Information-theoretic regularization for Multi-source Domain Adaptation

Geon Yeong Park Sang Wan Lee
KAIST
Daejeon, South Korea
{pky3436, sangwan}@kaist.ac.kr

A. Pseudocode

Due to the limited space, we provide the algorithm of **MIAN** in this Section. Details about training-dependent scaling of β_t are in Section E.

Algorithm 1: Multi-source Information-regularized Adaptation Networks (**MIAN**)

Input: mini-batch size for each domain m, Number of source domains N, Training iteration T.
$M = m(N+1)$, Set of domain labels
$\mathcal{V} = \{1, \dots, N+1\}$.
Output: Transferable Encoder F, Classifier C
for $t \leftarrow 1$ ***to*** T **do**
  $X = \{\mathbf{x}_i\}_{i=1}^{M} = X_{S_1} \bigcup \cdots \bigcup X_{S_N} \bigcup X_T$
  $Y = \{\mathbf{y}_i\}_{i=1}^{mN} = Y_{S_1} \bigcup \cdots \bigcup Y_{S_N}$
  Encode latent representation $\mathbf{z}_i = F(\mathbf{x}_i)$

  // Inner maximization
  Optimize discriminator h by the objective $L(h)$ in (16) using gradient descent.

  // Outer minimization
  $L(F, C) =$
  $-\frac{1}{mN} \sum_{\mathbf{y} \in \mathcal{Y}} \sum_{i:\mathbf{y}_i = \mathbf{y}} \left[\mathbb{1}^T_{[k=\mathbf{y}_i]} \log \hat{\mathbf{y}}_i \right]$
  $\beta_t = \beta_0 \cdot 2\big(1 - \frac{1}{1+exp(-\sigma \cdot t/T)}\big)$
  $L(F) = L(F, C) - \beta_t L(h)$;
  Optimize encoder F by the objective $L(F)$ using gradient descent.
  Optimize classifier C by the objective $L(F, C)$ using gradient descent.

B. Proofs

In this Section, we present the detailed proofs for Theorems 2, 3 and Lemma 2, explained in the main paper. Following [16], we provide a proof of Theorem 2 below for the sake of completeness.

B.1. Proof of Theorem 2

Theorem 2. *Let $P_Z(\mathbf{z})$ be the distribution of Z where $\mathbf{z} \in \mathcal{Z}$. Let h be a domain classifier $h : \mathcal{Z} \to \mathcal{V}$, where $\mathcal{Z}$ is the feature space and $\mathcal{V}$ is the set of domain labels. Let $h_{\mathbf{v}}(Z)$ be a conditional probability of V where $\mathbf{v} \in \mathcal{V}$ given $Z = \mathbf{z}$, defined by h. Then the following holds:*

$$I(Z;V) = \max_{h_{\mathbf{v}}(\mathbf{z}):\sum_{\mathbf{v}\in\mathcal{V}} h_{\mathbf{v}}(\mathbf{z})=1, \forall \mathbf{z}} \sum_{\mathbf{v}\in\mathcal{V}} P_V(\mathbf{v}) \mathbb{E}_{\mathbf{z}\sim P_{Z|\mathbf{v}}} \big[\log h_{\mathbf{v}}(\mathbf{z})\big] + H(V) \tag{1}$$

Proof. By definition,

$$\begin{aligned} I(Z;V) &= D_{KL}\big(P(Z,V) \parallel P(Z)P(V)\big) \\ &= \sum_{\mathbf{v}\in\mathcal{V}} P_V(\mathbf{v}) \mathbb{E}_{\mathbf{z}\sim P_{Z|\mathbf{v}}} \Big[\log \frac{P_{Z,V}(\mathbf{z},\mathbf{v})}{P_Z(\mathbf{z})}\Big] + H(V) \end{aligned} \tag{2}$$

