Contrastive Learning and Cycle Consistency-based Transductive Transfer Learning for Target Annotation

SHOAIB MERAJ SAMI
LCSEE Dept., West Virginia University, Morgantown, WV, USA

NASSER M. NASRABADI, Fellow, IEEE
LCSEE Dept., West Virginia University, Morgantown, WV, USA

RAGHUVEER RAO, Fellow, IEEE
Army Research Laboratory, Adelphi, MD, USA

***Abstract*—Annotating automatic target recognition images is challenging; for example, sometimes there is labeled data in the source domain but no labeled data in the target domain. Therefore, it is essential to construct an optimal target domain classifier using the labeled information of the source domain images. For this purpose, we propose a transductive transfer learning (TTL) network consisting of an unpaired domain translation network, a pre-trained source domain classifier, and a gradually constructed target domain classifier. We delve into the unpaired domain translation network, which simultaneously optimizes cycle consistency and modulated noise contrastive losses (MoNCE). Furthermore, the proposed hybrid CUT module integrated into the TTL network generates synthetic negative patches by noisy features mixup, and all the negative patches provide modulated weight into the NCE loss by considering similarity to the query. Apart from that, this hybrid CUT network considers query selection by entropy-based attention to specifying domain variants and invariant regions. The extensive analysis depicted that the proposed transductive network can successfully annotate civilian, military vehicles, and ship targets into the three benchmark ATR datasets. We further demonstrate the importance of each component of the TTL network through extensive ablation studies into the DSIAC dataset. The source code is available at https://github.com/ShoaibMerajSami/TTL_ATR.**

Index Terms—**Automatic Target Recognition, CycleGAN, CUT, MoNCE, QS-Attention, Feature mixup, Transductive Learning, Semi-supervised Learning.**

Manuscript received XXXXX 00, 0000; revised XXXXX 00, 0000; accepted XXXXX 00, 0000.." information, e.g. *(Corresponding author: M. Smith)*. Here you may also indicate if authors contributed equally or if there are co-first authors.

Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@boulder.nist.gov). University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu). for Metals, Tsukuba 305-0047, Japan (e-mail: author@nrim.go.jp).

Mentions of supplemental materials and animal/human rights statements can be included here.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

I. INTRODUCTION

Automatic target recognition (ATR) is essential for many civilian and military applications, such as target detection and vehicle classification [1]. It can help to alleviate civilian casualties, aid in rescue operations, and guide land and maritime vehicles. A large amount of labeled data [2] is required to properly implement ATR algorithms. However, annotating target vehicles are expensive, time-consuming, and cumbersome. Many semi-supervised and self-supervised algorithms have been proposed in the literature to alleviate the manual annotation process of data [3], [4]. Generally, ATR imagery is captured in the visible (VIS), mid-wave infrared (MWIR), long-wave infrared (LWIR), and synthetic aperture radar (SAR) domains [5], [6], [7], [8]. There are military scenarios where we have labeled images in one domain (source) but no annotated images in the other domain (target). Therefore, annotating target domain images is challenging when considering only the information from the source domain images. Transductive transfer learning (TTL) [9], [10], a semi-supervised learning algorithm can be used to annotate targets by considering the information of unpaired source domain images [11]. Therefore, utilizing the TTL framework to annotate vehicles and ships in different domains could be a promising approach.

Typically, a TTL framework consists of an unpaired image translation network and a classifier. A substantial part of the proposed TTL algorithm is an unpaired domain transformation method that transfers the source image to the target domain. This image-to-image (I2I) translation network should preserve the target chip's class information and its semantic structure. Therefore, constructing an I2I domain translation network is crucial to the TTL model. The unpaired I2I domain translation models can be divided into two major categories: conventional GAN-based methods and contrastive learning-based unpaired image translation methods, which maximize the mutual information between the different domains [12]. Traditional GAN-based unpaired I2I translation networks can be further divided into cycle consistency and one-way translation by bypassing cycle consistency. The cycle-consistency is used in different ways in the literature, such as image-to-latent, latent-to-image, and image-to-image domain. Moreover, recently, contrastive unpaired translation (CUT) learning [13] and its variants have outperformed the traditional unpaired I2I methods [14]. The different variants of CUT networks consider a query selection attention (QS-attention) module [15], re-weighting the negative patches [16], instance-wise negative example generation [17], learned self-similarity-based PatchNCE [18], and two-sided mapping by two PatchNCE networks [19].

