Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/46/20241615

Sequential domain adaptive network in objection detection

Ning Xie

College of Mathematics and Computer Science, Zhejiang A&F University Hangzhou, China

19157736552@163.com

Abstract. Recent advancements in object detection models heavily depend on annotated data, which often requires substantial time and resources to gather. To tackle this challenge, the field of unsupervised domain adaptive object detection focuses on adapting detectors trained on a labeled source domain for effective use in an unlabeled target domain. However, this process faces challenges such as high computational demands and instability in training, due to the significant discrepancies between domains. Our paper presents an efficient Sequential Domain Adaptive Network (SDAN) designed to incrementally minimize these domain differences. The network effectively mitigates the disparities in both initial and subsequent phases. We utilize the Fourier transform for spectrum switching, eliminating the need for training in this stage. To further address the issue of domain shift, the model employs unsupervised adversarial learning for the alignment of varying feature distributions. Extensive experiments demonstrate that our model exhibits strong performance against the most advanced domain adaptive object detection models currently available.

Keywords: Object detection, Unsupervised domain adaptation, Fourier transform, Adversarial learning

1. Introduction

Objection Detection, aiming to localize and classify the objects, holds significant societal implications across various domains, encompassing security like surveillance, and automated driving. Most objection detection methods[1][2] rely on processing large amounts of densely annotated data, which can be quite time-consuming and require a significant amount of manual effort. Moreover, the performance of models trained on specific data domains often diminishes significantly when applied to different data domains, a decline attributed to large data biases and discrepancies between different domains like colors, weather, and lighting conditions[3][4][5].

Overcoming these limitations can be achieved through the method of Unsupervised Domain Adaptation (UDA)[6][7]. This technique allows a model initially trained on an annotated source domain to effectively operate in a different, unlabeled target domain. However, numerous methods were invented to solve one complicated problem about semantic segmentation[8][9][10], while only a limited number of methods were designed on another complicated task- objection detection[11]. What's more, although UDA can mitigate several problems, it's not without its drawbacks when it meets large domain gaps[12], leading to unsteady training processes and sub-optimal results.

In our research, we introduce a Sequential Domain Adaptation Network(SDAN) shown in 'figure 1', which simplifies complex domain adaptation tasks into a two-step process but without a significant

© 2024 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: 10.54254/2755-2721/46/20241615

increase in computational demands. Initially, the model employs the Fast Fourier Transform (FFT) technique[18] for a Fourier Transform, replacing the source domain's low-frequency components with those from the target domain. This phase effectively minimizes the distributional differences between source and target datasets while eschewing the use of deep neural networks or adversarial training. Following this, the model integrates an Uncertainty-Aware Adversarial Learning approach, which mitigates the misalignment of styles between the source and target domains while preserving semantic content.