Let us constrain the term inside the log by $h_{\mathbf{v}}(\mathbf{z}) = \frac{P_{Z,V}(\mathbf{z},\mathbf{v})}{P_Z(\mathbf{z})}$ where $h_{\mathbf{v}}(\mathbf{z})$ represents the conditional probability of $V = \mathbf{v}$ for any $\mathbf{v} \in \mathcal{V}$ given $Z = \mathbf{z}$. Then we have: $\sum_{\mathbf{v}\in\mathcal{V}} h_{\mathbf{v}}(\mathbf{z}) = 1$ for all possible values of $\mathbf{z}$ according to the law of total probability. Let $\mathbf{h}$ denote the collection of $h_{\mathbf{v}}(\mathbf{z})$ for all possible values of $\mathbf{v}$ and $\mathbf{z}$, and $\boldsymbol{\lambda}$ be the collection of $\lambda_{\mathbf{z}}$ for all values of $\mathbf{z}$. Then, we can construct the Lagrangian function by incorporating the constraint $\sum_{\mathbf{v}\in\mathcal{V}} h_{\mathbf{v}}(\mathbf{z}) = 1$ as follows:

$$\begin{aligned} L(\mathbf{h}, \boldsymbol{\lambda}) = &\sum_{\mathbf{v}\in\mathcal{V}} P_V(\mathbf{v}) \mathbb{E}_{\mathbf{z}\sim P_{Z|\mathbf{v}}} \Big[log\big(h_{\mathbf{v}}(\mathbf{z})\big)\Big] + H(V) \\ &+ \sum_{\mathbf{z}\in\mathcal{Z}} \lambda_{\mathbf{z}} \Big(1 - \sum_{\mathbf{v}\in\mathcal{V}} h_{\mathbf{v}}(\mathbf{z})\Big) \end{aligned} \tag{3}$$

We can use the following KKT conditions:

$$\frac{\partial L(\mathbf{h}, \boldsymbol{\lambda})}{\partial h_{\mathbf{v}}(\mathbf{z})} = P_V(\mathbf{v}) \frac{P_{Z|\mathbf{v}}(\mathbf{z})}{h^*_{\mathbf{v}}(\mathbf{z})} - \lambda^*_{\mathbf{z}} = 0, \ \forall (\mathbf{z}, \mathbf{v}) \in \mathcal{Z} \times \mathcal{V} \tag{4}$$

$$1 - \sum_{\mathbf{v}\in\mathcal{V}} h^*_{\mathbf{v}}(\mathbf{z}) = 0, \quad \forall \mathbf{z} \in \mathcal{Z} \tag{5}$$

Solving the two equations, we have $1 - \sum_{\mathbf{v}\in\mathcal{V}} \frac{P_V(\mathbf{v})P_{Z|\mathbf{v}}(\mathbf{z})}{\lambda^*_{\mathbf{z}}} = 0$ such that $\lambda^*_{\mathbf{z}} = P_Z(\mathbf{z})$ for all $\mathbf{z}$. Then for all the possible values of $\mathbf{z}$,

$$\begin{aligned} h^*_{\mathbf{v}}(\mathbf{z}) &= \frac{P_{Z,V}(\mathbf{z},\mathbf{v})}{P_Z(\mathbf{z})} \\ &= P_{V|\mathbf{z}}(\mathbf{v}), \end{aligned} \tag{6}$$

where the given $h^*_{\mathbf{v}}(\mathbf{z})$ is same as the term inside log in (2). Thus, the optimal solution of concave Lagrangian function (3) obtained by $h^*_{\mathbf{v}}(\mathbf{z})$ is equal to the mutual information in (2). The substitution of $h^*_{\mathbf{v}}(\mathbf{z})$ into (2) completes the proof. □

Our framework can further be applied to segmentation problems because it provides a new perspective on pixel space [19, 20, 13] and segmentation space [22] adaptation. The generator in pixel space and segmentation space adaptation learns to transform images or segmentation results from one domain to another. In the context of information regularization, we can view these approaches as limiting information $I(\hat{X}; V)$ between the generated output $\hat{X}$ and the domain label V, which is accomplished by involving the encoder for pixel-level generation. This alleviates the domain shift in a raw pixel level. Note that one can choose between limiting the feature-level or pixel-level mutual information. These different regularization terms may be complementary to each other depending on the given task.

B.2. Proof of Theorem 3

Theorem 3. *Let $P_{Z|\mathbf{x},\mathbf{v}}(\mathbf{z})$ be a conditional probabilistic distribution of Z where $\mathbf{z} \in \mathcal{Z}$, defined by the encoder F, given a sample $\mathbf{x} \in \mathcal{X}$ and the domain label $\mathbf{v} \in \mathcal{V}$. Let $R_Z(\mathbf{z})$ denotes a prior marginal distribution of Z. Then the following inequality holds:*

$$\begin{aligned} &I(Z; X, V) \leq \mathbb{E}_{\mathbf{x},\mathbf{v}\sim P_{X,V}}\big[D_{KL}[P_{Z|\mathbf{x},\mathbf{v}} \parallel R_Z]\big] + H(V) \\ &+ \max_{h_{\mathbf{v}}(\mathbf{z}):\sum_{\mathbf{v}\in\mathcal{V}} h_{\mathbf{v}}(\mathbf{z})=1,\forall \mathbf{z}} \sum_{\mathbf{v}\in\mathcal{V}} P_V(\mathbf{v})\mathbb{E}_{P_{\mathbf{z}\sim Z|\mathbf{v}}}\big[\log h_{\mathbf{v}}(\mathbf{z})\big] \end{aligned} \tag{7}$$

