Consensus-Clustering-Based Automatic Distribution Matching for Cross-Domain Image Steganalysis

Ju Jia, Meng Luo, Siqi Ma, Lina Wang, and Yang Liu

Abstract—Image steganalysis is a technique to detect whether an image contains hidden information. Although the existing cross-domain steganalysis methods have been presented to narrow the distribution gap between different domains, it is still challenging to effectively capture the transferable steganalysis representations under the condition of severe distribution shifts. To address this issue, we propose a novel consensus-clustering-based automatic distribution matching scheme, called CADM, which can automatically and accurately match inconsistent distributions in cross-domain steganalysis scenarios. First, the original steganalysis features are clustered by the spatially constrained fuzzy c-means (SCFCM) algorithm with controllable parameters to fully perceive and mine inherent structural relationships. Subsequently, the cluster consensus knowledge is derived from the perspective of intra-domain and inter-domain to facilitate the clustering and the matching. In this way, the representations of weak stego signals can be augmented by identifying cluster centers that can be combined across domains. Ultimately, the cycle-consistent optimization and adaptation is achieved by gradually adjusting the learning strength of well-aligned and poorly-aligned samples to promote the positive transfer of overlapped clusters and prevent the negative transfer of outlier clusters. Furthermore, extensive experiments on various benchmark databases for cross-domain steganalysis demonstrate the superiority of CADM over the current state-of-the-art methods.

Index Terms—Consensus clustering, transferable representations, structural relationships, automatic distribution matching, cross-domain steganalysis

1 Introduction

IMAGE steganography aims to embed secret messages into images by partially replacing pixel values or transform coefficients while maintaining the visual imperceptibility and statistical undetectability [1], [2], [3], [4], [5], [6]. According to the design strategy of embedding modifications, the steganographic methods can be divided into the conventional nonadaptive and modern adaptive steganography, such as nsF5 [1], J-UNIWARD [2], UERD [3], and J-MiPOD [6]. Steganalysis schemes are considered as the countermeasures for detecting steganography, which try to discover the small traces caused by modification operations to determine the existence of hidden information [7], [8]. Nowadays, a large number of steganalysis schemes have emerged to capture steganographic embedding operations by constructing manual features or integrating deep features, mainly including JRM [9], DCTR [10], and SRNet [11]. Fig. 1 gives a typical application scenario of steganography and steganalysis.

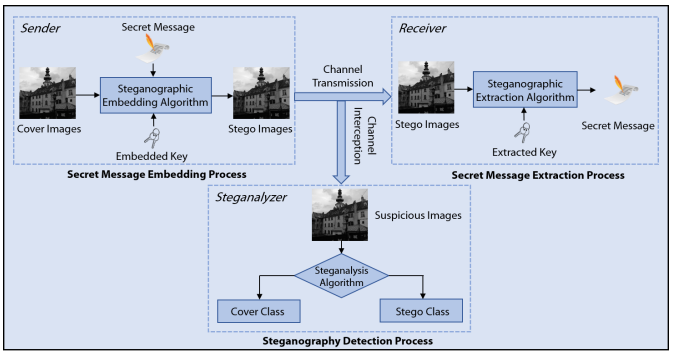


Fig. 1: An example illustrates the implementation process of steganography and steganalysis. Note that although there is no difference in visual observation between the cover images and the stego images, their steganalysis features will be significantly different in statistical distribution.

Although great progresses have been achieved for steganalysis under the condition of distribution matching [7], [8], [9], [10], [11]. However, in more realistic scenarios where the statistical distributions between the training data (i.e., source domain) and testing data (i.e., target domain) are significantly different, which will inevitably lead to the performance degradation of steganography detection [12]. Moreover, in the field of steganalysis, there are many factors that may cause the domain mismatch and distribution shift [13], [14], [15], such as steganographic algorithm, quality factor, payload rate, sample proportion, background content, and so on. From our observation, the main challenging lies in that these factors resulting in the mismatched steganalysis are diverse and complicated, so it is much more difficult to deal with the distribution discrepancy across domains.

- *J. Jia and Y. Liu are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798. E-mail: {jia.ju, yangliu}@ntu.edu.sg.*
- *M. Luo is with the Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115 USA. E-mail: mengluowork@gmail.com.*
- *S. Ma is with the School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2612, Australia. E-mail: siqi.ma@adfa.edu.au.*
- *L. Wang is with the Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China. E-mail: lnawang@163.com.*

Manuscript received XX, 2021; revised XX, 2022.
(Corresponding author: Meng Luo.)

Although some existing methods can effectively handle the distribution mismatch issues in other research areas [16], [17], [18], these approaches are designed specifically under the condition of multiple semantics and multiple classifications. In contrast, steganalysis is essentially a binary classification problem. In this way, directly employing these methods in other fields to address the cross-domain steganalysis problems easily fail to obtain satisfactory results. Moreover, steganalysis focuses on the use of high-frequency residuals to enhance the signal-to-noise ratio of steganographic signals, which can promote the discriminability between the cover sample and the stego sample. However, the combination of co-occurrences of different filter residuals is arbitrarily selected to construct high-dimensional steganalysis representations in cross-domain detection tasks that may introduce the redundant knowledge and thus aggravate the negative transfer. Therefore, the mismatch problems in steganalysis have remarkable differences compared with cross-domain transfer tasks in other fields. For all the above reasons, it is becoming crucial to design a reasonable distribution matching scheme according to these characteristics of mismatched steganalysis.

Recently, a number of research works have been conducted to improve the cross-domain steganalysis performance using distribution adaptation methods [12], [13], [15], [19]. These studies aim to exploit the knowledge in the source domain to assist in predicting the target domain, where the source and target data have similar but different distributions. Most existing approaches are designed to capture and learn transferable knowledge from the source domain, where the main challenge is how to effectively distinguish and utilize the reliable samples (i.e., matched samples across domains) and the unreliable samples (i.e., mismatched samples across domains) in the unaligned distribution space. To achieve this goal, remarkable efforts have been made to perform the sample selection from different perspectives, such as designing suitable criteria [13], [15], incorporating additional estimator [19], and introducing novel optimizer [12].

The previous practices for mismatched steganalysis often pay attention to deal with the distribution discrepancy problems and attempt to capture domain-invariant representations. Unfortunately, these approaches are unable to accurately discover the latent transferable knowledge and explicit correlations between different views due to the distribution shifts. Therefore, most of them tend to suffer from the following drawbacks: (1) these steganography detection methods separately implement feature preprocessing and model training, which makes it difficult to obtain a global optimal solution; (2) the intra-domain and inter-domain discrepancies are treated equally and the associated information (e.g., intrinsic steganographic structure) across domains is not fully exploited, which further results in lower discriminative and transferable representations; (3) the cover and stego samples in the source and target domains may be misaligned because of the disturbing or useless information (e.g., redundant features and outlier samples).

To address the above issues, a novel consensus-clustering-based automatic distribution matching scheme for cross-domain image steganalysis, CADM, is proposed in this paper, which consists of structural relationship exploration, cluster consensus matching, cycle-consistent optimization and adaptation. Fig. 2 illustrates the difference between our scheme and previous methods. We first deal with mismatched data to explore steganographic structure relations by the spatially constrained fuzzy c-means (SCFCM) clustering, and impose a new variation-aware structure loss to unify the steganalysis features across different domains. Then, the cluster consensus matching is proposed to from the perspective of intra-domain and inter-domain to calculate the domain loss, which helps to select and leverage knowledge from the source cluster that is most similar to the target cluster to achieve the adaptive filling of distribution gaps. Finally, the predicted values of the classifiers are input to the joint adaptation layer to adjust the alignment loss, while the cycle-consistent optimization and adaptation can be adopted to further improve the generalization performance by minimizing the total loss. In this way, CADM can facilitate the positive transfer of shared clusters and prevent the negative transfer of outlier clusters through implicit steganographic cue discovery and exploitation. Moreover, extensive experiments show that our CADM outperforms other methods on the challenging cross-domain steganalysis benchmark datasets, which have validated the effectiveness of the proposed scheme. The main contribution of this paper can be briefly summarized as follows:

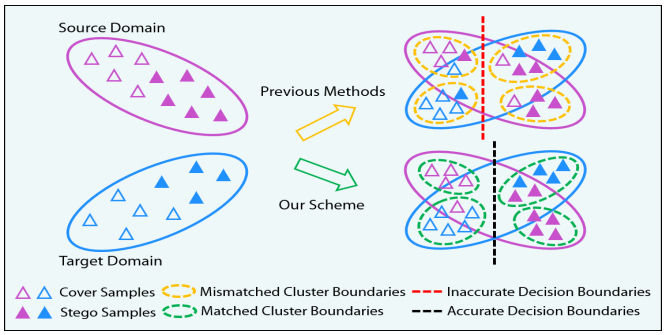


Fig. 2: The difference between the previous methods and the scheme proposed in this paper. Under the condition of distribution shifts, the previous methods tend to produce mismatched cluster boundaries during clustering due to the ignorance of exploring intrinsic structural relationships. However, our scheme aims to make better use of consensus clustering by forming overlapping clusters on both the source and target clusters, thus it can generate matched cluster boundaries automatically and accurately.

1) To capture structural relationships from the potential domains, a spatially constrained fuzzy c-means clustering that can fully exploit both the correlation and complementarity is introduced, which is conducive to the discovery of domain-invariant steganographic modification cues.
2) The cluster consensus matching is proposed to improve the quality of steganalysis features from two levels, i.e., the intra-domain level, which determines the number of clusters to make the clustering more reasonable, and the inter-domain level, which identifies cluster centers that can be combined across domains.
3) A cycle-consistent optimization and adaptation strategy is designed to enhance the comprehensive performance via encouraging a collaboration between the source and target clusters, which also promotes the generalization and transferability of the knowledge learned from the cover and stego samples.

4) To the best of our knowledge, this work is the first time that a flexible and controllable framework has been proposed for cross-domain steganalysis. Comprehensive experiments demonstrate that our proposed CADM can deeply excavate latent information and automatically perform distribution matching to curb the threat of negative transfer.

The rest of this article is organized as follows. Section 2 provides the related work and preliminary knowledge. Section 3 describes the proposed CADM scheme for cross-domain image steganalysis. Section 4 validates the effectiveness and superiority of our scheme by sufficient experiments and various comparison approaches. The analysis and discussion of CADM are given in Section 5. Finally, the conclusions and future work are summarized in Section 6.

2 Related Work and Preliminaries

2.1 Transferable Knowledge Discovery

The existing solutions address the distribution shift problems by exploring latent domains to discover the transferable knowledge, which can be classified into four categories, namely: 1) instance-level methods; 2) classifier-level methods; and 3) feature-level methods. The instance-based methods usually attempt to assign larger weights to significant samples and smaller weights to unimportant ones. For example, Xia *et al.* [20] proposed a flexible instance weighting framework to potentially correct the distribution bias by adjusting the dominant factors in domain adaptation. The classifier-level adaptation is to train a series of classifier combinations from source and target domains, and then to achieve the final detection of the target domain by selecting the optimal ensemble classifier. Following this, Ren *et al.* [21] leveraged multiple auxiliary classifiers to process the source and target data to further mitigate the distribution discrepancy from the perspective of classifier property. In addition, as one of the most commonly used techniques, the feature-based methods are encouraged to learn domain-invariant or domain-shared feature representations by aligning the distribution differences across domains. This idea has attracted increasing attention in the recent years and various approaches have been proposed, including designing handcrafted features (shallow models) [22] and learning deep features (deep architectures) [23], [24], [25]. There are some recent works that add adaptation layers or subnetworks for domain shift problem [26], [27]. Specifically, Li *et al.* [28] proposed a deep residual correction network to match the feature distributions across different domains. However, most of the above methods learn transferable representations from a limited source domain, which is often too ideal to be satisfied in real scenarios. Therefore, our proposed CADM considers a more practical situation, which can effectively deal with a more challenging distribution shift problem.