In general, the CUT model only considers a one-way translation [14]. However, in the TTL models, there is a need for a two-way translation, similar to DCLGAN [19]. This is because the reconstructed source domain images are used to feed into the pretrained source domain classifier, which backpropagates categorical cross-entropy [20] loss into the TTL network. Therefore, our proposed network is optimized for both PatchNCE [14] and cycle-consistency[21] losses.

The proposed TTL model needs to learn the patch and pixel-based correlation among the visible, infrared, and SAR domains. Furthermore, the IR domain target lacks color information compared to the visible domain target [22]. Therefore, the network ought to discern both the domain-specific and invariant regions. This incident is alluded to in the QS-Attn-CUT [15] network by choosing the appropriate query. In the CUT network, negative patches play an important role because all negative patches ought not to have a similar contrasting force [16]. It should be based on the similarity between negative patches and query, which can be performed by modulated contrast NCE loss [16]. Therefore, to benefit from query selection and modulated contrastive force, we simultaneously used QS-attention [15] and MoNCE loss-based CUT network. Furthermore, the generation of negative patches [23] plays an essential role in contrastive learning [24], [23], and CUT networks [17]. Hence, we employed noisy feature mixup-based synthetic patches [25] for robustness and to create more variations of negative patches.

In this research, we propose the TTL model that will annotate unlabeled target domain images using the information from the source domain images. The main contributions of this work can be summarized as follows:

1) We propose a CycleGAN-based TTL model to annotate target domain images by optimizing the cycle consistency, identity, adversarial, and categorical cross-entropy losses.
2) We further use different variants of CUT-based transductive transfer learning networks to annotate ATR targets. Our robust TTL model takes into account the QS-Attn-based query selection, synthetic negative patch generation, and MoNCE loss-based two CUT network, alongside the preserving cycle consistency. The novelty of our proposed method lies in synthetic negative patch generation and utilizing the MoNCE and cycle consistency losses into the QS-Attn-CUT network.
3) Experiments on two military vehicle datasets and one ship target dataset demonstrates the effectiveness of the proposed model.

Preliminary results of our proposed transductive CycleGAN algorithm is published in [26], introducing the annotation of unlabeled ATR data in the DSIAC dataset. The primary contribution of the current work is to propose a contrastive learning-based transductive learning approach, critically evaluate our algorithms' performance on three ATR datasets, and compare our method with other state-of-the-art methods. Additionally, we provide a detailed description of the hardness of negative patches, attention-based query selection, and synthetic patch generation to improve the annotation performance of the proposed contrastive TTL algorithm which was not reported in [26].

The remainder of our work is developed as follows: Section II presents a literature review related to our study; Section III explains the proposed TTL network architecture and describes each of its modules; Section IV illustrates the effectiveness of the proposed network and each of its components with organized experiments; and finally, in Section V, we conclude our paper.

II. Related Work

This section will discuss the relevant work regarding automatic target recognition, transductive learning, and unpaired image translation network.

A. Automatic Target Recognition

The automatic target recognition literature can be divided into: (a) target detection and (b) target vehicle classification task. Different deep learning methods, such as Faster R-CNN [27], SSD [28], YOLO [29], etc., have been widely employed to detect various civilian and military vehicles in various ATR datasets [1]. The ATR target classification algorithms can be expressed as: i) feature-based and ii) model-based approaches. The performance of the feature-based algorithm is lower than the model-based approaches. The convolutional neural networks (CNNs) [1], generative adversarial networks (GANs) [30], recurrent neural networks (RNN) [31], and autoencoders [32] are widely used in the model-based ATR classification method. Nasrabadi [1] used a deep neural network to classify FLIR images in the Comanche (Boeing–Sikorsky, USA) dataset [33]. In [34], authors proposed Meta-UDA to detect unlabeled MWIR target images by using the labeled visible domain images in the DSIAC dataset [35]. Ding et al. [36] used different variants of augmentation, such as translation, speckle noise augmentation, and generating pose synthetic images to classify SAR images. Furthermore, Wang et al. [37] used a coupled network aggregating a despeckling subnetwork and a neural network classifier for classifying noisy SAR target images.