Proof. Based on the chain rule for mutual information,

$$\begin{aligned} &I(Z; X, V) \\ &= I(Z; V) + I(Z; X \mid V) \\ &= H(V) + I(Z; X \mid V) \\ &+ \max_{h_{\mathbf{v}}(\mathbf{z}):\sum_{\mathbf{v}\in\mathcal{V}} h_{\mathbf{v}}(\mathbf{z})=1,\forall \mathbf{z}} \sum_{\mathbf{v}\in\mathcal{V}} P_V(\mathbf{v})\mathbb{E}_{\mathbf{z}\sim P_{Z|\mathbf{v}}}\big[\log h_{\mathbf{v}}(\mathbf{z})\big], \end{aligned} \tag{8}$$

where the latter equality is given by Theorem 2. Then,

$$\begin{aligned} &I(Z; X \mid V) \\ &= \mathbb{E}_{\mathbf{v}\sim P_V}\Big[\mathbb{E}_{\mathbf{z},\mathbf{x}\sim P_{Z,X|\mathbf{v}}}\Big[\log \frac{P_{Z,X|\mathbf{v}}(\mathbf{z},\mathbf{x})}{P_{Z|\mathbf{v}}(\mathbf{z})P_{X|\mathbf{v}}(\mathbf{x})}\Big]\Big] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{v}\sim P_{X,V}}\Big[\mathbb{E}_{\mathbf{z}\sim P_{Z|\mathbf{x},\mathbf{v}}}\Big[\log \frac{P_{Z|\mathbf{x},\mathbf{v}}(\mathbf{z})}{P_{Z|\mathbf{v}}(\mathbf{z})}\Big]\Big] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{v}\sim P_{X,V}}\Big[\mathbb{E}_{\mathbf{z}\sim P_{Z|\mathbf{x},\mathbf{v}}}\big[\log P_{Z|\mathbf{x},\mathbf{v}}(\mathbf{z})\big]\Big] \\ &\quad - \mathbb{E}_{\mathbf{v}\sim P_V}\Big[\mathbb{E}_{\mathbf{z}\sim P_{Z|\mathbf{v}}}\big[\log P_{Z|\mathbf{v}}(\mathbf{z})\big]\Big] \\ &\leq \mathbb{E}_{\mathbf{x},\mathbf{v}\sim P_{X,V}}\Big[\mathbb{E}_{\mathbf{z}\sim P_{Z|\mathbf{x},\mathbf{v}}}\big[\log P_{Z|\mathbf{x},\mathbf{v}}(\mathbf{z})\big]\Big] \\ &\quad - \mathbb{E}_{\mathbf{v}\sim P_V}\Big[\mathbb{E}_{\mathbf{z}\sim P_{Z|\mathbf{v}}}\big[\log R_Z(\mathbf{z})\big]\Big] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{v}\sim P_{X,V}}\Big[\mathbb{E}_{\mathbf{z}\sim P_{Z|\mathbf{x},\mathbf{v}}}\Big[\log \frac{P_{Z|\mathbf{x},\mathbf{v}}(\mathbf{z})}{R_Z(\mathbf{z})}\Big]\Big] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{v}\sim P_{X,V}}\Big[D_{KL}\big[P_{Z|\mathbf{x},\mathbf{v}} \parallel R_Z\big]\Big] \end{aligned} \tag{9}$$

The second equality is obtained by using $P_{Z,X|\mathbf{v}}(\mathbf{z},\mathbf{x}) = P_{X|\mathbf{v}}(\mathbf{x})P_{Z|\mathbf{x},\mathbf{v}}(\mathbf{z})$. The inequality is obtained by using:

$$D_{KL}[P_{Z|\mathbf{v}} \parallel R_Z] = \mathbb{E}_{\mathbf{z}\sim P_{Z|\mathbf{v}}}\big[\log P_{Z|\mathbf{v}}(\mathbf{z}) - \log R_Z(\mathbf{z})\big], \tag{10}$$

where $R_Z(\mathbf{z})$ is a variational approximation of the prior marginal distribution of Z. The last equality is obtained from the definition of KL-divergence. The substitution of (9) into (8) completes the proof. □