2.2 Cross-Domain Distribution Alignment

Since the multiple source domain adaptation (MSDA) is an extension of the single source domain adaptation (SSDA), it can not only explore intra-domain correlation from a single source domain, but also capture inter-domain correlation from multiple source domains [29]. In practice, we are more likely to obtain a source dataset containing multiple domains, while acting on an unlabeled target dataset, which enables us to transfer multi-source knowledge representations from the source to target. However, the previous SSDA approaches, straightforward merging multiple different source domains into a single source domain, are prone to the negative transfer for the MSDA problems. Thus, many researchers have paid more attention to investigating how to effectively address MSDA problems in visual perception scenarios. Recently, some representative MSDA methods are proposed, such as moment matching network (MMN) [30], adversarial domain aggregation network (ADAN) [17], and multi-source distilling domain adaptation (MDDA) [31]. All these MSDA methods mainly rely on a deep feature learning network to equivalently transform the multiple sources and target data into the common subspace. A pre-trained deep model aims to align the inconsistent distributions of source-target data, which is a common way adopted in MMN and MDDA. MMN dynamically incorporates the moment component into deep network based on the error bound to alleviate the domain discrepancy, while MDDA pre-trains and fine-tunes a classifier to adjust the target distribution to the source ones using a weighting strategy. ADAN constructs an adapted domain for each source while performing the alignment operation at the pixel-level towards the target, and then guarantees diverse adapted domains more closely aggregated at the feature-level. Different from these works, our CADM explores the structural relationships to capture the steganographic structure knowledge by the SCFCM clustering. By leveraging the cluster consensus matching across different domains, CADM can effectively perform the adaptive filling of distribution gaps between the source and target clusters. Furthermore, our scheme is designed to minimize the total loss through the cycle-consistent optimization and adaptation, which can guarantee automatic distribution matching and thus substantially different from other methods.

2.3 Domain Consistency Evaluation

The main challenge of domain consistency evaluation is how to determine the number of source and target clusters in steganalysis scenarios. The reason is that the number of clusters in the source domain and target domain in other research areas is usually determined according to the number of corresponding classes. However, steganalysis is a binary classification problem, so it is obviously unreasonable to directly obtain the number of clusters based on the number of categories. To tackle this issue, one feasible solution is to utilize existing clustering quality metrics to acquire the number of clusters. Unfortunately, these methods are designed for single-domain settings and cannot fully take cross-domain correlation knowledge into consideration. Thus, we introduce a criterion, domain consistency evaluation, which exploits the sample-level consistency to determine the number of clusters in source and target domains, thereby constructing discriminative clusters.

As illustrated in Fig. 3, for each sample from the source cluster, we attempt to explore and match the overlapped clusters in the target domain, and further calculate the consensus degree between them, that is, the proportion of samples holding consistent labels across domains. The

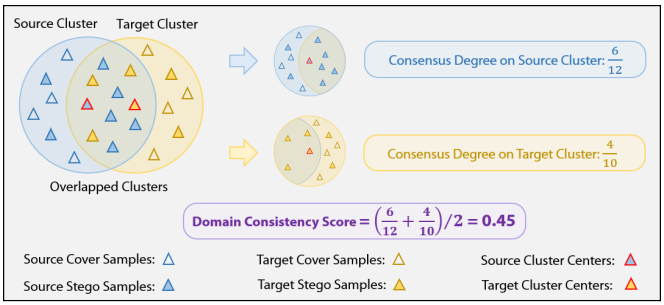


Fig. 3: Illustration of domain consistency score. For each sample from overlapped clusters, we aim to search for the nearest cluster in the other domain. Then, the domain consistency score is computed as the proportion of samples that achieve consensus.

consistency of paired clusters can be evaluated by collecting samples that achieve consensus.

Definition 1 (Consensus Degree): Given a pair of overlapped clusters $\{\mathbf{x}_{s_p}^i\}_{i=1}^{n_{sp}}$ and $\{\mathbf{x}_{t_q}^j\}_{j=1}^{n_{tq}}$ with the corresponding cluster center samples $\mathbf{z}_{s_p}$ and $\mathbf{z}_{t_q}$, s_p and t_q denote the p-th and q-th clusters in the source and target domains, respectively, and n_{s_p} and n_{t_q} are the number of samples in the corresponding clusters. We intend to measure the degree of consensus at the sample level from two complementary perspectives, namely the source perspective and the target perspective. To obtain the consensus degree on the source perspective, we calculate the similarity between each sample in the p-th source cluster and all target cluster centers $\{\mathbf{z}_{t_1}, ..., \mathbf{z}_{t_Q}\}$:

$$f_{s_p}(i,q) = S(\mathbf{x}_{s_p}^i, \mathbf{z}_{t_q}) = \frac{\mathbf{x}_{s_p}^i \cdot \mathbf{z}_{t_q}}{\left\|\mathbf{x}_{s_p}^i\right\| \left\|\mathbf{z}_{t_q}\right\|}, \ i \in [1, n_{s_p}], \ q \in [1, Q], \tag{1}$$

where Q is the total number of clusters in the target domain. Subsequently, the consensus degree on the p-th source cluster can be calculated by the proportion of samples that achieve consensus:

$$CD_{s_p} = \frac{\sum_{i=1}^{n_{s_p}} \mathbb{I}\{\arg\max_{\mathbf{y}_{s_p}^i} (f_{s_p}(i,q)) = \mathbf{y}_{t_q}\}}{n_{s_p}}, \tag{2}$$

where $\mathbf{y}_{s_p}^i$ is the label vector of source sample $\mathbf{x}_{s_p}^i$, and $\mathbf{y}_{t_q}$ denotes the label vector of target cluster center sample $\mathbf{z}_{t_q}$. $\mathbb{I}(\cdot)$ is an indicator function, indicating that if $\arg\max_{\mathbf{y}_{s_p}^i} (f_{s_p}(i,q)) = \mathbf{y}_{t_q}$ is true, the value of $\mathbb{I}\{\arg\max_{\mathbf{y}_{s_p}^i} (f_{s_p}(i,q)) = \mathbf{y}_{t_q}\}$ is 1, otherwise 0. Analogously, the consensus degree on the q-th target cluster can be computed by CD_{t_q}.

Definition 2 (Domain Consistency Score): The domain consistency score DCS of this overlapped clusters is obtained by the average of scores from two perspectives as follows:

$$DCS(s_p, t_q) = \frac{CD_{s_p} + CD_{t_q}}{2}. \tag{3}$$

More generally, the total domain consistency score is calculated as the mean of consensus degree for all overlapped pairs of clusters.

To specify the number of source and target clusters P and Q, the multiple clusterings are performed with different P and Q to obtain the optimal number of clusters according to the domain consistency score. Concretely, among the various instantiations of P and Q, the domain consistency score is calculated for each one, and then the instantiation of P and Q with the highest score is selected for subsequent experiments.

3 The Proposed CADM Scheme

In this section, we describe the proposed CADM algorithm for cross-domain image steganalysis tasks in detail. The architecture and overview of CADM are illustrated in Fig. 4.

3.1 Basic Notations and Concepts

In this article, the source and target domains are represented by subscript s and t. The datasets in the source domain are described as $D_s = \{(\mathbf{X}_{s_p}, \mathbf{y}_{s_p})\}_{p=1}^{P}$, where P denotes the total number of clusters in the source domain. $(\mathbf{X}_{s_p}, \mathbf{y}_{s_p})$ is the sample-label data pair in the p-th source cluster, where $\mathbf{X}_{s_p} \in \mathbb{R}^{d \times n_{s_p}}$ (d-dimensional data) denotes the sample matrix, and $\mathbf{y}_{s_p}$ represents the label vector. In addition, n_{s_p} is the number of samples in the p-th source cluster. The datasets of target domain are denoted as $D_t = \{(\mathbf{X}_{t_q}, \mathbf{y}_{t_q})\}_{q=1}^{Q}$, where $\mathbf{X}_{t_q} \in \mathbb{R}^{d \times n_{t_q}}$ is the data matrix for the limited labeled samples, $\mathbf{y}_{t_q}$ is the label vector corresponding to the $\mathbf{X}_{t_q}$, and Q denotes the total number of clusters in the target domain. The number of samples in the q-th target cluster is represented as n_{t_q}. Given a matrix $\mathbf{B} = [b_{ij}]$, we denote $\mathbf{b}^i$ as its i-th row vector and $\mathbf{b}_j$ as its j-th column vector. The l_1-norm, l_2-norm, and Frobenius norm of the matrix $\mathbf{B} \in \mathbb{R}^{g \times h}$ are denoted as $\|\mathbf{B}\|_1 = \sum_{i=1}^{g}\sum_{j=1}^{h}|b_{ij}|$, $\|\mathbf{B}\|_2 = (\sum_{i=1}^{g}\sum_{j=1}^{h}|b_{ij}|^2)^{\frac{1}{2}}$, and $\|\mathbf{B}\|_F = (\sum_{j=1}^{h}\|\mathbf{b}_j\|_2^2)^{\frac{1}{2}}$, respectively. For a quick reference, we summarize the commonly used notations and their descriptions throughout the paper in Table 1.

3.2 Structural Relationship Exploration

To discover the hidden structural relationships from the mismatched data, we propose an improved method called spatially constrained fuzzy c-means (SCFCM) clustering. The key insight is to fully leverage some measures to capture variation relations between the source and target samples. The conventional fuzzy c-means with the approximation partitioning approach works well when the available samples obey the approximate distribution, while the spatial clustering performance degrades when some samples are corrupted by various cross-domain interferences. In this case, the distribution difference across domains cannot be accurately described so that the partition matrix always contains imprecise (unreliable or uncertain) components with high probability. Moreover, unreliable samples are the main factors that lead to negative transfer, so it is necessary to eliminate the interference in the process of exploring structural relations. As a consequence, we consider spatial constraints based on the minimization of a cost function that can better explore the structural relationships of the data in the distribution shift.