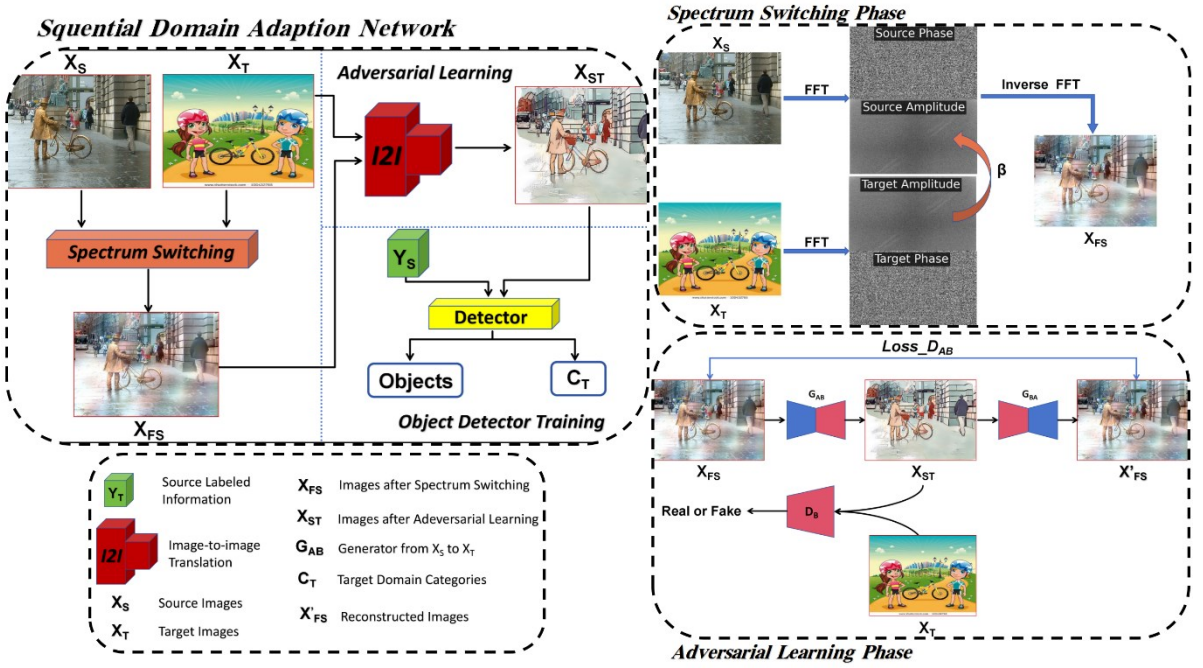


Figure 1. The proposed successive Domain Adaptive Architecture. The algorithm consists of two domain adaptive phases and one object detection phase: In Spectrum Switching, we employ the Fourier transform to replace source style information with target style information to minimize distribution differences. In the Adversarial Learning Phase, we use G_{AB} in CycleGAN to produce the generated target style source image. In the third step, we use those images with source tags to train the Yolov8 algorithm and test it on target images without any tags.

We evaluate our method on two domains PASCAL VOC 2012[13] and Clipart1k[14] each characterized by significant domain gaps. The proposed sequential method demonstrates our SDAN has superior performance over current state-of-the-art methods, indicating that our approach overcomes existing limitations with enhanced efficiency in domain adaptation. Our contributions can be the following: First, we have introduced an efficient sequential method that requires minimal adversarial training and network resources, yet significantly improves the process of aligning feature distribution differences. Second, we have carried out comprehensive adaptation experiments across two domains, each characterized by significant domain gaps, and have successfully achieved state-of-the-art performance.

2. Related Work

2.1. Object Detection

Recently, most state-of-the-art object detection models have generally split into two streams: proposal-based and proposal-free models. Proposal-based methods like Fast R-CNN[15] first generate region proposals and then classify them for final detection results. To further accelerate the process and reduce

DOI: 10.54254/2755-2721/46/20241615

computational consumption, proposal-free models like YOLO[16] eschew separate proposals by retrospecting the coordinates of predefined anchors and simultaneously classifying categories. However, most of them still use a large number of labeled trainings which requires a high labor cost. What's more, due to a large number of discrepancies, this model can overfit labeled source data while having bad performance in unlabeled target data. Therefore, in this paper, we focus on the invention of cross-domain object detection technology.

2.2. Domain Adaption

Domain Adaption addresses the challenge of domain shift by bridging the gap between a well-labeled source domain and a weakly-labeled target domain. Domain adversarial learning[17] plays a key role in domain adaptive methods. Most recent approaches typically involve employing adversarial learning to align the differing data distributions of the source and target domains. Furthermore, owing to the potential loss of semantic information in spatial translations, the application of the Fast Fourier Transform (FFT) technique[18] has gained great interest. By selectively modifying frequency components, it can conserve semantic content to the maximum extent[10]. Therefore, in this paper, we leverage this frequency domain adaptation approach in the first stage of our process to minimize the cross-domain discrepancy.

3. Methods

This section describes our approach, including two sections on problem definition and the specific structure of the model.

We propose to disintegrate domain adaptation tasks into two subtasks, denoted as S1 and S2. We define source images, intermediate images, and target images as $X_S \in R^{H\times W\times 3}$, $X_D \in R^{H\times W\times 3}$ and $X_T \in R^{H\times W\times 3}$ respectively. The object labels related to X_S can be denoted as $Y_S = \{C_S, B_S\}$. The $C_S \in (1, C)^N$ and $B_S \in (1, C)^{N\times 4}$ represent object classes C and bounding box coordinates respectively, where H, W, N represent image height, image width, and the class number of objects respectively. The objective of conventional domain adaptation $G(X_S, X_T)$ can be expressed in SDAN as equation (1):

$$G(X_S, X_T) = S_1(X_S, X_D) + S_2(X_D, X_S) \tag{1}$$

This objective aims to generate a high-performing cross-domain detector G by utilizing $\{X_S, Y_S, X_T\}$ with the assistance of X_D.

