Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33194

A Review of Feature Alignment Methods Based on Adversarial Learning and Discrepancy Maximization in Visual Unsupervised Domain Adaptation

Haiming Zheng

School of Computer Science and Technology, Tianjin University, Tianjin, China
zhenghm@tju.edu.cn

Abstract. Unsupervised domain adaptation addresses domain shift in visual recognition, enabling models trained on labeled source domains to transfer to unlabeled target domains. Among various approaches, feature alignment is a mainstream method, aiming to reduce feature-space distribution discrepancy and learn domain-invariant features. Thus, this paper provides a systematic review of feature alignment methods in visual unsupervised domain adaptation. Based on recent research in computer vision and machine learning, existing methods can be divided into two main paradigms, with one based on adversarial training and the other on discrepancy maximization. For each paradigm, this review elaborates on core principles, representative works, evolutionary trajectories, applicable scenarios, and limitations. Research shows that the two paradigms differ in mechanism, with adversarial training focusing on global distribution alignment and discrepancy maximization on local decision boundary optimization. Current research still encounters key challenges, including negative transfer, unstable training, and weak theoretical guarantees. Future directions may involve integrating self-supervised learning, Transformer architectures, and vision-language foundation models to build more robust and efficient domain adaptation models.

Keywords: Unsupervised Domain Adaptation, Feature Alignment, Adversarial Learning, Maximum Classifier Discrepancy

1. Introduction

Due to the distribution gap, or "domain shift," unsupervised domain adaptation is challenging. This issue prevents models trained on the labeled source domain from generalizing well to the unlabeled target domain. To address this, unsupervised domain adaptation leverages labeled source data and unlabeled target data to learn a model that performs well on the target domain. Among the various approaches, feature alignment has become the most mainstream technical path, aiming to bring the feature distributions of the source and target domains closer in the feature space, forcing the feature extractor to learn discriminative features that are invariant to domain attributes and relevant only to semantic categories. Based on alignment objectives, feature alignment methods can be divided into two categories, with adversarial training employing a domain discriminator and feature extractor to align overall feature distributions, while discrepancy-based methods leverage multiple classifiers'

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33194

output differences to achieve fine-grained decision boundary adjustments. This review summarizes discriminative feature alignment methods, distinguishing between distribution alignment through adversarial training and boundary alignment through discrepancy maximization, and discusses their core principles, representative works, and developmental trajectories. Besides, it addresses major challenges, including training stability, negative transfer, and theoretical guarantees, and discusses potential future directions that integrate self-supervised learning, Transformers, and vision-language foundation models, providing useful references for research in this field.

2. Principles and methods of discriminative feature alignment

2.1. Representation shift and decision boundary shift

In unsupervised domain adaptation, the distribution discrepancy in feature space is reflected both in the overall shift of features and in local deviations near the decision boundary. Both types of shift jointly affect model performance on the target domain, with representation shift referring to the phenomenon where, after being mapped by the feature extractor, data from the source and target domains form different distributions in the feature space, as shown in Figure 1a [1]. And this shift typically arises from differences in low-level statistics between domains, such as illumination, color, or texture, which adversarial training methods address by aligning the overall feature distribution via a domain discriminator competing with a feature extractor, thereby extracting domain-invariant features [2]. In contrast, decision boundary shift is more subtle. As shown in Figure 1b, even when overall distributions align, boundary features can remain uncertain because adversarial training only matches the global distribution and ignores class information. Ben-David et al. proposed that the target domain error $\epsilon_T(h)$ can be bounded by the source domain error $\epsilon_S(h)$ plus the $\mathcal{H}\Delta\mathcal{H}$ distance, reflecting the prediction discrepancy of classifiers on the target domain and directly corresponds to the location of the decision boundary [3]. Based on this insight, discrepancy maximization methods can mitigate shifts near the decision boundary [1].