Let $\mathbf{X} = [\mathbf{x}_s^1, ..., \mathbf{x}_s^{n_s}, \mathbf{x}_t^1, ..., \mathbf{x}_t^{n_t}]$ be the matrix formed by a set of data samples of the vector space across different domains, where n_s and n_t denote the total number of samples in the source and target domain respectively. Although we have no prior information about the cluster number $c \in [1, n_s + n_t]$, we assume that the matrix $\mathbf{X}$ can be decomposed into c fuzzy clustering to obtain a fuzzy

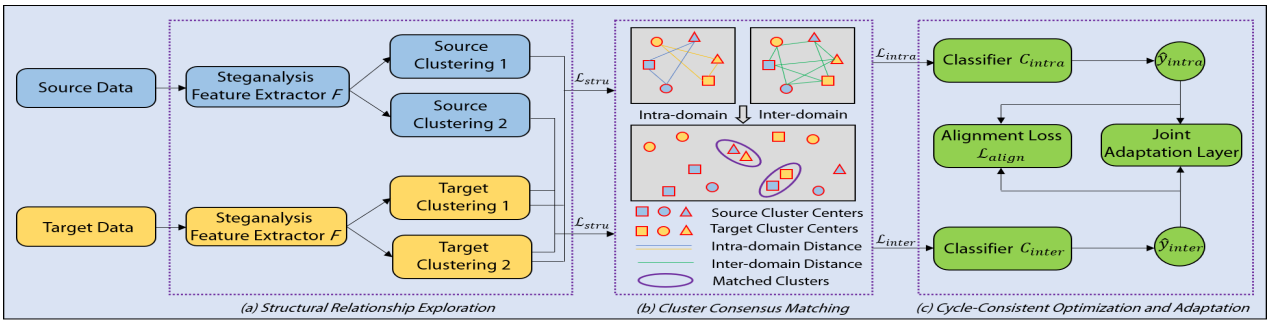


Fig. 4: Illustration of the proposed CADM scheme for cross-domain image steganalysis. It can be divided into three main parts: (a) the exploration of structural relationships is guided and conducted by the spatially constrained fuzzy c-means clustering from multiple potential candidates; (b) the cluster consensus matching from the intra-domain and inter-domain perspectives is leveraged to make the clusters more separated and thus identify the discriminative clusters and shared clusters; and (c) the cycle-consistent optimization and adaptation aims to optimize the steganalysis feature extractor F by minimizing the total loss $\mathcal{L}_{total}$, including structure loss $\mathcal{L}_{stru}$, intra-domain loss $\mathcal{L}_{intra}$, inter-domain loss $\mathcal{L}_{inter}$, and alignment loss $\mathcal{L}_{align}$, to make the diverse distributions keep consistent as large as possible in the overall training procedure. In the test phase, we first extract steganalysis features by optimal F for the given target samples. Then, the final prediction results are the average value of two estimates $\bar{y} = \frac{1}{2}(\hat{y}_{intra} + \hat{y}_{inter})$. Best viewed in color.

TABLE 1: Descriptions of Notations Used in the Paper

Notation	Description	Notation	Description
$\mathbf{x}_{s_p}^i$	The i-th sample in the p-th source cluster	$\mathbf{W}$	Intermediate matrix
$\mathbf{x}_{t_q}^j$	The j-th sample in the q-th target cluster	$\mathbf{P}$	Orthogonal matrix
$\mathbf{y}_{s_p}^i$	The i-th label vector of the p-th source cluster	$\boldsymbol{\varphi}_p$	Parameter vector of the feature extractor
$\mathbf{y}_{t_q}^j$	The j-th label vector of the q-th target cluster	$\boldsymbol{\varphi}_s$	Parameter vector of the structure loss
$\mathbf{z}_{s_p}$	Center vector of the p-th source cluster	$\boldsymbol{\varphi}_d$	Parameter vector of the domain loss
$\mathbf{z}_{t_q}$	Center vector of the q-th target cluster	$\boldsymbol{\varphi}_a$	Parameter vector of the alignment loss
n_{s_p}	Number of samples in the p-th source cluster	$\mathcal{L}_{stru}$	Structure loss
n_{t_q}	Number of samples in the q-th target cluster	$\mathcal{L}_d$	Domain loss
P	Total number of clusters in the source domain	$\mathcal{L}_{align}$	Alignment loss
Q	Total number of clusters in the target domain	ρ_l	The l-th tuning parameter to control the weight
d	Data dimension	k_l	The l-th basis kernel
c	Number of clusters	L	Total number of basis kernels
$\mathbf{I}$	Identity matrix	C_{intra}	Intra-domain classifier
$\mathbf{U}$	Partition matrix	C_{inter}	Inter-domain classifier
$\mathbf{V}$	Prototype matrix	$\lambda_1, \lambda_2, \lambda_3$	Trade-off parameters of the corresponding loss terms
CD_{s_p}	Consensus degree on the p-th source cluster	DCS	Domain consistency score
CD_{t_q}	Consensus degree on the q-th target cluster	SCFCM	Spatially constrained fuzzy c-means clustering
PCA	Principal component analysis	CLS	Constrained least squares
MMD	Maximum mean discrepancy	CCM	Cluster consensus matching
DMMD	Dynamic maximum mean discrepancy	RKHS	Reproducing kernel Hilbert space
SSDA	Single source domain adaptation	SGD	Stochastic gradient descent
MSDA	Multiple source domain adaptation	SIDA	Sample-imbalanced domain adaptation
QF	Quality factor	NA	No adaptation
JRM	JPEG domain rich model	GSL	Guide subspace learning
SRNet	Steganalysis residual network	IMFA	Iterative multi-order feature alignment
DCTR	Discrete cosine transform residual	THFSL	Transferable heterogeneous feature subspace learning
MEDA	Manifold embedded distribution alignment	ACFL	Adaptive cost-sensitive feature learning

c-partition. A fuzzy c-partition can be properly described in the form of a matrix, denoted as $\mathbf{U}$, which is called the partition matrix. The generic element of the partition matrix, u_{ik}, represents the membership degree of the sample point $\mathbf{x}_k$ in fuzzy cluster i. The idea of the fuzzy c-means algorithm is based on the minimization of the weighted sum of squared error, which can be formulated as:

$$J(\mathbf{X}, \mathbf{U}, \mathbf{V}; m) = \sum_{i=1}^{c} \sum_{k=1}^{n_s+n_t} (u_{ik})^m \|\mathbf{x}_k - \mathbf{v}_i\|_{\mathbf{A}_i}^2, \quad (4)$$

where $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_c] \in \mathbb{R}^{d \times c}$ is the prototype matrix, and m denotes the weighting exponent which can account for the fuzziness of c clusters. The metric matrix $\mathbf{A}_i$ is used to estimate the distance between the sample points and the prototypes.

The $D_{ik(\mathbf{A}_i)}^2$ represents the Mahalanobis metric, which can be calculated by a squared inner-product distance norm as follows:

$$D_{ik(\mathbf{A}_i)}^2 = \|\mathbf{x}_k - \mathbf{v}_i\|_{\mathbf{A}_i}^2 = (\mathbf{x}_k - \mathbf{v}_i)^T \mathbf{A}_i (\mathbf{x}_k - \mathbf{v}_i). \quad (5)$$

Since the standard fuzzy c-means algorithm relies on Euclidean metric, i.e., $\mathbf{A}_i = \mathbf{I}$ (identity matrix), $i = 1, 2, ..., c$, and it can be omitted later in this paper. The work in [32] reveals that the pair $(\mathbf{U}, \mathbf{V})$ can make the function $J(\mathbf{X}, \mathbf{U}, \mathbf{V}; m)$ locally minimum only if the following conditions are met simultaneously in Eq. (6) and Eq. (7).

$$u_{ik} = [\sum_{j=1}^{c} (\frac{D_{ik}}{D_{jk}})^{\frac{2}{m-1}}]^{-1}, 1 \le i \le c, 1 \le k \le n_s + n_t, \quad (6)$$

and

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n_s+n_t} (u_{ik})^m \mathbf{x}_k}{\sum_{k=1}^{n_s+n_t} (u_{ik})^m}, 1 \le i \le c. \quad (7)$$

Let $\mathbf{W} \in \mathbb{R}^{d \times (n_s+n_t)}$ and $\mathbf{P} \in \mathbb{R}^{d \times d}$ be intermediate and orthogonal matrices respectively. We consider the transformation shown in Eq. (8).

$$\mathbf{W} = \mathbf{P}\mathbf{X}. \quad (8)$$

Algorithm 1 Spatially Constrained Fuzzy C-Means Clustering

Input: data matrix $\mathbf{X}$, initial partition matrix $\mathbf{U}_0$, maximum error parameter α, maximum iteration number $MaxIter$;
Output: updated partition matrix $\mathbf{U}$, updated prototype matrix $\mathbf{V}$;

1: Randomly generate the initial partition matrix $\mathbf{U}_0 = [u_{ik}^0]$;
2: Calculate the initial prototype matrix $\mathbf{V}_0$, according to Eq. (7), using $\mathbf{U}_0$;
3: Iteration $T = 1$;
4: For $1 \leq k \leq n_s + n_t$, solve Eq. (10) to obtain $\mathbf{U} = [u_{ik}]$;
5: Update $\mathbf{V}$, using $\mathbf{U}$;
6: If $\max_{ik}|u_{ik} - u_{ik}^0| > \alpha$ and $T < MaxIter$;
(a) $\mathbf{U}_0 \leftarrow \mathbf{U}$; (b) $T \leftarrow T + 1$; (c) Go to step 4;
7: Return $\mathbf{U}$ and $\mathbf{V}$.

Replacing $\mathbf{x}_k$ by $\mathbf{P}\mathbf{x}_k$ in Eq. (7), we can obtain $\mathbf{v}_i'$. Thus, $\mathbf{V}' = \mathbf{P}\mathbf{V}$. Then, Eq. (4) can be transformed into the following form by setting $m = 2$:

$$J(\mathbf{W}, \mathbf{U}, \mathbf{V}') = \|\mathbf{W} - \mathbf{V}'\mathbf{U}\|_F^2 = \|\mathbf{P}(\mathbf{X} - \mathbf{V}\mathbf{U})\|_F^2 = J(\mathbf{X}, \mathbf{U}, \mathbf{V}). \tag{9}$$

In particular, if the matrix $\mathbf{P}$ performs a principal component analysis (PCA) on the sample matrix $\mathbf{X}$, we can alternatively calculate $\mathbf{U}$ by means of irrelevant $\mathbf{W}$, thus preventing the clustering process from being influenced by potential multicollinearity problems among selected original features. Furthermore, in the actual experiment, we consider using variance-adjusted prediction of the sample matrix $\mathbf{X}$ to estimate the partition matrix $\mathbf{U}$. In this way, our improved fuzzy clustering method can handle high-dimensional data with less computation.

The optimization of $J(\mathbf{X}, \mathbf{U}, \mathbf{V})$ in Eq. (9), given $\mathbf{V}$, is equivalent to minimizing

$$\|\mathbf{x}_k - \mathbf{V}\mathbf{u}_k\|_2^2, 1 \leq k \leq n_s + n_t, \text{s.t.} \|\mathbf{u}_\text{k}\|_1 = 1, \quad 0 \leq u_{ik} \leq 1, 1 \leq i \leq c. \tag{10}$$

This means that the partition matrix $\mathbf{U}$ can be obtained by solving a set of the constrained least squares (CLS) problems with both equality and inequality constraints. Formally, based on the spatially constrained fuzzy c-means clustering, we define the structure loss as:

$$\mathcal{L}_{stru}(\mathbf{x}_k, \boldsymbol{\varphi}_p, \boldsymbol{\varphi}_s) = \frac{1}{n_s}\sum_{i=1}^{c}\sum_{k=1}^{n_s}[J(\mathbf{x}_s^k, \mathbf{u}_k, \mathbf{v}_i)]^2 + \frac{1}{n_t}\sum_{i=1}^{c}\sum_{k=1}^{n_t}[J(\mathbf{x}_t^k, \mathbf{u}_k, \mathbf{v}_i) - 1]^2, \tag{11}$$

where $\boldsymbol{\varphi}_p$ denotes the parameter vector of the feature extractor, and $\boldsymbol{\varphi}_s$ is the parameter vector of the structure loss associated with $\mathbf{u}_k$ and $\mathbf{v}_i$. The procedure of SCFCM clustering is summarized in Algorithm 1. Note that in order to improve the search efficiency, we adopt a stop rule in actual application, that is, stopping the search once the consistency score drops continuously, and fixing the c once it keeps a certain value after a specific number of times. Its effectiveness has been proved experimentally in Section 5.1.

3.3 Cluster Consensus Matching

The main challenge of cross-domain knowledge transfer lies in how to separate reliable samples from unreliable samples across different distributions. Unlike the previous work [19] on identifying reliable samples at sample level, the goal of this paper is to mine reliable and unreliable samples simultaneously with discriminative clusters. Accordingly, a crucial question naturally arises: how to associate reliable clusters with similar distributions from both domains? To achieve this, we propose a cluster consensus matching (CCM) mechanism to link reliable samples from different clusters by mining consistency at the distribution level.