3.1. Spectrum Switching

In the first stage, as illustrated in the *Spectrum Switching Phase* of 'figure 1', the FFT algorithm converts images from the spatial domain to the frequency domain as equation (2). Here, the amplitude and the phase are defined as in equation (3) and equation (4), respectively. The variables u and v represent the spatial frequency coordinates in the x and y directions, respectively, where x and y represent the pixel coordinates.

$$F(u,v) = A(u,v)\cdot e^{i\phi(u,v)} \tag{2}$$

$$A(u,v) = |F(u,v)| = \left(\mathrm{Re}\big(F(u,v)\big)^2 + \mathrm{Im}\big(F(u,v)\big)^2\right)^{1/2} \tag{3}$$

$$\phi(u,v) = \arctan 2\left(\mathrm{Im}\big(F(u,v)\big), \mathrm{Re}\big(F(u,v)\big)\right) \tag{4}$$

Subsequently, the amplitude in the source domain is replaced with that in the target domain, controlled by the parameter β. This parameter β, ranging from 0 to 1, specifies the frequency range to be replaced from 0 to β. Choosing an appropriate β is crucial, as a too-large value can alter the semantic content of the source images. The final step involves using the Inverse FFT to reconstruct the source images in the target style. This overall process is notably efficient, as it does not require adversarial training or labeled data. The reconstructed images significantly reduce distribution discrepancies.

DOI: 10.54254/2755-2721/46/20241615

3.2. Adversarial Learning

The second stage, depicted in the *Adversarial Learning Phase* of 'figure 1', employs adversarial learning[6] to further reduce the feature discrepancies. Our aim is to refine GAB such that it generates images that the discriminators D_S find challenging to differentiate from real target images X_T.

The output images from the Spectrum Switching Phase serve as the source images for this subtask. The training process involves G_{AB}, which generates fake target images X_{ST} from source images. Subsequently, G_{BA} generates a reconstructed version of the source image, denoted as $X_S^{'}$. This allows for the calculation of discrepancies between X_S and $X_S^{'}$, which aids in the refinement of the G_{AB} model. The loss functions for these components are expressed as equation (5):

$$L_{S\to T}(G, D_T, X_s, X_T) = E\left[\log\left(D_T(X_T)\right)\right] + E\left[\log\left(1 - D_T\left(G(X_S)\right)\right)\right] \tag{5}$$

where G represents a generator from X_S to X_T. We also utilize reversed processes from the target domain to generate the source domain, which can further help us improve our GAB model. The loss of these parts can be shown as equation (6):

$$L_{S\to T}(G, D_T, X_s, X_T) = E\left[\log\left(D_T(X_T)\right)\right] + E\left[\log\left(1 - D_T\left(G(X_S)\right)\right)\right] \tag{6}$$

where G represents a generator from X_S to X_T. Therefore, the overall loss function L can be shown as equation (7):

$$L = L_{S\to T}(G, D_S, X_S, X_T) + L_{T\to S}(F, D_T, X_S, X_T) \tag{7}$$

Subsequently, we train the object detector using the synthetic source images X_{ST} styled to resemble the target, accompanied with their corresponding source label information $Y_S = \{C_S, B_S\}$. This process showcases the exemplary performance of our model.

4. Experiments

4.1. Datasets

Our methodology involves training our model across two divergent domains: utilizing PASCAL VOC2012[13] as the annotated source training set and Clipart1k[14]as the non-annotated target training set. For assessing the efficacy of various methods, we leverage the validation subset from the target dataset. Both datasets include the same object categories, maintaining uniformity in the object detection challenges.

• The PASCAL VOC2012[13] dataset presents a multifaceted real-world scenario, containing 5,717 images for training and 5,823 for validation. This dataset provides bounding box annotations over 20 unique classes.

• The Clipart1k[14] dataset comprises 1,000 caricature-like images, segregating 800 for training and 200 for validation. These artistically created images display stark contrasts to real-world imagery and are annotated with bounding boxes for the same 20 classes as found in PASCAL VOC2012[13].

4.2. Experiments Results

In this research, we employ YOLOv8 as a benchmark to evaluate the prowess of our proposed model, SDAN. Our assessment spans across markedly diverse data realms: PASCAL VOC 2012, a dataset reflective of real-world environments, and Clipart1k, which encompasses a number of handcrafted images. The relative performance of SDAN in comparison to other cutting-edge models is comprehensively presented(seen in table 1). Specifically targeting the task of adaptation from PASCAL VOC to CLIPART1K, we measure the Average Precision (AP%) for individual object categories and the Mean Average Precision (MAP%) across all categories within the CLIPART1K validation dataset. Additionally, we provide visual representations of some image detection outcomes in 'figure 2'.