The existing DA work on semantic segmentation tasks [12, 21] can be explained as the process of fostering close collaboration between the aforementioned information bottleneck terms. The only difference between Theorem 3 for $\mathcal{V} = \{0, 1\}$ and the objective function in [12] is that [12] employed the shared encoding $P_{Z|\mathbf{x}}(\mathbf{z})$ instead of $P_{Z|\mathbf{x},\mathbf{v}}(\mathbf{z})$, whereas some adversarial DA approaches use the unshared one [23].

B.3. Proof of Lemma 2

Lemma 2. *Let $d_{\mathcal{H}}(\mathcal{V}) = \frac{1}{N+1}\sum_{\mathbf{v}\in\mathcal{V}} d_{\mathcal{H}}(D_{\mathbf{v}}, D_{\mathbf{v}^c})$. Let $\mathcal{H}$ be a hypothesis class. Then,*

$$d_{\mathcal{H}}(\mathcal{V}) \leq \frac{1}{N(N+1)} \sum_{\mathbf{v},\mathbf{u}\in\mathcal{V}} d_{\mathcal{H}}(D_{\mathbf{v}}, D_{\mathbf{u}}). \tag{11}$$

Proof. Let $\alpha = \frac{1}{N}$ represents the uniform domain weight

for the mixture of domain $D_{\mathbf{v}^c}$. Then,

$$
\begin{aligned}
& d_{\mathcal{H}}(\mathcal{V}) \\
&= \frac{1}{N+1} \sum_{\mathbf{v} \in \mathcal{V}} d_{\mathcal{H}}(D_{\mathbf{v}}, D_{\mathbf{v}^c}) \\
&= \frac{1}{N+1} \sum_{\mathbf{v} \in \mathcal{V}} 2 \sup_{h \in \mathcal{H}} \Big| \mathbb{E}_{\mathbf{x} \sim P_{D^X_{\mathbf{v}}}} \left[\mathbb{I}\big(h(\mathbf{x}=1)\big) \right] \\
&\qquad\qquad - \mathbb{E}_{\mathbf{x} \sim P_{D^X_{\mathbf{v}^c}}} \left[\mathbb{I}\big(h(\mathbf{x}=1)\big) \right] \Big| \\
&= \frac{1}{N+1} \sum_{\mathbf{v} \in \mathcal{V}} 2 \sup_{h \in \mathcal{H}} \Big| \sum_{\mathbf{u} \in \mathcal{V}: \mathbf{u} \neq \mathbf{v}} \alpha \Big(\mathbb{E}_{\mathbf{x} \sim P_{D^X_{\mathbf{v}}}} \left[\mathbb{I}\big(h(\mathbf{x}=1)\big) \right] \\
&\qquad\qquad - \mathbb{E}_{\mathbf{x} \sim P_{D^X_{\mathbf{u}}}} \left[\mathbb{I}\big(h(\mathbf{x}=1)\big) \right] \Big) \Big| \\
&\leq \frac{1}{N+1} \sum_{\mathbf{v} \in \mathcal{V}} \sum_{\mathbf{u} \in \mathcal{V}: \mathbf{u} \neq \mathbf{v}} \alpha \cdot 2 \sup_{h \in \mathcal{H}} \Big| \mathbb{E}_{\mathbf{x} \sim P_{D^X_{\mathbf{v}}}} \left[\mathbb{I}\big(h(\mathbf{x}=1)\big) \right] \\
&\qquad\qquad - \mathbb{E}_{\mathbf{x} \sim P_{D^X_{\mathbf{u}}}} \left[\mathbb{I}\big(h(\mathbf{x}=1)\big) \right] \Big| \\
&= \frac{1}{N(N+1)} \sum_{\mathbf{v}, \mathbf{u} \in \mathcal{V}} d_{\mathcal{H}}(D_{\mathbf{v}}, D_{\mathbf{u}}),
\end{aligned}
\tag{12}
$$

where the inequality follows from the triangluar inequality and jensen's inequality.

□

C. Experimental setup

In this Section, we describe the datasets, network architecture and hyperparameter configuration.