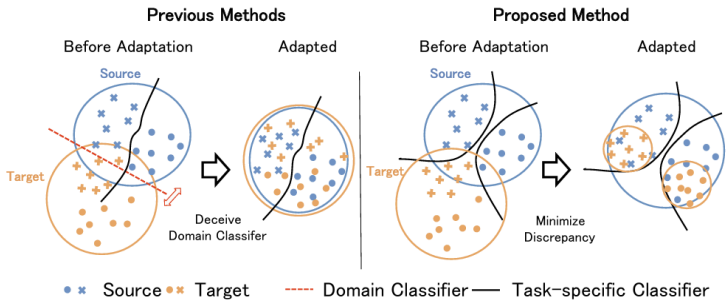


Figure 1. Representation shift and decision boundary shift. (a) global distribution alignment (ignoring decision boundary); (b) detection of features near the boundary [1]

2.2. Distribution alignment and boundary alignment strategies

Based on the analysis of representation shift and decision boundary shift, existing feature alignment methods can be broadly categorized into two classes, each targeting different types of distribution discrepancies. In practice, they are complementary, forming a global-to-local adaptation strategy. Specifically, distribution alignment methods primarily address overall feature shift by reducing the global distribution discrepancy between source and target domains through adversarial mechanisms. For instance, representative methods like DANN, ADDA, and CDAN leverage an adversarial game

DOI: 10.54254/2755-2721/2026.GU33194

between a domain discriminator and a feature extractor to learn domain-invariant features, thereby mitigating global representation shift [2,4,5]. In contrast, boundary alignment targets local shifts by minimizing prediction gaps among classifiers on the target domain. Moreover, Maximum Classifier Discrepancy (MCD) aligns features near the decision boundary by iteratively reducing classifier prediction gaps [1]. Relatedly, methods like MMCD_DA, SLOT, and MDD extend this idea [6-8]. Increasingly, studies combine both paradigms to align global distributions and refine features near the decision boundary, thus improving target-domain generalization.

3. Adversarial training methods and distribution alignment mechanisms

3.1. Principles of adversarial learning and domain-invariant features

Adversarial learning stems from Generative Adversarial Networks (GANs), centered on a zero-sum game between a generator and a discriminator. The generator produces realistic samples to fool the discriminator, while the discriminator distinguishes real from generated samples [9]. By introducing this principle into domain adaptation, adversarial mechanisms can be used to encourage the feature extractor to learn domain-invariant features. Specifically, if the feature distributions of the source and target domains differ significantly, the domain discriminator can easily distinguish the domain of origin, resulting in a small classification loss. Conversely, when the feature distributions become similar, the discriminator struggles to identify the source, leading to a larger classification loss, with accuracy approaching random chance. The domain discriminator loss directly measures the distance between source and target feature distributions. Therefore, adversarial domain adaptation aims to train the feature extractor to learn domain-invariant features through an adversarial game with the discriminator. During training, the discriminator minimizes its loss to separate source and target features, while the feature extractor maximizes it to produce features the discriminator cannot distinguish. Through alternating optimization, the two eventually reach equilibrium. At this point, the domain discriminator can no longer distinguish feature origins, indicating that the feature extractor has learned domain-invariant features that are irrelevant to domain attributes and relevant only to semantic categories.

3.2. DANN architecture and adversarial mechanism

Introduced by Ganin et al., DANN first applied adversarial learning to domain adaptation [2]. In particular, comprising three components, it includes a feature extractor mapping inputs to features, a label predictor, and a domain discriminator separating source and target features. Placed between the feature extractor and domain discriminator, a Gradient Reversal Layer (GRL) lets the extractor obtain domain-invariant features. During backpropagation, the GRL reverses the discriminator gradient, making the extractor adjust features and reduce domain discrepancy, as shown in Figure 2.

DOI: 10.54254/2755-2721/2026.GU33194

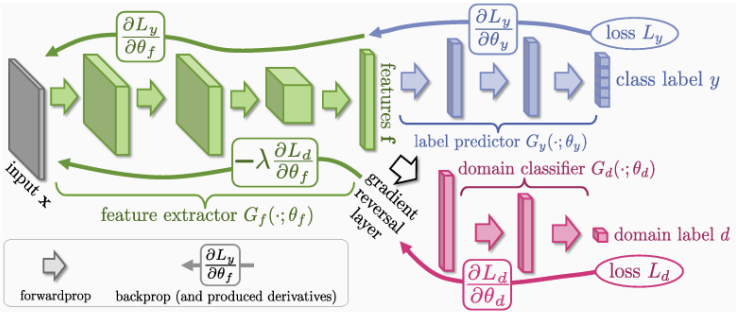


Figure 2. DANN architecture [2]

The GRL plays distinct roles in forward and backward propagation. During forward propagation, it acts as an identity mapping, passing the input unchanged. During backpropagation, it multiplies the gradient from the domain discriminator by a negative constant before passing it to the feature extractor. Consequently, when the discriminator loss is backpropagated, the GRL reverses the gradient, forcing the feature extractor to update parameters and raise the discriminator loss, making it harder to distinguish feature origins. In this design, DANN sets up an adversarial game: the discriminator minimizes its classification loss to separate source and target features, while the feature extractor maximizes it to produce indistinguishable features. Alternating optimization leads to equilibrium, where the discriminator can no longer tell feature origins, showing that the extractor has learned domain-invariant representations.