DOI: 10.54254/2755-2721/46/20241615

Table 1. Partial quantitative comparison of SDAN with state-of-the-art domain adaptive methods.

Methods	aero	bicy	bird	bott	car	cat	pers	plant	sheep	sofa	train	tv	table	mAP
Source-only	4.5	**75.2**	31.5	5.0	64.7	6.9	49.8	52.1	9.6	20.7	34.7	21.8	32.7	31.2
CRDA[20]	7.9	50.7	33.2	13.2	68.4	3.2	**68.8**	55.7	15.3	27.8	15.0	30.7	30.1	37.7
GPA[21]	12.5	54.8	40.6	17.1	**71.6**	4.5	63.4	58.9	13.9	11.4	45.0	31.9	21.9	38.6
UaDAN[11]	**35.0**	72.7	**41.0**	21.3	69.8	2.3	62.2	**61.3**	13.9	7.6	**48.6**	23.9	31.0	40.2
SDAN(Ours)	26.6	72.3	33.9	**49.3**	55.2	**23.2**	62.9	43.4	**21.6**	**36.5**	47.8	**45.6**	**46.4**	**43.9**

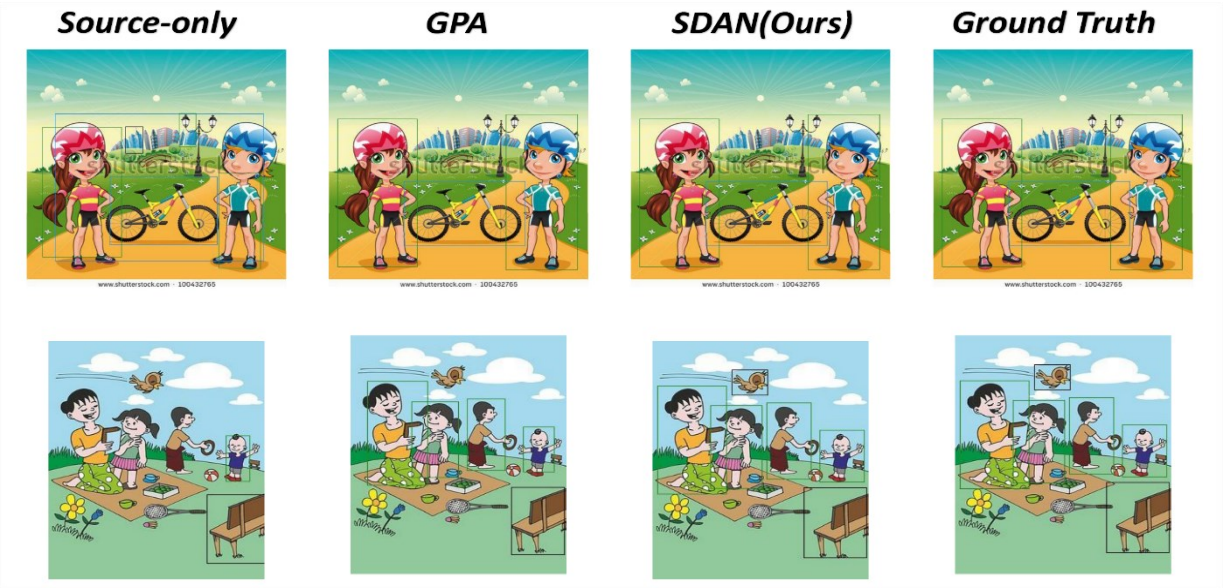


Figure 2. Visual Representations of Some Image Detection Outcomes.

5. Discussion

As illustrated in Fig, our SDAN model attains a peak accuracy with a mean Average Precision (mAP) of **43.9%**. Impressively, SDAN outshines in 8 specific categories, including 'bott', 'table', and 'tv', achieving the top mAP scores in these classes. Particularly noteworthy is the performance in categories like 'bott' and 'tv', where our model reaches an accuracy of 49.3%, surpassing other models by a significant 20% margin. Additionally, in 'figure 3', we utilize t-SNE[19] for a graphical depiction of how our model progressively reduces the disparities between the source and target domains. This visualization effectively demonstrates the sequential improvement stages inherent in SDAN's design. Consequently, these results demonstrate the superiority of SDAN over current leading domain adaptive object detection models.

Nevertheless, two experimental data fields belong to a less-represented category. This poses a challenge as model performance can substantially decline when handling a broad spectrum of classes. Additionally, while the initial stage of our model does not require training, the process of converting images in the second stage is time-consuming. Thus, there remains ample scope for future advancements in developing models that are both more robust and user-friendly.

