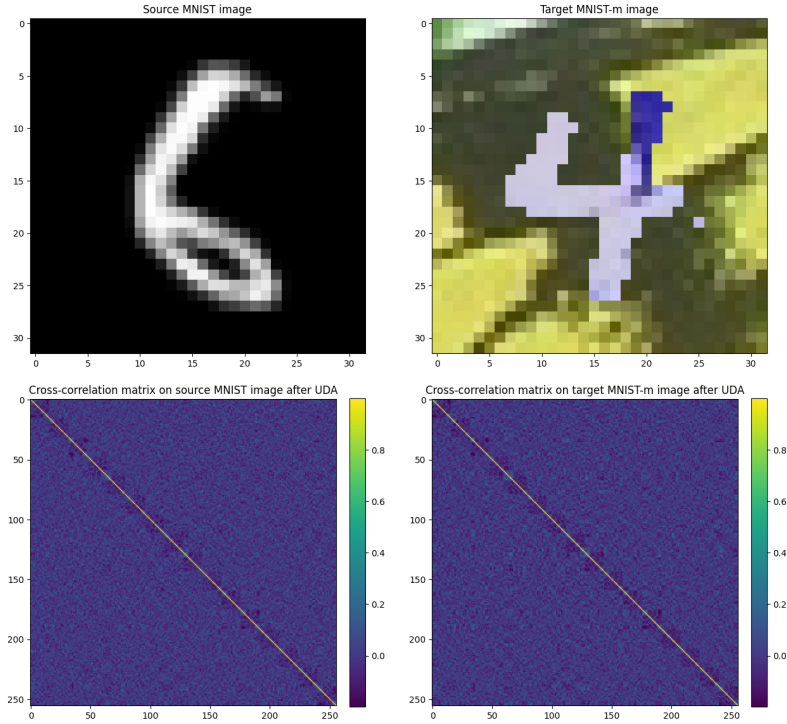


Figure 10: Example of MNIST (source domain) and MNIST-M (target domain) with the corresponding cross-correlation matrices after UDA.

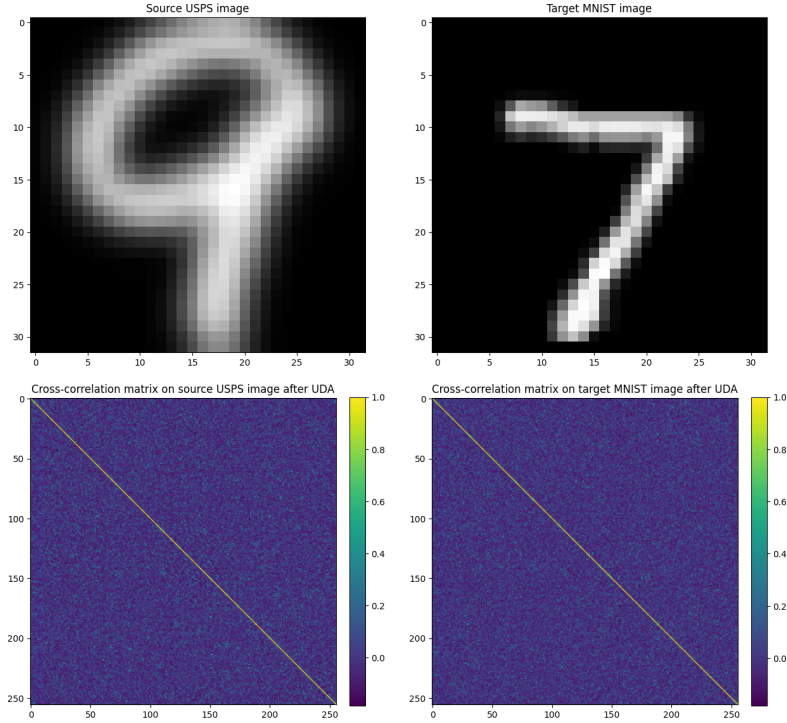


Figure 11: Example of USPS (source domain) and MNIST (target domain) with the corresponding cross-correlation matrices after UDA.

A.6 EXPLORATION OF THE CAPACITY OF f'

In the main paper, we reported that the best classification performance regarding MNIST to USPS, USPS to MNIST, MNIST to MNIST-M was achieved when using the last block of a ResNet-18, which corresponds to approximately 42% of the overall network (Table 1). Since the theoretical guarantees of our method rely on the validity of Assumption 1—in particular, on the representational capacity of f' —we designed an additional experiment to explicitly examine the impact of the capacity of f' . The goal was to determine whether the method remains effective across a wide range of f' sizes, including configurations with significantly reduced capacity compared to the one used in the main paper.