As shown in Fig. 4, for each cluster center, we calculate the intra-domain and inter-domain distance to search for the matched cluster center in the other domain. If two clusters achieve consensus, i.e., both serve as the matched centers for each other simultaneously, then such a pair of clusters is regarded as reliable clusters. The intuition behind is that reliable clusters tend to have larger intra-domain distances but smaller inter-domain distances, which is beneficial to perform the cluster consensus matching. In addition, to guarantee this assumption, the domain consistency evaluation based on sample-level consensus is used to improve the effectiveness of CCM, which has been described in detail in Section 2.3.

The maximum mean discrepancy (MMD) is a distance function defined between probability distributions in a particular metric space. We describe an improved MMD metric which is called the dynamic MMD (DMMD), to measure the similarity by calculating the distance between the source and target distributions. Moreover, each source domain is given a weight based on minimizing a distance measure between the probability density functions of the source and target domains.

The MMD is defined as the squared distance between the means of different data distributions in a reproducing kernel Hilbert space (RKHS) using function $\phi(\cdot)$ and can be formulated as:

$$\text{MMD}^2(s_p, t_q) = \sup_{\|\phi\|_\mathcal{H} \leq 1} \|E_{\mathbf{x}_{s_p} \sim s_p}[\phi(\mathbf{x}_{s_p})] - E_{\mathbf{x}_{t_q} \sim t_q}[\phi(\mathbf{x}_{t_q})]\|_\mathcal{H}^2, \tag{12}$$

where $E_{\mathbf{x}_\gamma \sim \gamma}[\cdot]$ describes the expectation with regard to the distribution γ, $\gamma \in \{s_p, t_q\}(p \in [1, P], q \in [1, Q])$, and $\|\phi\|_\mathcal{H} \leq 1$ represents a series of functions in the unit ball of a RKHS $\mathcal{H}$. $D_{s_p} = \{\mathbf{x}_{s_p}^i\}_{i=1}^{n_{s_p}}$ and $D_{t_q} = \{\mathbf{x}_{t_q}^j\}_{j=1}^{n_{t_q}}$ denote the sample sets extracted from the distributions s_p and t_q, respectively. An empirical evaluation method of MMD is presented as:

$$\text{MMD}^2(D_{s_p}, D_{t_q}) = \|\frac{1}{n_{s_p}}\sum_{i=1}^{n_{s_p}}\phi(\mathbf{x}_{s_p}^i) - \frac{1}{n_{t_q}}\sum_{j=1}^{n_{t_q}}\phi(\mathbf{x}_{t_q}^j)\|_\mathcal{H}^2, \tag{13}$$

where $\phi(\cdot)$ represents the feature map corresponding to the kernel map $k(\mathbf{x}_{s_p}, \mathbf{x}_{t_q}) = \langle\phi(\mathbf{x}_{s_p}), \phi(\mathbf{x}_{t_q})\rangle$. The kernel $k(\mathbf{x}_{s_p}, \mathbf{x}_{t_q})$ is usually calculated by a convex combination of basis kernels as follows:

$$k(\mathbf{x}_{s_p}, \mathbf{x}_{t_q}) = \sum_{l=1}^{L}\rho_l k_l(\mathbf{x}_{s_p}, \mathbf{x}_{t_q}), \text{s.t.} \rho_l \geq 0, \sum_{l=1}^{L}\rho_l = 1, \tag{14}$$

where ρ_l is the l-th tuning parameter, $k_l(\mathbf{x}_{s_p}, \mathbf{x}_{t_q})$ denotes the l-th basis kernel, and L represents the total number of basis kernels.

However, the traditional MMD is not robust to the class weight bias, which can be interpreted by further decomposi-

tion of $p_{s_p}(\mathbf{x}_{s_p})$ and $p_{t_q}(\mathbf{x}_{t_q})$ into conditional distributions,

$$\begin{aligned} p_\gamma(\mathbf{x}_\gamma) &= p(y_\gamma{=}\varepsilon_c)p(\mathbf{x}_\gamma|y_\gamma{=}\varepsilon_c) + p(y_\gamma{=}\varepsilon_s)p(\mathbf{x}_\gamma|y_\gamma{=}\varepsilon_s) \\ &= w_\gamma^{\varepsilon_c} p(\mathbf{x}_\gamma|y_\gamma{=}\varepsilon_c) + w_\gamma^{\varepsilon_s} p(\mathbf{x}_\gamma|y_\gamma{=}\varepsilon_s), \end{aligned} \tag{15}$$

where ε_c and ε_s represent the cover samples and the stego samples, respectively. Specifically, $w_{s_p}^{\varepsilon_c} = p(y_{s_p} = \varepsilon_c)$ and $w_{s_p}^{\varepsilon_s} = p(y_{s_p} = \varepsilon_s)$ represent the prior probabilities (i.e., weights and biases) of the p-th source cluster for the cover class and the stego class, respectively.

To suppress the influence of class weight bias across domains, we construct a reference source cluster distribution $p_{s_p,\beta_\varepsilon}(\mathbf{x}_{s_p})$ to compare the difference between the source and target clusters, where ε denotes the category of samples. For this purpose, we ensure that $p_{s_p,\beta_\varepsilon}(\mathbf{x}_{s_p})$ has the same class weight as the target cluster, but maintains the class conditional distribution in the p-th source cluster. Let $\beta_\varepsilon = \frac{w_{t_q}^\varepsilon}{w_{s_p}^\varepsilon}$, we can define $p_{s_p,\beta_\varepsilon}(\mathbf{x}_{s_p})$ to alleviate the effect of class weight bias as follows:

$$p_{s_p,\beta_\varepsilon}(\mathbf{x}_{s_p}) = \beta_{\varepsilon_c} w_{s_p}^{\varepsilon_c} p(\mathbf{x}_{s_p}|y_{s_p}{=}\varepsilon_c) + \beta_{\varepsilon_s} w_{s_p}^{\varepsilon_s} p(\mathbf{x}_{s_p}|y_{s_p}{=}\varepsilon_s). \tag{16}$$

Therefore, the empirical equation of DMMD between the source and target conditional distributions can be expressed as:

$$\begin{aligned} \mathrm{DMMD}^2(D_{s_p}, D_{t_q}) = \| &\frac{1}{\sum_{i=1}^{n_{s_p}} \beta_{y_{s_p}^i}} \sum_{i=1}^{n_{s_p}} \beta_{y_{s_p}^i} \phi(\mathbf{x}_{s_p}^i) \\ &- \frac{1}{\sum_{j=1}^{n_{t_q}} \beta_{y_{t_q}^j}} \sum_{j=1}^{n_{t_q}} \beta_{y_{t_q}^j} \phi(\mathbf{x}_{t_q}^j) \|_2^2, \end{aligned} \tag{17}$$

Because the formulation of DMMD in Eq. (17) is defined according to the whole source and target data, it is unsuitable for achieving the neural network-based deep cluster adaptation through mini-batch stochastic gradient descent (SGD). Assuming $n_{s_p} = n_{t_q} = \tau$, we give an approximation of the linear time complexity for DMMD in Eq. (18).

$$\mathrm{DMMD}^2(D_{s_p}, D_{t_q}) = \frac{2}{\tau} \sum_{i=1}^{\frac{\tau}{2}} f(\mathbf{r}_i), \tag{18}$$

where $\mathbf{r}_i$ is a quad-tuple operator defined as $\mathbf{r}_i = (\mathbf{x}_{s_p}^{2i-1}, \mathbf{x}_{s_p}^{2i}, \mathbf{x}_{t_q}^{2i-1}, \mathbf{x}_{t_q}^{2i})$. Then, $f(\mathbf{r}_i)$ can be represented as:

$$\begin{aligned} f(\mathbf{r}_i) = &\beta_{y_{s_p}^{2i-1}} k(\mathbf{x}_{s_p}^{2i-1}, \mathbf{x}_{s_p}^{2i}) + \beta_{y_{s_p}^{2i}} k(\mathbf{x}_{t_q}^{2i}, \mathbf{x}_{t_q}^{2i-1}) \\ &- \beta_{y_{s_p}^{2i-1}} k(\mathbf{x}_{s_p}^{2i-1}, \mathbf{x}_{t_q}^{2i}) - \beta_{y_{s_p}^{2i}} k(\mathbf{x}_{s_p}^{2i}, \mathbf{x}_{t_q}^{2i-1}). \end{aligned} \tag{19}$$

The approximation is in summation form and thus can be easily optimized by mini-batch SGD. Specifically, we define the domain loss based on DMMD, which can be represented as:

$$\begin{aligned} \mathcal{L}_d(\boldsymbol{\varphi}_p, \boldsymbol{\varphi}_d) = &\frac{1}{2} \sum_{p=1}^{P} \sum_{p'=1(p\neq p')}^{P} \mathrm{DMMD}(D_{s_p}, D_{s_{p'}}) \\ &+ \frac{1}{2} \sum_{q=1}^{Q} \sum_{q'=1(q\neq q')}^{Q} \mathrm{DMMD}(D_{t_q}, D_{t_{q'}}) \\ &+ \sum_{p=1}^{P} \sum_{q=1}^{Q} \mathrm{DMMD}(D_{s_p}, D_{t_q}), \end{aligned} \tag{20}$$

where $\boldsymbol{\varphi}_d$ denotes the parameter vector of the domain loss associated with the input distributions. The first two terms in Eq. (20) represents the intra-domain loss $\mathcal{L}_{intra}$ and the last term measures the inter-domain loss $\mathcal{L}_{inter}$. In this way, DMMD quantitatively measures feature distributions across different domains to select suitable source knowledge, thus can effectively promote the positive transfer of the useful source data and control the negative transfer of the redundant source data.

Empirically, we find that cluster consensus matching tends to divide the samples from similar distributions into multiple clusters at the beginning, which is also called as over-clustering. The reason is that in order to obtain a higher consistency score, more precise and fine-grained matching between clusters is required. Therefore, at the beginning, cluster consensus matching is more inclined to small clusters with stable samples (i.e. less affected by the distribution shift), which may temporarily make the number of clusters much larger than the final one. With the progress of deep adaptation and alignment, the number of clusters gradually decreases and converges to a finite fixed number after a period of training.

3.4 Cycle-Consistent Optimization and Adaptation

To improve the generalization ability, we present the intra-domain classifier C_{intra} to exploit useful information among source domains for knowledge transfer. Since the inter-domain classifier C_{inter} is only trained on the loss between the source and target domains, the knowledge from source domains is likely to be redundant and may easily cause negative transfer. Intuitively, the learning process for C_{intra} and C_{inter} can be regarded as the discovery and integration of the transferable knowledge across domains, thus they play a prominent role in reducing the alignment loss. In addition, we introduce the joint adaptation layer [33] to ensure the mutual learning between classifiers C_{intra} and C_{inter} through knowledge adaptation. The calculation of alignment loss based on the above analysis can be written as:

$$\begin{aligned} \mathcal{L}_{align}(\boldsymbol{\varphi}_p, \boldsymbol{\varphi}_d, \boldsymbol{\varphi}_a) = &\frac{1}{(n_s+n_t)^2} \sum_{k=1}^{n_s+n_t} \sum_{k'=1}^{n_s+n_t} \| p_{inter}(y|\mathbf{x}_k) \\ &- p_{intra}(y|\mathbf{x}_{k'}) \|_2^2, \end{aligned} \tag{21}$$

where $p_{inter}(y|\mathbf{x}_k)$ and $p_{intra}(y|\mathbf{x}_{k'})$ respectively denote the predicted category distribution of the sample data by the classifier C_{inter} and C_{intra}, and $\boldsymbol{\varphi}_a$ is the parameter vector of the alignment loss associated with the classifiers.