6. Conclusion

The article introduces a novel approach for unsupervised domain adaptation using a sequential domain adaptation network. The method includes a two-stage algorithm that effectively bridges significant disparities between two domains. It begins with spectrum switching, which uniquely operates without training in the frequency domain, followed by adversarial learning that enhances the algorithm's effectiveness. The superior performance of this approach is demonstrated through a series of

DOI: 10.54254/2755-2721/46/20241615

experiments across two highly distinct feature data domains. Future research endeavors will concentrate on scaling this methodology to encompass cross-domain datasets including an extensive range of categories.

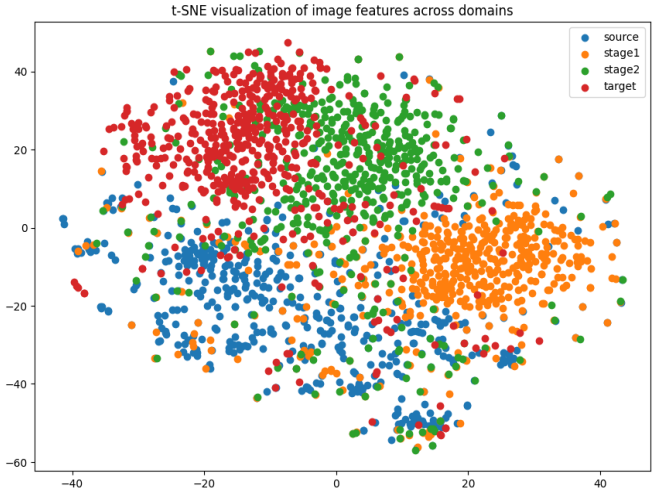


Figure 3. Visualization of features distribution via t-SNE[19]. We randomly choose 500 images of each domain. Each dot represents an image, demonstrating how our algorithm sequentially aligns different distributions. 'stage1' refers to the image features following Spectrum Switching, while 'stage2' denotes the image features after undergoing Adversarial Learning.

References

[1] Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. Proceedings of the IEEE (2023)

[2] Zhao, Z.-Q., Zheng, P., Xu, S.-t., Wu, X.: Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems 30(11), 3212–3232 (2019)

[3] Liu, W., Ren, G., Yu, R., Guo, S., Zhu, J., Zhang, L.: Image-adaptive yolo for object detection in adverse weather conditions. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 1792–1800 (2022)

[4] Liu, X., Li, W., Yang, Q., Li, B., Yuan, Y.: Towards robust adaptive object detection under noisy annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14207–14216 (2022)

[5] Xu, Y., Sun, Y., Yang, Z., Miao, J., Yang, Y.: H2fa r-cnn: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14329–14339 (2022)

[6] Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle- consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)

[7] Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)

[8] Lo, S.-Y., Oza, P., Chennupati, S., Galindo, A., Patel, V.M.: Spatio-temporal pixel-level contrastive learning-based source-free domain adaptation for video semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10534–10543 (2023)

[9] Truong, T.-D., Le, N., Raj, B., Cothren, J., Luu, K.: Fredom: Fairness domain adaptation approach to semantic scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19988–19997 (2023)

[10] Huang, J., Guan, D., Xiao, A., Lu, S.: Rda: Robust domain adaptation via fourier adversarial attacking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8988–8999 (2021)

[11] Guan, D., Huang, J., Xiao, A., Lu, S., Cao, Y.: Uncertainty-aware unsupervised domain adaptation in object detection. IEEE Transactions on Multimedia (2021)

DOI: 10.54254/2755-2721/46/20241615

[12] Hsu, H.-K., Yao, C.-H., Tsai, Y.-H., Hung, W.-C., Tseng, H.-Y., Singh, M., Yang, M.-H.: Progres- sive domain adaptation for object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 749–757 (2020)
[13] Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision 111, 98– 136 (2015)
[14] Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5001–5009 (2018)
[15] Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
[16] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
[17] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research 17(1), 2096–2030 (2016)
[18] Nussbaumer, H.J., Nussbaumer, H.J.: The Fast Fourier Transform. Springer (1982)
[19] Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research 9(11) (2008)
[20] Xu, C.-D., Zhao, X.-R., Jin, X., Wei, X.-S.: Exploring categorical regularization for domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11724–11733 (2020)
[21] Zheng, Y., Huang, D., Liu, S., Wang, Y.: Cross-domain object detection through coarse-to-fine feature adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13766–13775 (2020)