Generally, training CNN classifiers require a large amount of labeled training data; few-shot learning is used in ATR vehicle classification to alleviate the need for a large amount of training data. Bi-LSTM-based prototypical few-shot learning [38], hybrid inference [39], and meta-learning [40] based on few-shot learning have been employed in the MSTAR SAR dataset. Furthermore,

autoencoder and supervised constraint [32] were exploited to classify the small samples in the ATR dataset. Also, in [30] a semi-supervised method was proposed to classify the ATR SAR images where the original and GAN-generated synthetic images were used to construct an ATR classifier [30].

B. Transductive Transfer Learning

Transductive transfer learning (TTL) is an effective tool for classifying unlabeled data in various applications. For instance, the authors in [41] suggested a cross-domain aspect label propagation-based TTL method for opinion mining and sentiment analysis. TTL methods based on RNN and LSTM were investigated for optical character recognition (OCR) [42]. The authors considered that the OCR and text incorporate specific details in the high-dimensional space. In [43], a linear discriminant analysis-based TTL network was used for emotion recognition in the wild dataset. In [44] a TTL technique utilizing a VGGFace16-Net was exploited for cross-domain expression recognition. In this work, authors optimized cross-entropy loss of the source domain, and regression loss of the target domain. A transductive transfer learning method [45] using cluster correspondence in the domain adaptation using manifold alignment was used to detect zero-day attacks in intrusion detection systems. Furthermore, a TTL model [46] was employed for gesture classification using an electromyographic dataset. A TTL model was developed to detect epilepsy using electroencephalogram signals by exploiting Takagi-Sugeno-Kang's fuzzy logic systems [47], [48]. In [49], [50] a TTL model was used for document and text classification using genetic programming, and distributional correspondence indexing . The crop classification is vital for farming and nutrition management, while occasionally, there remains a small amount of ground truth data. To unravel this matter, the authors utilized a transductive transfer learning-based approach [51] for crop classification. A TTL network based on the deep forest classifier was utilized for cross-domain transfer learning and for measuring accuracy across different datasets, including MNIST, USPS, Amazon, DSLR, Webcam, and Caltech-256 datasets [52].

C. Unpaired I2I Domain Translation

The unpaired domain translation literature is vast [53]. The CycleGAN [21] paper was the breakthrough in this field that consists of bijection mapping by preserving cycle-consistency. In many works, cycle consistency was used in different aspects [54], [55], [56]. Instead of cycle-consistency loss, one-way unpaired image translation was also used in several papers [57], [58], [59]. Nevertheless, those approaches employed specific distance functions or full image statistics. However, CUT [14] can improve image quality and faster convergence by replacing the cycle-consistency information maximization between input and output patches between the source and target domain. The different variants of CUT are a de facto method for unpaired I2I translation. The CUT network uniformly contrasts all the negative patches that are not effective. Optimal transport and self-similarity-based MoNCE loss-based CUT was proposed to alleviate this issue [16]. Hard negative generation [17], learned self-similarity [18] based CUT was also exploited in this field. Entropy-based query selection attention for specifying the domain-specific and domain-invariant information has also been utilized [15]. DCLGAN used two-way translation by considering a CUT network [19].
Furthermore, the denoising diffusion model was successfully exploited in the image-to-image translation method [60], [61]. It was also used for unpaired domain translation in several fields [62], [63].

III. Methodology

In this section, we will describe our proposed transductive CycleGAN and hybrid contrastive learning based unpaired image-to-image translation (H-CUT) network. The first generator of our proposed TTL model transforms the source domain images into target domain images. Then, a second generator transforms the synthetic images back into the source domain images. Moreover, there are dedicated source and target domain CNN-based ATR classifiers that backpropagate their categorical cross-entropy losses to help to train the whole TTL model. Our proposed method is depicted in Figure 1.

A. Generative Adversarial Network (GAN)

GAN [64] is a powerful tool for different types of image generation, image translation, and super-resolution. A generator and a discriminator are the building blocks of a GAN architecture. The GAN generates images from a random variable z; the generator function can be depicted as $G : z \rightarrow y$. The generator $G(z; \theta_g)$ is optimized for its parameters θ_g. On the other hand, the discriminator function D is designed to distinguish between the real image distribution and the generated image distribution. In practice, the discriminator is a simple binary classifier that classifies whether the image is real (y) or fake ($G(z)$). In GAN, the generator and the discriminator compete with each other as a two-player mini-max game [64]. The GAN objective function can be expressed by $\mathcal{L}(G, D)$:

$$\begin{aligned}\mathcal{L}(G, D) = \min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] \\ +\mathbb{E}_{z\sim p_z(z)}[1 - \log D(G(z))],\end{aligned} \tag{1}$$

where $p_{\text{data}}(x)$ is the probability distribution of the data and $p_z(z)$ is a Gaussian distribution for the random variable z.