In order to minimize the total loss, a cycle-consistent optimization and adaptation strategy is proposed to make the mismatched distributions close. Therefore, our ultimate goal can be transformed into solving the following optimization problem of the objective function [34]:

$$\begin{aligned} \min_{\boldsymbol{\varphi}_p, \boldsymbol{\varphi}_s, \boldsymbol{\varphi}_d, \boldsymbol{\varphi}_a} &\lambda_1 \mathcal{L}_{stru}(\mathbf{x}_k, \boldsymbol{\varphi}_p, \boldsymbol{\varphi}_s) + \lambda_2 \mathcal{L}_d(\boldsymbol{\varphi}_p, \boldsymbol{\varphi}_d) \\ &+ \lambda_3 \mathcal{L}_{align}(\boldsymbol{\varphi}_p, \boldsymbol{\varphi}_d, \boldsymbol{\varphi}_a), \end{aligned} \tag{22}$$

where λ_1, λ_2, and λ_3 represent trade-off parameters for the corresponding loss functions, respectively.

We first exploit the SCFCM to explore the structural relationships of mismatched data, then design the DMMD to measure the distribution discrepancy across different clusters, and finally develop the cycle-consistent optimization and adaptation to promote the generalization ability and

Algorithm 2 Consensus-Clustering-Based Automatic Distribution Matching

Input: source data $D_s = \{(\mathbf{X}_{s_p}, \mathbf{y}_{s_p})\}_{p=1}^{P}$, target data $D_t = \{(\mathbf{X}_{t_q}, \mathbf{y}_{t_q})\}_{q=1}^{Q}$, steganalysis feature extractor F, classifiers $\{C_{intra}, C_{inter}\}$, trade-off parameters $\{\lambda_1, \lambda_2, \lambda_3\}$, parameter vectors $\{\boldsymbol{\varphi}_p, \boldsymbol{\varphi}_s, \boldsymbol{\varphi}_d, \boldsymbol{\varphi}_a\}$;
Output: final steganalysis feature extractor F, predicted labels $\hat{y}_{intra}$ and $\hat{y}_{inter}$;

```
Initialize model parameters and extract feature vectors;
for k = 1 → N do;
    Calculate the structure loss L_stru to explore the intrinsic relationships of mismatched data by Eq. (11);
    Estimate the domain loss L_d based on DMMD to achieve the cluster consensus using Eq. (20);
    Obtain the prediction value ŷ_intra based on the intra-domain loss L_intra;
    Obtain the prediction value ŷ_inter based on the inter-domain loss L_inter;
    Compute the alignment loss L_align according to ŷ_intra and ŷ_inter by Eq. (21);
    while not converge do
        for each batch do
            Minimize the total loss L_total through Eq. (22);
        end for
        Update the parameters and steganalysis feature extractor F.
    end while
end for
```

prediction accuracy for cross-domain steganography detection. The most important part of our proposed CADM lies in the effectiveness of cluster partitioning and cluster matching, with performance relying on the SCFCM and CCM. Furthermore, the total loss in Eq. (22) is optimized with a gradient reversal layer (GRL) [35], which can reverse the gradient of the loss function when backpropagating to the steganalysis feature extractor F. These steps are repeated until the convergence conditions are satisfied, and thus the various parameters of CADM scheme are updated to overcome the distribution discrepancy between the training data and test data. The complete procedure of CADM is summarized in Algorithm 2.

3.5 Computational Complexity

The computational complexity of our proposed algorithm consists of the following three major parts:

1) Using the SCFCM clustering algorithm to explore the structural relationships, the time-consuming steps mainly include the PCA on the sample matrix and the solution of CLS, which takes $O((n_s^2 + n_t^2)d + n_s^2 n_t^2 m)$ time to obtain the fuzzy c-partition.
2) The cluster consensus matching is accomplished by measuring the distribution distance between the source and target clusters, whose time complexity is $O(\frac{1}{2}(P+Q)^2 T_1)$ to calculate the intra-domain loss and the inter-domain loss, while T_1 is the number of iterations of maintaining the consistency.
3) The objective function in Eq. (22) is implemented to perform the cycle-consistent optimization and adaptation, which takes $O((n_s + n_t)(T_2 + \log(n_s + n_t)))$ time to obtain the optimal result, and T_2 is the number of iterations that meet the convergence condition in Algorithm 2.

We assume $m, d, P, Q, T_1, T_2 \ll n_s + n_t$, the overall computational complexity of this algorithm can be simplified as $O(n_s^2 n_t^2 m + (n_s + n_t)\log(n_s + n_t))$. Notably, it can be seen from the results that CADM is different from those existing algorithms whose computational complexity will increase dramatically with the increase of feature dimensions. The computational complexity of the SCFCM clustering is related to the feature dimensions, while the complexity of the cluster consensus matching and the cycle-consistent distribution alignment is independent of the feature dimensions. Moreover, the last two parts are major components of overall computational complexity. As a result, the complexity of CADM is very little affected by the growth of the feature dimensions, which is also an advantage that distinguishes from other algorithms.

4 Experiments

4.1 Experimental Setup

1) Datasets: The images used in our experiments are taken from the BOSSbase 1.01 database [36] and the BOWS-2 database [37]. Each of the databases is composed of 10,000 pieces of portable gray map (PGM) images with a resolution of 512×512 pixels. In order to reduce computing burden, we resize the input images to the size of 256×256. When training the steganalysis networks, all images are separated into three non-overlapping groups, namely 50% for the training set, 10% for the validation set, and 40% for the testing set. We find the optimal hyperparameters according to the performance on the validation set, and then assess the effectiveness of the proposed CADM scheme on the testing set.

2) Steganographic Algorithms: The datasets are JPEG compressed with different quality factors (QF) of 75, 85, and 95 to obtain the cover images. We modify the DCT coefficients in the JPEG domain to embed secret information based on classical non-adaptive and adaptive steganographic methods, including nsF5 [1], J-UNIWARD [2], UERD [3], and J-MiPOD [6]. In the following text, we apply simplified symbols ns, J, U, and M to describe the nsF5, J-UNIWARD, UERD, and J-MiPOD steganographic algorithms, respectively. Since the performance of steganography detection is closely related to the amount of payloads, we use a series of relative payloads of 0.1, 0.2, and 0.3 bits per nonzero ac DCT coefficient (bpnzac) to explore their impact on cross-domain steganalysis.

3) Scenario Settings: We consider the following four scenario settings: (1) no adaptation (NA): it represents to train with the source domain and test using the target domain directly; (2) single source domain adaptation (SSDA): it learns meaningful representations from a single source domain to promote the performance of classification and detection in the target domain; (3) multiple source domain adaptation (MSDA): MSDA is an extension of SSDA, which captures transferable knowledge from multiple source domains for the recognition in the target domain; (4) sample-imbalanced domain adaptation (SIDA): it performs the match and adaptation operation on the training set, where the numbers of cover and stego samples have significant differences, to achieve cross-domain steganalysis.

4) Implementation Details: Our proposed framework and loss function are provided in Section 3. In the following, we

give the implementation details of CADM during training the cross-domain steganalysis model. We pre-train and fine-tune the SRNet on the BOSSbase and BOWS-2 datasets as the initial steganalysis feature extractor, which can capture the crucial cues caused by steganographic modifications. Due to the limitation of GPU memory and hardware resources, we utilize a batch size of 16 images per domain for all iterations. All layers are initialized by the Xavier initializer [38] to ensure the magnitude of the gradients approximately the same. In the training phase, we adopt Adam optimizer with a weight decay 2×10^{-4}. For the BOSSbase, the initial learning rate is set to 10^{-4} which is scaled by a factor of 10 at 6k iterations. For the experiments on the BOWS-2, the initial learning rate is set to 10^{-3}, and the learning rate is decayed by a factor of 0.1 at 1k iterations. According to the above settings, our proposed model is trained to minimize the total loss function mentioned in Section 3.4. After the cross-domain steganalysis model is well trained, the cross-domain detection results can be obtained by Algorithm 2.

5) Evaluation Metrics: The accuracy Acc and F_1-measure F_1 [14] of the steganography detection model on the testing set are employed as the evaluation criteria. The expression of Acc is as follows:

$$Acc = \frac{N_{correct}}{N_{testing}}, \tag{23}$$

where $N_{correct}$ is the number of samples correctly identified by the steganography detection model, and $N_{testing}$ denotes the total number of samples in the testing set.

The F_1-measure is defined as the harmonic mean of recall (Rec) and precision (Pre). Particularly, it is a widely adopted metric to estimate the performance of classifiers in imbalanced data, with higher scores indicating better performance. The relevant calculation equations are as follows:

$$Rec = \frac{TP}{TP + FN}, \tag{24}$$

$$Pre = \frac{TP}{TP + FP}, \tag{25}$$

$$F_1 = 2 * \frac{Pre * Rec}{Pre + Rec}, \tag{26}$$

where TP is the number of stego samples correctly identified by steganalysis model, FP denotes the number of cover samples incorrectly detected as stego samples, and FN represents the number of stego samples incorrectly identified as cover samples.

4.2 Comparison with Prior Arts

To evaluate the proposed scheme, our CADM is compared with the following recent state-of-the-art baselines on mismatched steganographic datasets. These methods can be divided into two categories depending on whether they are designed for matched or mismatched distribution conditions. 1) Effective steganalysis approaches for matched scenarios: JPEG domain rich model (JRM) [9], discrete cosine transform residual (DCTR) [10], and steganalysis residual network (SRNet) [11]. 2) Distribution adjustment and adaptation schemes for mismatched scenarios: manifold embedded distribution alignment (MEDA) [39], guide subspace learning (GSL) [40], iterative multi-order feature alignment (IMFA) [13], transferable heterogeneous feature subspace learning (THFSL) [12], adaptive cost-sensitive feature learning (ACFL) [14], and multiperspective progressive structure adaptation (MPSA) [19].

4.3 Experimental Results and Analysis

We compare our scheme with previous state-of-the-arts methods in four cases, i.e., NA, SSDA, MSDA, and SIDA. For NA cases, we select the classical steganalysis approach as a baseline to explore the impact of mismatched data on its performance. For SSDA scenarios, we consider the mismatched steganographic algorithm (MSA) and mismatched quality factor (MQF) respectively. In addition, we also conduct cross-dataset experiments to evaluate the performance in the single source domain case. For MSDA settings, we perform a set of experiments to understand how the number of source domains influences the detection performance. For SIDA situations, we first conduct baseline experiments on imbalanced data distribution in non-cross-domain conditions, and then further perform experiments on imbalanced data distribution in cross-domain conditions.

1) Results on Mismatched Steganographic Algorithm (MSA): The following experimental results reveal that the proposed CADM for the MSA problem is reasonable and feasible. The source and target samples have the same quality factor and payload rate, but the steganographic methods are different in the training and testing sets. Specifically, the BOSSbase dataset is applied to produce cover samples using QF = 75 or 95, and subsequently nsF5, UERD or J-MiPOD steganography techniques are employed to generate stego samples with a payload rate of 0.3 bpnzac. The detailed experimental results are given in Table 2. Comparing no adaptation methods and domain adaptation approaches, it can be seen that on average the domain adaptation approaches achieve better performance, which is obviously different from the no adaptation situation. Notably, for the baseline method without using adaptation strategy (i.e., JRM-NA), our proposed CADM improves the average accuracy from 63.2% to 77.0%. This is because the adaptation operations can narrow the distribution gap to some extent. Moreover, CADM can fully explore the combination of diverse clusters and perfectly perform the cluster matching through the consensus clustering to achieve better cross-domain detection performance.