C.1. Datasets

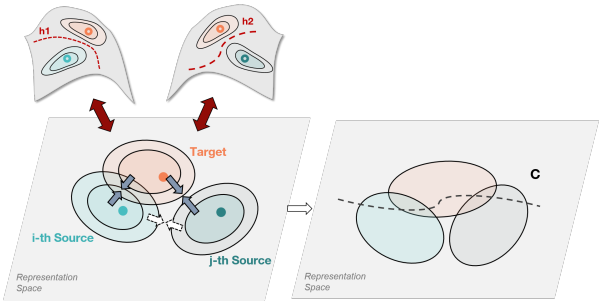
We validate the Multi-source Information-regularized Adaptation Networks (**MIAN**) with the following benchmark datasets: Digits-Five, Office-31 and Office-Home. Every experiment is repeated four times and the average accuracy in target domain is reported.

Digits-Five [15] dataset is a unified dataset including five different digit datasets: MNIST [8], MNIST-M [4], Synthetic Digits [4], SVHN, and USPS. Following the standard protocols of unsupervised MDA [26, 15], we used 25000 training images and 9000 test images sampled from a training and a testing subset for each of MNIST, MNIST-M, SVHN, and Synthetic Digits. For USPS, all the data is used owing to the small sample size. All the images are bilinearly interpolated to 32×32.

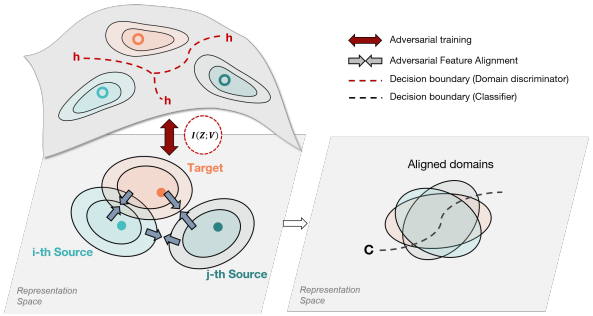
Office-31 [17] is a popular benchmark dataset including 31 categories of objects in an office environment. Note that it is a more difficult problem than Digits-Five, which includes 4652 images in total from the three domains: Amazon, DSLR, and Webcam. All the images are interpolated to 224×224 using bicubic filters.

Office-Home [24] is a challenging dataset that includes 65 categories of objects in office and home environments. It includes 15,500 images in total from the four domains: Artistic images (Art), Clip Art(Clipart), Product images (Product), and Real-World images (Realworld). All the images are interpolated to 224×224 using bicubic filters.

C.2. Architectures



(a) Existing works

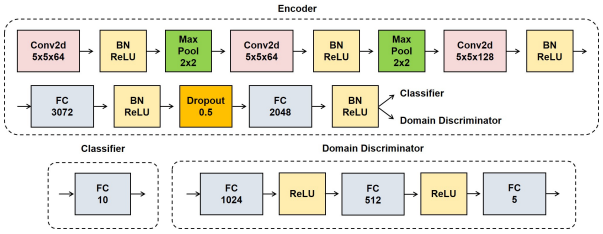


(b) Proposed model

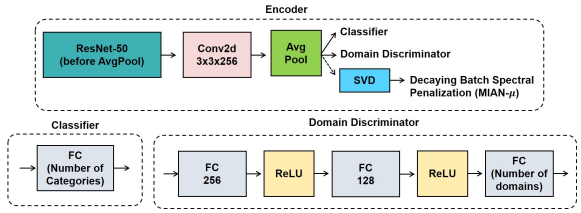
Figure 1: Comparison of existing and proposed MDA models. **(a)** Existing multiple-discriminator based methods align each pairwise source and target domain but may fail due to the disintegration of domain-discriminative knowledge. It also may suffer from unstable optimization and lack of resource-efficiency. **(b)** Our proposed model mitigates suggested problems by unifying domain discriminators.

Simulation setting For the Digits-Five dataset, we use the same network architecture and optimizer setting as in [15]. For all the other experiments, the results are based on ResNet-50, which is pre-trained on ImageNet. The domain discriminator is implemented as a three-layer neural network. Detailed architecture is shown in Figure 2.