3.3. Limitations and improvements of DANN

In domain adaptation, DANN aligns only the global distribution while ignoring class-level features, and its shared feature extractor limits model expressiveness. To address the limitation of a shared feature extractor, ADDA separates the feature extractors for source and target domains [4]. Unlike DANN, which employs a shared encoder, ADDA trains separate feature extractors for each domain. The source encoder and classifier are pretrained on source data; the source encoder is then fixed, and the target encoder is adversarially trained to align its output with the source. This approach removes the parameter-sharing constraint, letting each domain learn independent features, and its staged training is more stable than end-to-end training. To enhance class-level alignment, CDAN incorporates class information into adversarial training [5]. Unlike DANN, where the discriminator sees only features, CDAN feeds a joint representation of features and classifier predictions into the discriminator. In particular, multilinear conditioning captures cross-covariance through the outer product of features and predictions, thus allowing the domain discriminator to identify conditional distribution differences across classes. Meanwhile, entropy conditioning downweights samples with high prediction uncertainty, enabling the discriminator to focus on easily transferable samples and ignore uncertain ones. By adding class information, CDAN extends global distribution alignment to class-level conditional alignment, outperforming DANN and ADDA on multiple datasets. Besides, DANN has been extended with DSN, which models domain-private and shared components to capture domain-specific information and learn domain-invariant features. A difference loss enforces orthogonality between components, purifying shared features [10]. Meanwhile, MADA introduces class-specific domain discriminators for multi-class tasks, each aligning features for one class, achieving finer class-level alignment than a single discriminator [11]. These methods address DANN's limitations and advance adversarial domain adaptation.

DOI: 10.54254/2755-2721/2026.GU33194

4. Decision boundary-driven feature alignment and its extensions

4.1. Maximum classifier discrepancy and decision boundary alignment

Through adversarial training, global distribution alignment between source and target domains is achieved by opposing a domain discriminator to a feature extractor. However, they consider only global distribution and neglect feature adjustments near class decision boundaries. Since the domain discriminator only distinguishes domain origins and does not consider class information, the feature extractor may still generate ambiguous features near decision boundaries, leading to classification errors. To address this, Saito et al. proposed Maximum Classifier Discrepancy (MCD), which uses prediction differences between two classifiers on the target domain to identify ambiguous regions and guide the feature extractor, pulling target features toward the source decision boundary [1]. This effectively compensates for the limitations of global distribution alignment, which often overlooks class boundaries. As illustrated in Figure 3, samples where the two classifiers agree are considered confident, while those where they disagree are near the decision boundary. During training, MCD alternates between maximizing and minimizing classifier discrepancy, revealing ambiguous regions and pulling them toward confident areas, thus complementing global distribution alignment with boundary refinement. Moreover, according to the domain adaptation theory proposed by Ben-David et al., when both classifiers perform correctly on the source domain, their prediction discrepancy approaches zero, and the $\mathscr{H}\Delta\mathscr{H}$ distance can be approximated by the upper bound of prediction discrepancy on the target domain [3]. Based on this, MCD effectively estimates and reduces this distance by maximizing and minimizing classifier discrepancy, lowering the target domain error.

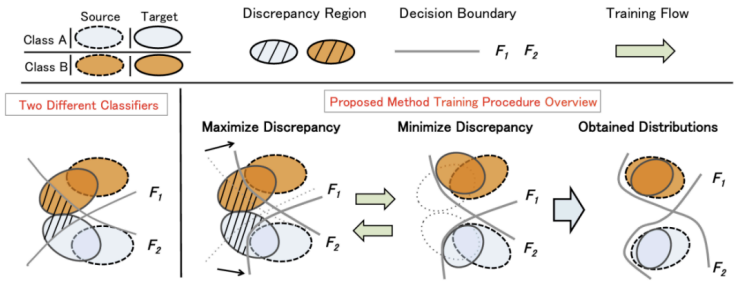


Figure 3. Schematic of MCD boundary alignment mechanism [1]

4.2. MCD architecture and three-step training strategy

Initializing the two classifiers differently allows MCD to form distinct decision boundaries. Its network structure consists of a feature extractor and these two classifiers. The training process follows a three-step alternating strategy, where each step updates either the feature extractor or the classifiers, collectively achieving boundary-driven feature alignment, as shown in Table 1 [1].