TABLE 2: Comparison of Detection Accuracy on the Mismatched Steganographic Algorithm Setting

Source	Target	JRM-NA	MEDA	GSL	IMFA	THFSL	MPSA	Ours
75-U	75-ns	0.787	0.794	0.812	0.811	0.841	0.904	**0.929**
75-ns	75-U	0.679	0.692	0.721	0.709	0.718	0.789	**0.814**
75-U	75-M	0.520	0.541	0.574	0.561	0.583	0.641	**0.659**
75-M	75-U	0.593	0.602	0.638	0.652	0.664	0.725	**0.764**
95-U	95-ns	0.822	0.834	0.865	0.843	0.876	**0.940**	0.938
95-ns	95-U	0.627	0.648	0.667	0.666	0.664	0.739	**0.769**
95-U	95-M	0.511	0.529	0.534	0.519	0.512	0.538	**0.563**
95-M	95-U	0.616	0.632	0.651	0.664	0.693	0.762	**0.785**
75-ns	75-M	0.524	0.528	0.540	0.538	0.551	0.619	**0.641**
75-M	75-ns	0.685	0.703	0.729	0.737	0.762	0.843	**0.872**
95-ns	95-M	0.518	0.523	0.526	0.521	0.536	0.541	**0.567**
95-M	95-ns	0.702	0.728	0.769	0.798	0.831	0.935	**0.943**
Average		0.632	0.646	0.669	0.668	0.686	0.748	**0.770**

2) Results on Mismatched Quality Factor (MQF): The performance of the proposed scheme for the MQF problem is investigated by the following experiments. We select cover samples with the QF = {75, 85, 95} from the BOSSbase dataset and then the J-UNIWARD or UERD steganography methods are used to generate stego samples with a payload

rate of 0.2 bpnzac. The detection results are reported in Table 3, from which we observe that MPSA ranks second (70.7%) on average, but our proposed CADM is still 2.4% higher than MPSA and thus achieves state-of-the-art performance. Concretely, except for a few special cases, we can find that our CADM scheme outperforms the other methods on all the MQF tasks, and the accuracy of our method is 11.8% higher than that of JRM-NA. The reason may be that the exploration of the latent transferable knowledge through the intra-domain and inter-domain cluster consensus in this paper can automatically match inconsistent distributions in cross-domain steganalysis scenarios.

TABLE 3: Comparison of Detection Accuracy on the Mismatched Quality Factor Setting

Source	Target	JRM-NA	MEDA	GSL	IMFA	THFSL	MPSA	Ours
75-U	85-U	0.646	0.675	0.686	0.674	0.623	0.745	**0.792**
85-U	75-U	0.654	0.668	0.633	0.671	0.673	0.748	**0.780**
75-U	95-U	0.579	0.614	0.622	0.603	0.539	0.671	**0.694**
95-U	75-U	0.662	0.690	0.672	0.688	0.657	0.751	**0.793**
75-J	85-J	0.613	0.652	0.655	0.643	0.627	**0.718**	0.695
85-J	75-J	0.605	0.647	0.598	0.640	0.621	0.722	**0.779**
75-J	95-J	0.544	0.583	0.564	0.645	0.531	0.639	**0.667**
95-J	75-J	0.623	0.639	0.631	0.633	0.693	0.682	**0.724**
85-U	95-U	0.620	0.652	0.663	0.655	0.562	0.708	**0.731**
95-U	85-U	0.601	0.648	0.659	0.632	0.606	**0.729**	0.703
85-J	95-J	0.588	0.621	0.628	0.609	0.548	0.682	**0.696**
95-J	85-J	0.565	0.609	0.615	0.598	0.591	0.693	**0.723**
Average		0.608	0.642	0.636	0.641	0.606	0.707	**0.731**

3) Results on Cross-Dataset Experiments with BOSSbase and BOWS-2: We adopt two different settings to evaluate the effectiveness of the proposed CADM in cross-dataset scenarios with BOSSbase and BOWS-2. In the first setting, the samples with a certain QF = $\{75, 95\}$ in BOSSbase are used as the source set, while the samples with the same QF in BOWS-2 are employed as the target set. In the second setting, the source and target sets are swapped. The stego samples are obtained by non-adaptive and adaptive steganography methods with a payload rate of 0.3 bpnzac. The experimental results are illustrated in Fig. 5, from which we can see that the proposed CADM outperform the state-of-the-art methods in all cases. It is obvious that the detection results are significantly better under 75-ns condition than others, while the detection results are very poor under 95-M condition. The reason is that nsF5 is a conventional non-adaptive steganography algorithm with low security, which makes it relatively easy to detect. However, J-MiPOD is a modern adaptive steganography algorithm with strong concealment, which makes it extremely difficult to detect.

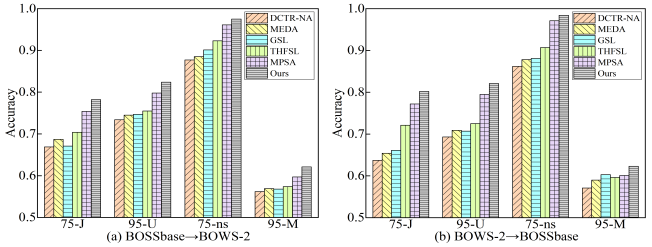


Fig. 5: Comparison of different methods on cross-dataset settings with BOSSbase and BOWS-2.

4) Evaluation Study of Multiple Source Domain Adaptation: To further demonstrate the effectiveness of our proposed CADM, we conduct a series of experiments to reveal the relationship between the number of source domains and detection performance. The experimental data are collected from six different domains with the payload rate of 0.3 bpnzac, including 75-ns, 75-J, 75-U, 95-ns, 95-J and 95-U. In order to guarantee that the single source domain methods can be performed under the condition of multiple source domain scenarios, we adopt the domain selection strategy as a preprocessing step according to reference [41]. In this way, we record them as M-MEDA, M-GSL and M-THFSL in the subsequent experiments. From Fig. 6, it shows the best detection results with various target domains that can be achieved when the number of latent source domains is set as two or three in almost all cases. The reasons can be attributed to the following two aspects. On the one hand, an appropriate increase in the number of latent source domains can boost the adaptation by exploring complementary information among different domains. On the other hand, an excessive number of latent source domains may introduce interference information and cause negative transfer, which will lead to the degradation of performance in cross-domain steganalysis. Table 4 shows the results of different target domains on their best performance. From the results, we can also see that the proposed CADM is superior to several other state-of-the-art methods in each case.

TABLE 4: Cross-domain Steganalysis Accuracy (%±STD) with Different Methods on Four Target Datasets

Compared Methods	75-ns	75-U	95-U	95-J
DCTR-NA	83.9±2.9	70.7±4.6	67.1±2.5	64.8±2.7
M-MEDA	84.6±2.0	73.3±2.8	67.9±0.9	67.0±1.6
M-GSL	87.8±1.2	74.8±1.9	66.4±1.7	67.3±2.3
M-THFSL	83.4±2.5	73.1±3.2	70.3±1.4	68.5±3.8
MPSA	93.2±1.7	78.9±2.4	77.5±2.1	75.5±1.9
Ours	**95.7±1.4**	**80.5±2.1**	**79.6±0.8**	**77.2±1.1**

5) Detection Performance on Imbalanced Data Distribution in Non-Cross-Domain Conditions: To evaluate the performance of our proposed scheme and other methods under imbalanced data distribution, we conduct a series of experiments with varying imbalanced ratio using the BOSSbase dataset on non-cross-domain scenarios. Since stego samples are generally less than cover samples in practical applications, the number of stego samples is set as 500 while the number of cover samples is selected from the set {500,1000,...,5000}. Accordingly, we can obtain ten training datasets with different degrees of imbalanced ratios. Moreover, the testing samples are randomly collected from the remaining datasets. Due to the fact that F_1-measure is a more appropriate performance evaluation criteria than accuracy in the imbalanced distribution. Therefore, Fig. 7 shows the detection performance using F_1-measure. The experimental results indicate that: (1) the proposed CADM is significantly superior to other counterpart methods in the F_1-measure; and (2) the performance of the method (i.e., ACFL and CADM) specifically designed for addressing the imbalanced data distribution is relatively stable in the face of various imbalanced ratios, while the performance of other approaches decreases dramatically with the increase of the proportion of cover and stego samples. The reason is that the limited information and uneven distribution of the minority class makes it difficult to detect, which may lead to misclassification and degradation of model performance. However, our proposed CADM can automatically match the inconsistent distribution to alleviate the influence of the imbalanced distribution.

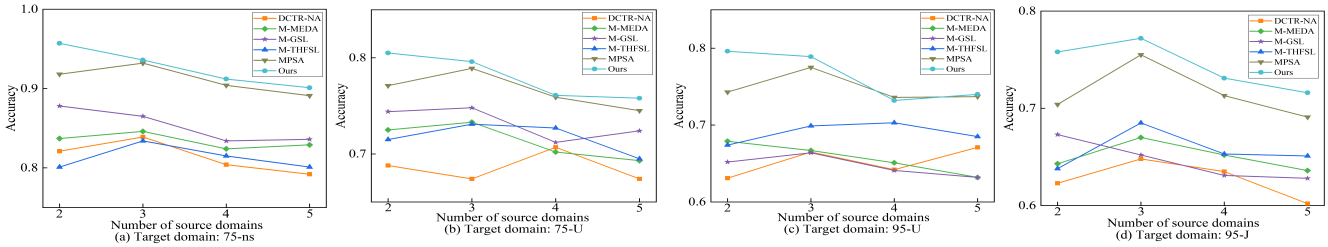


Fig. 6: Accuracy of cross-domain steganalysis with different number of source domains under the condition of MSDA.

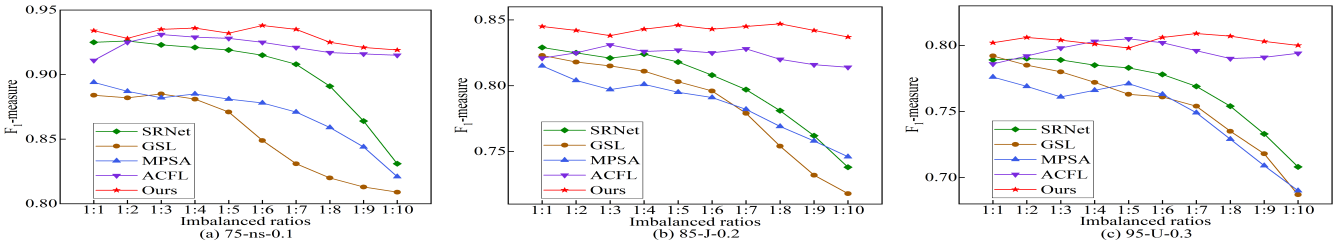


Fig. 7: Performance of steganography detection with different imbalanced ratios under non-cross-domain conditions. (a) 75-ns-0.1: using nsF5 with QF = 75 and embedding rate of 0.1 bpnzac. (b) 85-J-0.2: using J-UNIWARD with QF = 85 and embedding rate of 0.2 bpnzac. (c) 95-U-0.3: using UERD with QF = 95 and embedding rate of 0.3 bpnzac.