We compare our method with the following state-of-the-art domain adaptation methods: Deep Adaptation Network (**DAN**, [10]), Joint Adaptation Network (**JAN**, [11]), Manifold Embedded Distribution Alignment (**MEDA**, [25]), Domain Adversarial Neural Network (**DANN**, [5]), Domain-Specific Batch Normalization (**DSBN**, [2]), Batch Spectral Penalization (**BSP**, [3]), Adversarial Discriminative Domain Adaptation (**ADDA**, [23]), Maximum Classifier Discrepancy



(a) Encoder, domain discriminator, and classifier used in Digits-Five experiments



(b) Encoder, domain discriminator, and classifier used in Office-31 and Office-Home experiments

Figure 2: Network architectures. BN denotes Batch Normalization [6] and SVD denotes differentiable SVD in PyTorch for **MIAN-γ** (Section E)

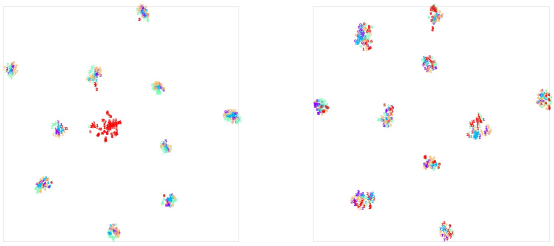
(**MCD**, [18]), Deep Cocktail Network (**DCTN**, [26]), and Moment Matching for Multi-Source Domain Adaptation (**M^{3}SDA**, [15]).

Hyperparameters Details of the experimental setup are summarized in Table 1. Other state-of-the-art adaptation models are trained based on the same setup except for these cases: **DCTN** show poor performance with the learning rate shown in Table 1 for both Office-31 and Office-Home datasets. Following the suggestion of the original authors, $1e^{-5}$ is used as a learning rate with the Adam optimizer [7]; **MCD** show poor performance for the Office-Home dataset with the learning rate shown in Table 1. $1e^{-4}$ is selected as a learning rate. For both the proposed and other baseline models, the learning rate of the classifier or domain discriminator trained from the scratch is set to be 10 times of those of ImageNet-pretrained weights, in Office-31 and Office-Home datasets. More hyperparameter configurations are summarized in Table 2 (Section E)

D. Additional results

Visualization of learned latent representations. We visualized domain-independent representations extracted by the input layer of the classifier with t-SNE (Figure 3). Before the adaptation process, the representations from the target domain were isolated from the representations from each source domain. However, after adaptation, the representations were well-aligned with respect to the class of digits, as opposed to the domain.

Hyperparameter sensitivity. We conducted the analysis on hyperparameter sensitivity with degree of regularization β. The target domain is set as Amazon or Art, where the value β_0 changes from 0.1 to 0.5. The accuracy is high when β_0 is approximately between 0.1 and 0.3. We thus choose $\beta_0 = 0.2$ for Office-31, and $\beta_0 = 0.3$ for Office-Home.



(a) Before adaptation (b) After adaptation

Figure 3: t-SNE visualization **(a)** before and **(b)** after adaptation. Representations from target domain (SVHN) are shown in red. Digit class labels are shown with corresponding numbers.

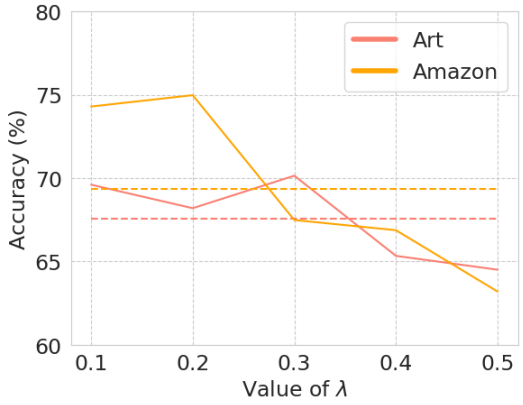


Figure 4: Analysis on hyperparameter sensitivity.

E. Decaying Batch Spectral Penalization

In this Section, we provides details on the Decaying Batch Spectral Penalization (DBSP) which expands **MIAN** into **MIAN-γ**.

E.1. Backgrounds

There is little motivation for models to control the complex mutual dependence to domains if reducing the entropy of representations is sufficient to optimize the value of $I(Z;V) = H(Z) - H(Z \mid V)$. If so, such implicit entropy minimization substantially reduce the upper bound of

Table 1: Experimental setup. The batch size for each domain is reported.

Dataset	Optimization method	Learning rate	Momentum	Batch size	Iteration
Digits-Five	Adam	$2e^{-4}$	(0.9, 0.99)	128	50000
Office-31	mini-batch SGD	$1e^{-3}$	0.9	16	25000
Office-Home	mini-batch SGD	$1e^{-3}$	0.9	16	25000

$I(Z;Y)$, potentially leading to a increase in optimal joint risk λ^*. In other words, the decrease in the entropy of representations may occur as the side effect of $I(Z;V)$ regularization. Such unexpected side effect of information regularization is highly intertwined with the hidden deterioration of discriminability through adversarial training [3, 9].