In this experiment, we defined ψ as the full ResNet-18 backbone with an embedding size of 256. The function f was implemented as a representation head composed of two linear layers with a ReLU activation in between (mapping from 256 to f_{dim} , and then from f_{dim} to 32). The classifier g was fixed as a linear layer from 32 to $n_{classes} = 10$. We pre-trained the model on MNIST source domain with $f_{dim} \in \{16, 32, 64, 128, 256\}$ and subsequently performed MNIST to MNIST-M UDA training by initializing f' as a copy of f . Thus, the only factor that varies across experiments is f_{dim} , directly controlling the capacity of f' .

The results, summarized in Table 5, reveal two main trends. First, performance consistently improves as the capacity of f' increases, highlighting the benefits of a richer representation. Second, and more importantly, even with minimal capacity—e.g., $f_{dim} = 16$, where f' accounts for only 0.17% of the total network parameters—the method still achieves a strong accuracy of 97% on the target domain. For comparison, when f' corresponds to 42% of the network (as in the main paper), the accuracy reaches 98.2%. These findings demonstrate that our method is highly stable with respect to the capacity of f' and remains effective even under severe capacity constraints.

f_{dim}	16	32	64	128	256
Δ parameters	0.17%	0.33%	0.66%	1.33%	2.66%
Target Accuracy	0.9702	0.9708	0.9709	0.9716	0.9734

Table 5: Ablation study on f' capacity. Test accuracy on the MNIST-M target domain was evaluated following unsupervised domain adaptation from the MNIST source domain using the proposed method FARR with various complexity of f' taking $f_{dim} \in \{32, 64, 128, 256\}$ as described in section F. For each parameter f_{dim} , we report is the ratio between the number of parameters of f' and the number of parameters of ψ (Δ parameters) and the accuracy on the test set of the target domain (Target Accuracy).

A.7 DATASETS

- **MNIST** LeCun et al. (1998): 28×28 grayscale images of handwritten digits (0–9), with 60,000 training and 10,000 test samples. Standard benchmark for digit classification.
- **MNIST-M** Ganin et al. (2016): Colored digits created by overlaying MNIST digits on random patches from BSDS500 images. Used for unsupervised domain adaptation to introduce style variability.
- **USPS** Denker et al. (1989): 16×16 grayscale images of handwritten digits, totaling 9,298 samples. Dataset for UDA experiments with limited style diversity.
- **DRIVE** Staal et al. (2004): 40 retinal images (20 training, 20 test) with manually annotated blood vessels. Each image is 565×584 pixels. Used for vessel segmentation tasks in medical imaging.
- **CHASEDB1** Hunter et al. (2013): 28 high-resolution retinal images (14 training, 14 test) with annotated vessel masks. Images are 999×960 pixels. Another benchmark for retinal vessel segmentation.
- **CHAOS** Kavur et al. (2021): Combined CT and MR images of healthy abdominal organs. Dataset includes multiple organs (liver, spleen, kidneys) with manual segmentations. Used for multi-organ segmentation tasks.
- **LiTS** Bilic et al. (2019): Liver Tumor Segmentation dataset containing 131 CT scans with liver and tumor annotations. Standard benchmark for liver and tumor segmentation in abdominal CT imaging.

A.8 TRAINING DETAILS

We provide the detailed training protocols to ensure reproducibility. Unless otherwise stated, all experiments were conducted on a single Nvidia V100 GPU with 32GB memory. For each dataset pair, we adopted a four-fold split: $\frac{3}{4}$ of the data for training and validation, and $\frac{1}{4}$ for testing. All models were trained with the Adam optimizer. For all experiments, the hyperparameters α and γ were determined through hyperparameter search using Optuna Akiba et al. (2019).

A.8.1 DIGIT CLASSIFICATION: MNIST→MNIST-M, MNIST→USPS, MNIST→USPS

- **Backbone:** ResNet18 with 32 channels.
- **Architecture:** f' defined as the last layer block of the final layer group of ResNet-18, g as the classification head. An ablation study on other decomposition can be found in Appendix A.9.
- **Augmentation:** None.
- **Hyperparameters:** Learning rate $1e-3$, batch size 32. α and γ can be found in Table 6.

A.8.2 RETINAL VASCULATURE SEGMENTATION: DRIVE→CHASEDB1

- **Backbone:** 5-stage U-Net, 64 channels at stage 1.
- **Architecture:** f' corresponds to the last stage of the U-Net, g is the final 1×1 convolution.
- **Preprocessing:** Images resized to 512×512 pixels.

Table 6: Hyperparameters α and γ for each UDA task.