6) Detection Performance on Imbalanced Data Distribution in Cross-Domain Conditions: In order to further verify the ability of the proposed CADM to handle the mismatched distributions, we perform more complex experiments on imbalanced data distribution in cross-domain conditions. We adopt mismatched steganographic algorithm and quality factor to construct the source domain and target domain, where the imbalanced ratio of the stego sample and the cover sample is set to 1:10 in the source training dataset. Table 5 shows the F_1-measure results on the imbalanced training data in cross-domain steganography detection scenarios. It can be found that our scheme is obviously better than other comparison methods. Although in scenarios where there are only inconsistent or imbalanced distributions (i.e., domain inconsistency or data imbalance issues), some comparison methods have achieved promising detection results. However, in more complex scenarios where both problems exist at the same time, the performance of comparison methods is severely degraded. The reason may be that these methods have limited ability to deal with the distribution shifts and cannot automatically perform distribution matching. In addition, our proposed CADM can gain about 3.2% average F_1-measure improvement compared with the second best method, which demonstrates that mining and aligning the suitable components based on consensus clustering can achieve superior performance in mismatched steganalysis tasks.

TABLE 5: Results of Steganalysis Using F_1-measure on Imbalanced Data Distribution under Cross-Domain Conditions

Source	Target	SRNet-NA	GSL	IMFA	THFSL	ACFL	MPSA	Ours
75-U	85-ns	0.632	0.651	0.634	0.659	0.672	0.703	**0.746**
85-ns	75-U	0.535	0.537	0.554	0.562	0.583	0.634	**0.673**
75-U	85-J	0.518	0.526	0.529	0.548	0.560	0.591	**0.629**
85-J	75-U	0.506	0.502	0.517	0.531	0.529	0.574	**0.603**
85-U	75-ns	0.613	0.658	0.691	0.684	0.687	0.738	**0.752**
75-ns	85-U	0.501	0.531	0.528	0.526	0.524	0.592	**0.625**
85-U	75-J	0.539	0.554	0.563	0.543	0.562	0.587	**0.604**
75-J	85-U	0.516	0.525	0.529	0.547	0.563	0.621	**0.653**
75-ns	85-J	0.531	0.523	0.540	0.562	0.564	0.595	**0.637**
85-J	75-ns	0.539	0.561	0.584	0.596	0.581	0.629	**0.662**
85-ns	75-J	0.527	0.549	0.536	0.528	0.542	0.573	**0.611**
75-J	85-ns	0.624	0.678	0.664	0.687	0.697	0.728	**0.749**
Average		0.548	0.566	0.572	0.581	0.589	0.630	**0.662**

5 Analysis and Discussion

5.1 Ablation Study

We train the model on three representative tasks (i.e., Task1: 85-J→75-M, Task2: 75-U→95-ns, and Task3: 75-U&85-M&95-ns→75-J) with the payload rate of 0.2 bpnzac from the BOWS-2 dataset to perform ablation analysis, where we investigate how the influence of different components on the performance of the proposed scheme from various perspectives.

Effect of Domain Consistency Score. To better investigate domain consistency score, we carry out a series of experiments to understand its mechanism. We consider the consistency scores of the source domain, the target domain, and the entire domain, which are denoted by the symbols DCS_s, DCS_t, and DCS, respectively. As shown in Fig. 8(a), as the number of clusters c increases, the curves of DCS_s and DCS_t show the same trends, that is, DCS_s and DCS_t both decrease. However, the consistency score of the entire domain, which consists of the source domain and the target domain, is gradually increasing. The reason can be explained in the following two aspects. On the one hand, with the increase of c, the source and target samples are scattered to generate more and smaller clusters, which leads to a decrease in DCS_s and DCS_t. On the other hand, as more fine-grained clusters are formed, the distribution matching can be better implemented between the source and target clusters, which explains the increase of DCS.

Effect of Cluster Consensus Matching. In order to reveal the characteristics of cluster consensus matching in the training process, we record the change of the domain consistency score as the number of iterations increases. In Fig. 8(b), we visualize the evolution of the domain consistency score as the training progresses. As expected, the domain consistency score reaches a steady state after the initial several iterations, which implies that our scheme can quickly find the optimal number of clusters. In addition, this indicates that a search for the number of clusters is only required in the early phase.

Effect of SCFCM Clustering. Fig. 8(c) shows the evolution of the cluster number c during training under three

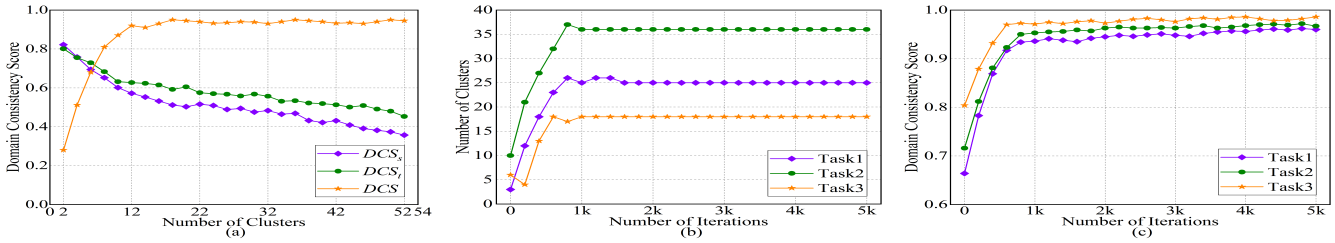


Fig. 8: Impact assessment of our proposed scheme. (a) The variation curve of domain consistency score with the number of clusters c on 85-J→75-M (0.2 bpnzac). (b) The evolution of the number of clusters c as training progresses. (c) The evolution of domain consistency score as training progresses.

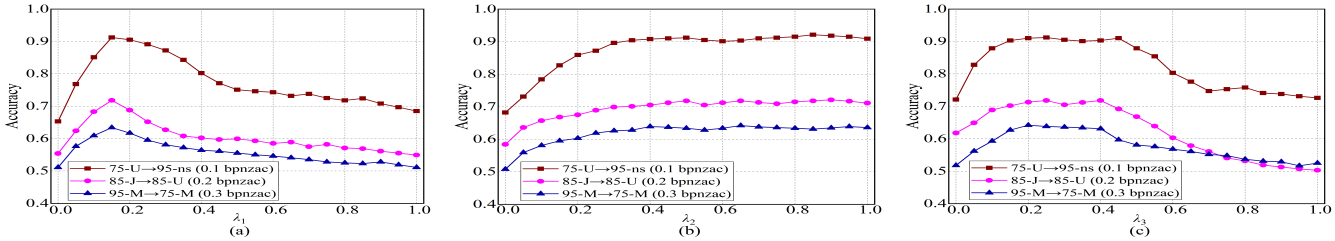


Fig. 9: Influence of parameters λ_1, λ_2 and λ_3 on the performance of CADM, respectively.

tasks. In these experiments, we do not implement the proposed stop rule (see Section 3.2). As illustrated in the figure, the number of clusters gradually converges to the stationary and optimal value after several rounds of searches, which is similar to the convergence trend of the domain consistency score in Fig. 8(b). This also means that the search for c is only necessary in the initial stage of training, which proves that the proposed stop rule used to shorten the training time is reasonable.

Effect of Different Modules. We further conduct a series of experiments to evaluate the contribution of each module in our proposed scheme. The effectiveness is investigated by changing one module while fixing the others. In our experiments, the influence of different modules can be estimated by three major constraint terms, including the structure loss (SL) term, domain loss (DL) term, and alignment loss (AL) term. The experiment results are reported in Table 6 by using the proposed model without (w/o) the SL term, without (w/o) the DL term and without (w/o) the AL term. From the results, we observe that the detection performance can be boosted by using the constraint terms with SL, DL and AL, which also validates the feasibility of CADM in cross-domain steganalysis.

TABLE 6: Influence of Different Modules on Three Representative Tasks by Deleting One Module While Fixing the Others

Detection Tasks	CADM	w/o SL	w/o DL	w/o AL
Task1	**0.679**	0.564	0.590	0.634
Task2	**0.886**	0.792	0.801	0.822
Task3	**0.743**	0.653	0.676	0.698
Average	**0.769**	0.670	0.689	0.718

5.2 Parameter Sensitivity

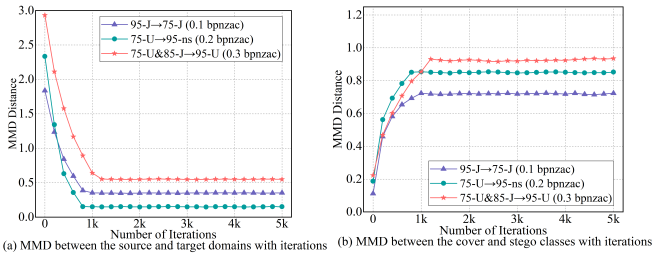
We conduct comprehensive experiments to observe the effect of parameters λ_1, λ_2, and λ_3 on the performance of our scheme. The results on 75-U→95-ns (0.1 bpnzac), 85-J→85-U (0.2 bpnzac), and 95-M→75-M (0.3 bpnzac) tasks are presented in Fig. 9. The impact of parameters on the performance of CADM is analyzed by utilizing the control variable method. Fig. 9(a) shows the average accuracy using different values of λ_1 on three representative tasks, where λ_2 = 0.5 and λ_3 = 0.35. It can be seen from the experimental results that with the increase of λ_1, the performance of the scheme enhances in the initial phase and then decreases progressively. To be specific, the best performance is achieved when λ_1 is set to 0.15. Fig. 9(b) provides the average accuracy with different values of λ_2 under the condition of λ_1 = 0.15 and λ_3 = 0.35. From the results, when λ_2 is greater than 0.3, the performance almost remains at a desired level. Fig. 9(c) gives the average accuracy using different values of λ_3 with the setting of λ_1 = 0.15 and λ_2 = 0.5. Within a wide range of $\lambda_3 \in [0.2, 0.4]$, the performance of CADM achieves the best and only varies in a narrow range, which indicates that our scheme is robust to the selection of λ_3 in this interval.

5.3 Convergence Evaluation

The convergence of the CADM scheme is evaluated by performing extensive experiments on BOSSbase (i.e., 95-J→75-J) and BOWS-2 (i.e., 75-U→95-ns and 75-U&85-J→95-U). Fig. 10(a) describes the reduction of MMD distance between the source and target domain during the training process, which demonstrates that the discrepancy across different domains can be effectively eliminated by our scheme. Fig. 10(b) presents the MMD distance between the cover and stego classes. We find that the value of MMD increases monotonically with cycle-consistent iterations. In other words, our scheme can improve the discrimination of representations in cross-domain steganalysis. Moreover, the response curve of MMD distance can achieve the steady state after a certain number of iterations, which proves that the proposed CADM can converge.

5.4 Time Complexity

We have recorded the average execution time of all the comparison methods under the distribution shift conditions. The experiments of computational time are conducted on the Windows 10 operating system, Intel(R) Core(TM) i7-8700 CPU @3.20 GHz, 500G Solid State Drives, 16 GB DDR3



(a) MMD between the source and target domains with iterations (b) MMD between the cover and stego classes with iterations

Fig. 10: Convergence evaluation of the proposed CADM scheme. During the iterative optimization and adaptation, the MMD between the source and target domains can be minimized, while the MMD between the cover and stego classes can be maximized.

TABLE 7: Execution Time of Comparison Methods on the 75-U → 85-ns Dataset under the Distribution Shift Conditions

Method	SRNet-NA	GSL	IMFA	THFSL	ACFL	MPSA	Ours
Runtime (s)	749.2	1783.8	1967.1	1479.3	1186.8	1251.0	1036.4

RAM, MATLAB R2020a and Nvidia Titan X Pascal GPU. The experimental results on the 75-U → 85-ns mismatched datasets with the imbalanced ratio of 1:10 and the embedding rate of 0.3 bpnzac are reported in Table 7. From the experimental results, it can be seen that the calculation time of our proposed CADM is less than or roughly equal to the other compared methods, except for the no adaptation method (i.e., SRNet-NA). Therefore, it can address the problem of cross-domain steganalysis within a reasonable range of running time. In addition, to further reduce time complexity, we can deploy and execute a parallel computing architecture for our solution on advanced software and hardware resources.