Based on these insights, we employ the SVD-entropy $H_{SVD}(\mathbf{Z})$ [1] of a representation matrix $\mathbf{Z}$ to assess the richness of the latent representations during adaptation, since it is difficult to compute $H(Z)$. Note that while $H_{SVD}(\mathbf{Z})$ is not precisely equivalent to $H(Z)$, $H_{SVD}(\mathbf{Z})$ can be used as a proxy of the level of disorder of the given matrix [14]. In future works, it would be interesting to evaluate the temporal change in entropy with other metrics. We found that $H_{SVD}(\mathbf{Z})$ indeed decreases significantly during adversarial adaptation, suggesting that some eigenfeatures (or eigensamples) become redundant and, thus, the inherent feature-richness diminishes (Figure 5a). To preclude such deterioration, we employ Batch Spectral Penalization (BSP) [3], which imposes a constraint on the largest singular value to solicit the contribution of other eigenfeatures. The overall objective function in the multi-domain setting is defined as:

$$\min_{F,C} \; L(F,C) + \beta \hat{I}(Z;V) + \gamma \sum_{i=1}^{N+1} \sum_{j=1}^{k} s_{i,j}^2, \qquad (13)$$

where β and γ are Lagrangian multipliers and $s_{i,j}$ is the jth singular value from the ith domain. We found that SVD entropy of representations is severely deteriorated especially in the early stages of training (Figure 5a), suggesting the possibility of over-regularization. The noisy domain discriminative signals in the initial phase [5] may distort and simplify the representations. To circumvent the impaired discriminability in the early stages of the training, the discriminability should be prioritized first with high γ and low β, followed by a gradual decaying and annealing in γ and β, respectively, so that a sufficient level of domain transferability is guaranteed. Based on our temporal analysis, we introduce the training-dependent scaling of β and γ by modifying the progressive training schedule [5]:

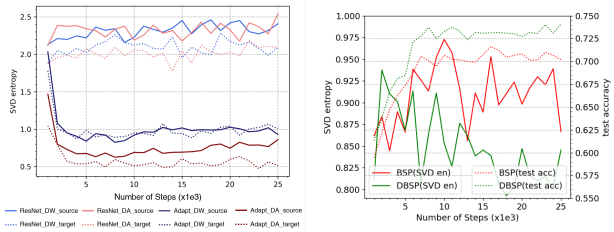
$$\begin{aligned} \beta_p &= \beta_0 \cdot 2\big(1 - \frac{1}{1 + exp(-\sigma \cdot p)}\big) \\ \gamma_p &= \gamma_0 \cdot \big(\frac{2}{1 + exp(-\sigma \cdot p)} - 1\big), \end{aligned} \qquad (14)$$

where β_0 and γ_0 are initial values, σ is a decaying parameter, and p is the training progress from 0 to 1. We refer to this version of our model as **MIAN-γ**. Note that **MIAN** only includes annealing-β, excluding DBSP. For the proposed method, β_0 is chosen from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ for Office-31 and Office-Home dataset, while $\beta_0 = 1.0$ is fixed in Digits-Five. γ_0 is fixed to $1e^{-4}$ following [3].

Table 2: Hyper parameters configuration. Annealing-β is not adopted in the Digits-Five experiment. Decaying batch spectral penalization is not adopted in the **MIAN**.

Dataset(Model)	β_0	γ_0	σ	k
Digits-Five (**MIAN**)	1.0	N/A	N/A	N/A
Office-31 (**MIAN**)	0.1	N/A	10.0	N/A
Office-31 (**MIAN-γ**)	0.2	0.0001	10.0	1
Office-Home (**MIAN**)	0.3	N/A	10.0	N/A
Office-Home (**MIAN-γ**)	0.3	0.0001	10.0	1

E.2. Experiments



(a) (b)

Figure 5: **(a)**: SVD-entropy analysis. (Office-31; Source domain: DSLR) **(b)**: Comparisons between BSP and DBSP. (Office-31; DSLR $\rightarrow$ Amazon)