Table 1. Three-step training strategy of MCD

Step	Target Updated	Operation
Step A	Feature extractor + both classifiers	Train simultaneously on source data to ensure correct classification of source samples
Step B	Both classifiers (feature extractor fixed)	Maximize prediction discrepancy (L1 distance) on target domain
Step C	Feature extractor (both classifiers fixed)	Minimize prediction discrepancy on target domain

DOI: 10.54254/2755-2721/2026.GU33194

During training,Step A builds a classification foundation by training the feature extractor and classifiers on the source domain, maintaining consistent decision boundaries. Following this, Step B maximizes prediction discrepancy on the target domain, highlighting ambiguous samples near the decision boundary. Then, Step C minimizes the discrepancy, pulling samples back into confident regions and progressively aligning target features near the decision boundary. Existing studies have shown that MCD significantly outperforms adversarial training methods such as DANN on tasks such as transferring from SVHN to MNIST and from SYN SIGNS to GTSRB [1].

4.3. Extensions and improvements of MCD

Inspired by the success of MCD, discrepancy maximization methods have been extended in several directions, including multi-classifier expansion, discrepancy metric improvements, and integration with other mechanisms.In multi-classifier design, increasing the number and diversity of classifiers enables better capture of ambiguous samples near decision boundaries. For example, MMCD_DA extends MCD to multiple classifiers and defines a discrepancy loss based on classifier dissimilarity, improving classification stability [6]. Similarly, the 3C method uses a triangular discrepancy loss and multi-classifier voting, boosting robustness to noise and reducing the risk of classifier collusion [1]. In discrepancy measurement, the Wasserstein distance replaces the traditional L1 distance, alleviating gradient vanishing in early training [12]. Building on this, SLOT integrates the sliced Wasserstein distance into MCD, converting high-dimensional distribution comparisons into integrals of one-dimensional Wasserstein distances via random projections, preserving geometric meaning while cutting computational cost [7]. In a different approach, MDD revisits discrepancy maximization theoretically, introducing a margin parameter to improve robustness and showing its close relation to target domain error [8]. Additionally, discrepancy maximization methods can be combined with other mechanisms to boost performance. For example, DIRT-T combines entropy minimization with virtual adversarial training, optimizing decision boundaries via natural gradients [13]. Co-regularized Alignment builds multiple feature spaces, aligns source and target distributions in each, and enforces consistency in target predictions via a co-regularization term, reducing misalignment risk [14]. Meanwhile, STAR models classifiers as Gaussian distributions and samples them stochastically, allowing flexible scaling without extra parameters and seamless integration into existing frameworks [15].

5. Mechanisms and performance comparison: distribution vs. boundary alignment

In domain adaptation, feature alignment varies in mechanism and objective, with DANN aligning global features via an adversarial game between domain discriminator and feature extractor. These methods mainly tackle representation shift but need careful tuning and can be unstable, whereas boundary alignment methods like MCD use classifier discrepancies to refine features near decision boundaries. Although the latter are more stable, they involve complex two-stage procedures and depend on classifier initialization. The differences between the two paradigms are highlighted, as shown in Table 2. Accordingly, task traits determine whether to use one strategy or combine both.

DOI: 10.54254/2755-2721/2026.GU33194

Table 2. Comparison of two feature alignment paradigms

Aspect	Distribution Alignment (Adversarial Training)	Boundary Alignment (Discrepancy Maximization)
Core Mechanism	Adversarial game between domain discriminator and feature extractor	Optimization of discrepancy between two classifiers
Alignment Objective	Global distribution matching	Decision boundary alignment
Addressed Issue	Representation shift	Decision boundary shift
Training Approach	End-to-end (with gradient reversal)	Two-stage alternation
Training Stability	Difficult, requires tuning	Relatively stable
Representative Works	DANN [1], ADDA [4], CDAN [5]	MCD [3], MMCD_DA [6], SLOT [7], MDD [8]