6 Conclusion

In this paper, we present an automatic distribution matching scheme based on consensus clustering to realize the recognition of cross-domain steganographic modifications. The main conclusions can be drawn from this research work as follows: 1) an effective SCFCM clustering can fully exploit both the correlation and complementarity from the original mismatched data, which provides a critical guidance to capture intrinsic structural relationships in steganography detection across domains; 2) the cluster consensus matching is designed from the perspective of intra-domain and inter-domain to guarantee that the distribution gaps are adaptively filled to promote the quality of steganalysis features, which can be extended to enhance the signal-to-noise ratio for researchers in other related fields, such as image forgery detection and localization; 3) this work offers a new perspective that the cycle-consistent optimization and adaptation can be leveraged to further boost the overall performance via encouraging a collaboration between the source and target clusters in cross-domain steganalysis; and 4) comprehensive experiments show that our proposed CADM is more applicable for steganography detection scenarios where the distributions are not aligned, and can automatically perform the distribution matching through consensus clustering to mitigate the risk of negative transfer. In the near future, we are planning to study the cluster consensus and cooperation mechanism that can accurately and rapidly explore structure representations to achieve the fine-grained matching for cross-domain steganalysis tasks.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (No. 2020YFB1805400); in part by the National Natural Science Foundation of China (Nos. 61876134, U1836112, and U1536204).

References

[1] J. J. Fridrich, T. Pevný, and J. Kodovský, "Statistically undetectable jpeg steganography: dead ends challenges, and opportunities," in *Proceedings of the 9th workshop on Multimedia & Security (MM&Sec)*, Sep. 2007, pp. 3–14.

[2] V. Holub, J. J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," *EURASIP Journal Information Security*, vol. 2014, 2014, Art. no. 1.

[3] L. Guo, J. Ni, W. Su, C. Tang, and Y. Shi, "Using statistical image model for JPEG steganography: Uniform embedding revisited," *IEEE Transactions Information Forensics and Security*, vol. 10, no. 12, pp. 2669–2680, 2015.

[4] I. J. Kadhim, P. Premaratne, P. J. Vial, and B. Halloran, "Comprehensive survey of image steganography: Techniques, evaluations, and trends in future research," *Neurocomputing*, vol. 335, pp. 299–326, 2019.

[5] J. Jia, Z. Xiang, L. Wang, and Y. Xu, "An adaptive JPEG double compression steganographic scheme based on irregular DCT coefficients distribution," *IEEE Access*, vol. 7, pp. 119 506–119 518, 2019.

[6] R. Cogranne, Q. Giboulot, and P. Bas, "Steganography by minimizing statistical detectability: The cases of JPEG and color images," in *IH&MMSec '20: ACM Workshop on Information Hiding and Multimedia Security*, 2020, pp. 161–167.

[7] T. Wu, W. Ren, D. Li, L. Wang, and J. Jia, "Jpeg steganalysis based on denoising network and attention module," *International Journal of Intelligent Systems*, to be published, doi: 10.1002/int.22749.

[8] W. Ren, L. Zhai, J. Jia, L. Wang, and L. Zhang, "Learning selection channels for image steganalysis in spatial domain," *Neurocomputing*, vol. 401, pp. 78–90, 2020.

[9] J. J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Transactions Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[10] V. Holub and J. J. Fridrich, "Low-complexity features for JPEG steganalysis using undecimated DCT," *IEEE Transactions Information Forensics and Security*, vol. 10, no. 2, pp. 219–228, 2015.

[11] M. Boroumand, M. Chen, and J. J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Transactions Information Forensics and Security*, vol. 14, no. 5, pp. 1181–1193, 2019.

[12] J. Jia, L. Zhai, W. Ren, L. Wang, Y. Ren, and L. Zhang, "Transferable heterogeneous feature subspace learning for JPEG mismatched steganalysis," *Pattern Recognition*, vol. 100, 2020, Art. no. 107105.

[13] X. Kong, C. Feng, M. Li, and Y. Guo, "Iterative multi-order feature alignment for JPEG mismatched steganalysis," *Neurocomputing*, vol. 214, pp. 458–470, 2016.

[14] J. Jia, L. Zhai, W. Ren, L. Wang, and Y. Ren, "An effective imbalanced JPEG steganalysis scheme based on adaptive cost-sensitive feature learning," *IEEE Transactions on Knowledge and Data Engineering*, to be published, doi: 10.1109/TKDE.2020.2995070.

[15] L. Yang, M. Men, Y. Xue, J. Wen, and P. Zhong, "Transfer subspace learning based on structure preservation for JPEG image mismatched steganalysis," *Signal Processing: Image Communication*, vol. 90, 2021, Art. no. 116052.

[16] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[17] S. Zhao, B. Li, X. Yue, Y. Gu, P. Xu, R. Hu, H. Chai, and K. Keutzer, "Multi-source domain adaptation for semantic segmentation," in *Annual Conference on Neural Information Processing Systems*, 2019, pp. 7285–7298.

[18] F. Zhu, Y. Wang, J. Zhou, C. Chen, L. Li, and G. Liu, "A unified framework for cross-domain and cross-system recommendations," *IEEE Transactions on Knowledge and Data Engineering*, to be published, doi: 10.1109/TKDE.2021.3104873.

[19] J. Jia, M. Luo, J. Liu, W. Ren, and L. Wang, "Multiperspective progressive structure adaptation for jpeg steganography detection across domains," *IEEE Transactions on Neural Networks and Learning Systems*, to be published, doi: 10.1109/TNNLS.2021.3054045.

[20] R. Xia, Z. Pan, and F. Xu, "Instance weighting with applications to cross-domain text classification via trading off sample selection bias and variance," in *International Joint Conference on Artificial Intelligence*, 2018, pp. 4489–4495.

[21] C. Ren, P. Ge, P. Yang, and S. Yan, "Learning target domain specific classifier for partial domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, to be published, doi: 10.1109/TNNLS.2020.2995648.

[22] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2066–2073.

[23] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *International Conference on Machine Learning*, vol. 37, 2015, pp. 97–105.

[24] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Annual Conference on Neural Information Processing Systems*, 2016, pp. 136–144.

[25] ——, "Deep transfer learning with joint adaptation networks," in *International Conference on Machine Learning*, vol. 70, 2017, pp. 2208–2217.

[26] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *European Conference on Computer Vision*, vol. 9908, 2016, pp. 597–613.

[27] R. Xu, Z. Chen, W. Zuo, J. Yan, and L. Lin, "Deep cocktail network: Multi-source unsupervised domain adaptation with category shift," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3964–3973.

[28] S. Li, C. H. Liu, Q. Lin, Q. Wen, L. Su, G. Huang, and Z. Ding, "Deep residual correction network for partial domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2329–2344, 2020.

[29] H. Wu, Y. Yan, G. Lin, M. Yang, M. K.-P. Ng, and Q. Wu, "Iterative refinement for multi-source visual domain adaptation," *IEEE Transactions on Knowledge and Data Engineering*, to be published, doi: 10.1109/TKDE.2020.3014697.

[30] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *International Conference on Computer Vision*, 2019, pp. 1406–1415.

[31] S. Zhao, G. Wang, S. Zhang, Y. Gu, Y. Li, Z. Song, P. Xu, R. Hu, H. Chai, and K. Keutzer, "Multi-source distilling domain adaptation," in *AAAI Conference on Artificial Intelligence*, 2020, pp. 12 975–12 983.

[32] Y. Guo and A. Sengür, "NCM: neutrosophic c-means clustering algorithm," *Pattern Recognition*, vol. 48, no. 8, pp. 2710–2724, 2015.

[33] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," in *International Conference on Learning Representations*, 2017.

[34] C. Lu, J. Tang, S. Yan, and Z. Lin, "Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 829–839, 2016.

[35] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*, vol. 37, 2015, pp. 1180–1189.

[36] T. Filler, T. Penvy, and P. Bas, "Boss (break our stegnography system)," http://www.agents.cz/boss/.

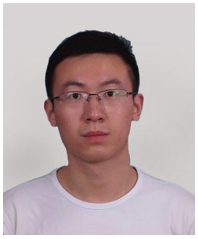
[37] P. Bas and T. Furon, "Image database of bows-2," http://bows2.ec-lille.fr/.

[38] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, vol. 9, 2010, pp. 249–256.

[39] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu, "Visual domain adaptation with manifold embedded distribution alignment," in *ACM Multimedia Conference on Multimedia Conference*, 2018, pp. 402–410.

[40] L. Zhang, J. Fu, S. Wang, D. Zhang, Z. Dong, and C. L. Philip Chen, "Guide subspace learning for unsupervised domain adaptation," *IEEE Transactions Neural Networks and Learning Systems*, to be published, doi: 10.1109/TNNLS.2019.2944455.

[41] H. S. Bhatt, A. Rajkumar, and S. Roy, "Multi-source iterative adaptation for cross-domain classification," in *International Joint Conference on Artificial Intelligence*, 2016, pp. 3691–3697.



Ju Jia received his Ph.D. degree in cyberspace security from Wuhan University, Wuhan, China, in 2021. He is currently a research fellow with the School of Computer Science and Engineering at Nanyang Technological University, Singapore. His research interests include: information hiding, multimedia data security, transfer learning, and domain adaptation.

Meng Luo received her B.Eng. degree in information security from Wuhan University, Wuhan, China, in 2015 and the Ph.D. degree in computer science from Stony Brook University, USA, in 2020. She is currently a postdoctoral research associate with Khoury College of Computer Sciences at Northeastern University, USA. Her research interests include mobile security, web security, and artificial intelligence in cybersecurity.

Siqi Ma received the B.S. degree in computer science from Xidian University, Xi'an, China in 2013 and Ph.D. degree in information system from Singapore Management University in 2018, respectively. She was a research fellow of distinguished system security group from CSIRO and then was a lecturer at University of Queensland. She is currently a senior lecturer of the University of New South Wales, Canberra Campus, Australia. Her research interests include data security, IoT security and software security.

Lina Wang received the B.S. degree from the Hefei University of Technology, Hefei, China, in 1986, and the M.S. and Ph.D. degrees from Northeastern University, Shenyang, China, in 1989 and 2001, respectively. She is currently a Professor in Cyber Science and Engineering School of Wuhan University. Her research interests include multimedia content security, data security, and machine learning methods in network security detection.

Yang Liu (Senior Member, IEEE) received the B.Comp. degree (Hons.) from the National University of Singapore (NUS) in 2005 and the Ph.D. degree from NUS and MIT, in 2010. He started his postdoctoral work in NUS and MIT. In 2012, he joined Nanyang Technological University (NTU). He is currently a Full Professor and the Director of the Cybersecurity Laboratory, NTU. He specializes in software verification, security, and software engineering. His research has bridged the gap between the theory and practical usage of formal methods and program analysis to evaluate the design and implementation of software for high assurance and security. By now, he has more than 270 publications in top tier conferences and journals. He received a number of prestigious awards, including the MSRA Fellowship, the TRF Fellowship, the Nanyang Assistant Professor, the Tan Chin Tuan Fellowship, the Nanyang Research Award, and eight best paper awards in top conferences, such as ASE, FSE, and ICSE.