Task	α	γ
MNIST $\rightarrow$ MNIST-M	0.80	5.0×10^{-2}
USPS $\rightarrow$ MNIST	0.55	2.1×10^{-2}
MNIST $\rightarrow$ USPS	0.60	1.32

- **Augmentation:** Rotation, zoom, cut-mix.
- **Hyperparameters:** Learning rate $1e-3$, $\alpha = 1.5 \times 10^{-1}$, $\gamma = 3.2 \times 10^{-3}$, batch size 4.

A.8.3 LIVER SEGMENTATION

We used the same U-Net backbone as in the retinal experiments. All 2D axial slices were resampled to a common pixel size of 1.8mm. Augmentations included rotation, zoom, random flipping, contrast adjustment, and additive noise.

CT $\rightarrow$ CBCT (Private dataset)

- **Dataset:** 15,827 CT slices and 13,024 CBCT slices.
- **Hyperparameters:** Learning rate $1e-3$, $\alpha = 2.4 \times 10^{-1}$, $\gamma = 6.2 \times 10^{-3}$, batch size 8.

CT $\rightarrow$ MR (LiTS $\rightarrow$ CHAOS)

- **Dataset:** 5,324 CT slices (LiTS) and 647 MR slices (CHAOS).
- **Hyperparameters:** Learning rate $1e-3$, $\alpha = 3.0 \times 10^{-1}$, $\gamma = 7.5 \times 10^{-3}$, batch size 8.

A.9 ABLATION STUDY ON NETWORK DECOMPOSITION FOR FARR

To better understand the flexibility of our approach, we conduct an ablation study on how the decomposition of the predictor $h(x) \equiv g(f(\psi(x)))$ into a feature extractor $\psi(\cdot)$, a representation head $f(\cdot)$, and a task-specific prediction head $g(\cdot)$ affects performance. Using ResNet-18 for the classification task, we vary the granularity of the representation head $f(\cdot)$. In particular, we compare using the entire `layer4` block (`self.layer4`) versus only its final sub-block (`self.layer4[1]`) as $f(\cdot)$, while adjusting the feature extractor $\psi(\cdot)$ accordingly. The size of f and f' consequently ranges from approximately 75% to 42% of the total number of parameters in the network. We conduct this experiment on USPS $\rightarrow$ MNIST using same hyperparameters (α , γ , learning rate, batch size) for both decompositions.

In Table 7, this experiment reveals that even shallow representation heads can yield strong alignment. However, the relative capacity of ψ and f also determines the strength of the adversarial signal provided by f' , and thus influences the overall feature alignment. This highlights the importance of balancing the expressiveness of the feature extractor and the adversary, and confirms the robustness of our method to architectural choices.

size of ψ	24.85%	57.7%
size of f	75.05%	42.2%
Target Accuracy	0.961	0.986

Table 7: Ablation study on network decomposition for FARR in USPS $\rightarrow$ MNIST UDA for classification. The size of ψ and f' represent the number of parameters of the two modules divided by the total number of parameters of the network. Test accuracy on the MNIST target domain was evaluated using the proposed method FARR with various decompositions for ψ , f , g . As g is fixed and represent 0.1% of the network size in both cases, we only report the size of ψ and f .

Table 8: Test accuracy on the MNIST target domain was evaluated following unsupervised domain adaptation from the USPS source domain using the proposed method FARR with various backbones.

Backbone	ResNet-18	ViT-S/14
Target Accuracy	0.986	0.953 ⁴

A.10 FLEXIBILITY OF ARCHITECTURE

In our main experiments, we adopt CNN backbones: ResNet-18 for digit classification and U-Net for segmentation tasks (e.g., DRIVE→CHASEDB1, CT→CBCT, CT→MR). To further demonstrate the architectural flexibility of FARR, we conducted an experiment using a transformer backbone by replacing ResNet-18 with DINOv2-Small Oquab et al. (2023) (ViT-S/14) on the USPS→MNIST classification task (Table 8). FARR achieved 95.3% accuracy with DINOv2-Small, compared to 98.6% with ResNet-18. DINOv2 was primarily used to leverage pretraining and provide a strong initialization for the source-only model, analogous to using ResNet-18 with a low learning rate. FARR was implemented with the same hyperparameters (α and γ) as for ResNet-18 (Appendix A.8), and the representation head f and adversary f' were defined as the last transformer block of ViT-S/14. While this configuration—mirroring the ResNet-18 setup—may not be optimal for the transformer, these results indicate that FARR can be applied beyond CNNs, supporting its architecture-agnostic nature.

⁴For a fair comparison with ResNet-18, a full hyperparameter search (α , γ , ψ and f split) should be performed as for ResNet-18