In terms of remaining challenges, three core issues persist across both paradigms. First, negative transfer arises when misaligned features mix across classes, degrading target performance. Besides, most methods depend on empirical designs without rigorous generalization analysis, while current approaches mainly target single-source, closed-set settings, thus limiting multi-source, continuous, or open-set adaptation. Moreover, adversarial training can be unstable and needs careful tuning, while discrepancy maximization demands efficient optimization strategies. At the feature learning level, self-supervised tasks such as rotation prediction or contrastive learning provide additional signals to support alignment. Transformer architectures replace convolutional networks, using self-attention to capture global image relations and improve feature expressiveness. At the model transfer level, vision-language foundation models like CLIP, pretrained on large image-text datasets, can be adapted via prompt learning or lightweight adapters for rapid transfer, reducing reliance on extensive target data. For paradigm expansion, efficient adaptation methods can be developed, including test-time adaptation, which updates models online with test data, and continuous domain adaptation, allowing models to adjust to evolving domains. At the theoretical level, rigorous frameworks can offer generalization bounds to guide algorithm design and hyperparameter tuning.

6. Conclusion

This paper reviews feature matching in visual unsupervised domain adaptation, further examining adversarial distribution matching and discrepancy-driven boundary matching. Adversarial methods, exemplified by DANN, align source and target features globally via a domain discriminator, with later models such as ADDA and CDAN enhancing flexibility and enabling fine-grained, class-level alignment. Discrepancy maximization methods, exemplified by MCD, optimize decision boundaries via classifier disagreements, with later extensions enhancing classifier diversity, refining metrics, and integrating additional mechanisms. Through comparative analysis, it is found that distribution alignment excels at mitigating representation shifts and capturing complex distributions, although it may suffer from training instability. In contrast, boundary alignment effectively handles decision boundary shifts with stable training, though implementation is slightly more complex. Importantly, these paradigms are not mutually exclusive, and recent research increasingly integrates them to balance global distribution alignment with local boundary refinement. However, there is currently a lack of quantitative performance comparisons across methods under unified experimental settings. Most work targets image classification, with limited attention to semantic segmentation or object

DOI: 10.54254/2755-2721/2026.GU33194

detection. Existing approaches mainly focus on single-source, closed-set scenarios, with restricted applicability to multi-source, continuous, or open-set domains. Future studies should benchmark methods, broaden task coverage,compare method similarities and differences, and explore solutions for complex domain adaptation scenarios to advance feature alignment and robust vision systems.

References

[1] Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3723-3732.
[2] Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In International conference on machine learning, 1180-1189.
[3] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79(1), 151-175.
[4] Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 7167-7176.
[5] Long, M., Cao, Z., Wang, J., & Jordan, M. I. (2018). Conditional adversarial domain adaptation. Advances in neural information processing systems, 31.
[6] Yang, Y., Kim, T., & Wang, G. (2021). Multiple classifiers based maximum classifier discrepancy for unsupervised domain adaptation. arXiv preprint arXiv: 2108.00610.
[7] Lee, C. Y., Batra, T., Baig, M. H., & Ulbricht, D. (2019). Sliced wasserstein discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 10285-10295.
[8] Zhang, Y., Liu, T., Long, M., & Jordan, M. (2019). Bridging theory and algorithm for domain adaptation. In International conference on machine learning, 7404-7413.
[9] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative adversarial nets. NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2, 2672-2680.
[10] Bousmalis, K., Trigeorgis, G., Silberman, N., et al. (2016). Domain separation networks. Advances in neural information processing systems, 29.
[11] Pei, Z., Cao, Z., Long, M., & Wang, J. (2018). Multi-adversarial domain adaptation. In Proceedings of the AAAI conference on artificial intelligence, 32(1).
[12] Shen, J., Qu, Y., Zhang, W., & Yu, Y. (2018). Wasserstein distance guided representation learning for domain adaptation. In Proceedings of the AAAI conference on artificial intelligence, 32(1).
[13] Shu, R., Bui, H. H., Narui, H., & Ermon, S. (2018). A dirt-t approach to unsupervised domain adaptation. arXiv preprint arXiv: 1802.08735.
[14] Kumar, A., Sattigeri, P., Wadhawan, K., et al. (2018). Co-regularized alignment for unsupervised domain adaptation. Advances in neural information processing systems, 31.
[15] Lu, Z., Yang, Y., Zhu, X., Liu, C., Song, Y. Z., & Xiang, T. (2020). Stochastic classifiers for unsupervised domain adaptation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 9111-9120.