SVD-entropy. We evaluated the degree of compromise of SVD-entropy owing to transfer learning. For this, DSLR was fixed as the source domain, and each Webcam and Amazon target domain was used to simulate low (DSLR$\rightarrow$ Webcam; *DW*) and high domain (DSLR$\rightarrow$Amazon; *DA*) shift conditions, respectively. SVD-entropy was applied to the representation matrix extracted from ResNet-50 and **MIAN** (denoted as *Adapt* in Figure 5a) with constant

$\beta = 0.1$. For accurate assessment, we avoided using spectral penalization. As depicted in the Figure 5a, adversarial adaptation, or information regularization, significantly decreases the SVD-entropy of both the source and target domain representations, especially in the early stages of training, indicating that the representations are simplified in terms of feature-richness. Moreover, when comparing the *Adapt_DA_source* and *Adapt_DW_source* conditions, we found that SVD-entropy decreases significantly as the degree of domain shift increases.

We additionally conducted analyses on temporal changes of SVD entropy by comparing BSP and decaying BSP (Figure 5b). SVD entropy gradually decreases as the degree of compensation decreases in DBSP which leads to improved transferability and accuracy. Thus DBSP can control the trade-off between the richness of the feature representations and adversarial adaptation as the training proceeds.

Ablation study of decaying spectral penalization. We performed an ablation study to assess the contribution of the decaying spectral penalization and annealing information regularization to DA performance (Table 3, 4). We found that the prioritization of feature-richness in early stages (by controlling β and γ) significantly improves the performance. We also found that the constant penalization schedule [3] is not reliable and sometimes impedes transferability in the low domain shift condition (Webcam, DSLR in Table 3). This implies that the conventional BSP may over-regularize the transferability when the degree of domain shift and SVD-entropy decline are relatively small.

Table 3: Ablation study of decaying batch spectral penalization and annealing information regularization (Office-31). For accurate assessment of extent to which performance improvement is caused by each strategies, γ is fixed as 0 in *Annealing-*β, and β is fixed as 0.1 in *Decaying-*γ. Results from *Annealing-*β and *Full version* are reported in main paper as **MIAN** and **MIAN-**γ, respectively.

Standards	Hyper parameters	Amazon	DSLR	Webcam	Avg
Baseline	$\beta = 0.1$ as a constant	69.98	99.48	98.13	89.20
Annealing-β (**MIAN**)	$\beta_0 = 0.1, \sigma = 10$	**74.65**	**99.48**	**98.49**	**90.87**
Decaying-γ	**BSP**: $\gamma = 1e^{-4}$ as a constant	74.73	98.65	96.24	89.87
	DBSP: $\gamma_0 = 1e^{-4}, \sigma = 10$	**75.01**	**99.68**	98.10	**90.93**
Full version (**MIAN-**γ)	$\beta_0 = 0.1, \gamma_0 = 1e^{-4}, \sigma = 10$	**76.17**	99.22	**98.39**	**91.26**

Table 4: Accuracy (%) on Office-Home dataset.

Standards	Art	Clipart	Product	Realworld	Avg
MIAN	69.39±0.50	63.05±0.61	79.62±0.16	80.44±0.24	73.12
MIAN-γ	**69.88±0.35**	**64.20±0.68**	**80.87±0.37**	**81.49±0.24**	**74.11**

References

[1] Orly Alter, Patrick O Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000. 5

[2] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019. 3

[3] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1081–1090, 2019. 3, 5, 6

[4] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014. 3

[5] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 3, 5

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 3

[9] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, pages 4013–4022, 2019. 5

[10] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015. 3

[11] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2208–2217. JMLR. org, 2017. 3

[12] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6778–6787, 2019. 2

[13] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018. 2

[14] Paul K Newton and Stephen A DeSalvo. The shannon entropy of sudoku matrices. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 466(2119):1957–1975, 2010. 5

[15] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019. 3, 4

[16] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fr-train: A mutual information-based approach to fair and robust training. *arXiv preprint arXiv:2002.10234*, 2020. 1

[17] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 3

[18] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 4

[19] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018. 2

[20] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018. 2

[21] Yuxuan Song, Lantao Yu, Zhangjie Cao, Zhiming Zhou, Jian Shen, Shuo Shao, Weinan Zhang, and Yong Yu. Improving unsupervised domain adaptation with variational information bottleneck. *arXiv preprint arXiv:1911.09310*, 2019. 2

[22] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 2

[23] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017. 2, 3

[24] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 3

[25] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S Yu. Visual domain adaptation with manifold embedded distribution alignment. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 402–410, 2018. 3

[26] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3964–3973, 2018. 3, 4