B. Unpaired Image Translation Networks Architecture

1. *CycleGAN*

In the unpaired image-to-image translation literature, CycleGAN [21] has been successfully exploited as a

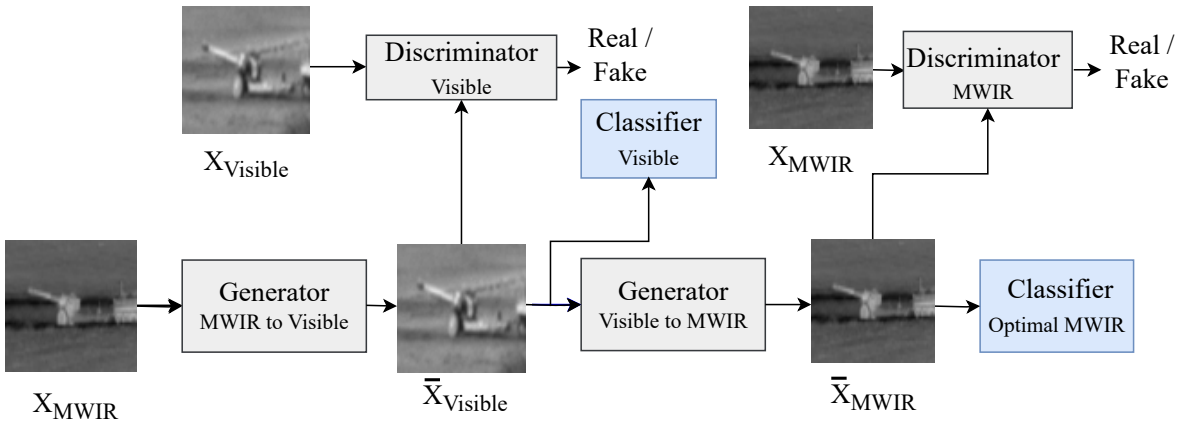


Fig. 1: Block diagram of the deep transductive transfer learning framework for automatic target classification.

powerful domain-to-domain transformation tool. This algorithm consists of two generators and two discriminators. The first generator (G) translates the domain X images into the domain Y images. The second generator (F) translates the domain Y images into the domain X images. The combined adversarial loss, denoted by $\mathcal{L}_{GAN}(G, F, D_x, D_y)$, for the CycleGAN's generators G and F can be expressed by:

$$\begin{aligned}\mathcal{L}_{GAN}(G, F, D_x, D_y) = \min_G \max_{D_y} \mathbb{E}_{y\sim p_{\text{data}}(y)}[\log D_y(y)] \\ +\mathbb{E}_{x\sim p_{\text{data}}(x)}[1 - \log D_y(G(x))] \\ + \min_F \max_{D_x} \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D_x(x)] \\ +\mathbb{E}_{y\sim p_{\text{data}}(y)}[1 - \log D_x(F(y))].\end{aligned} \tag{2}$$

The authors of the CycleGAN proposed that the transformation should be a bijection mapping. If there is no bijection mapping, then the generator (G) can map an image in an infinite way that follows a similar distribution of domain Y. In the CycleGAN paper, the authors introduced the cycle-consistency loss, which translates the domain X image to the domain Y image then the translated domain Y image is converted back to the domain X image by the second generator F. Mathematically; it can be described as a generator $G : X \rightarrow Y$ and another generator $F : Y \rightarrow X$. Also, the generators G and F should translate in the inverse domain, and mapping should be bijection [21]. The cycle consistency loss can be denoted as $\mathcal{L}_{Cycle}$ where η_1 and η_2 are the hyper-parameters:

$$\begin{aligned}\mathcal{L}_{Cycle}(G, F) =& \eta_1 * \mathbb{E}_{x\sim p_{\text{data}}(x)} \left(\|F(G(x)) - x\|_1\right) \\ &+\eta_2 * \mathbb{E}_{y\sim p_{\text{data}}(y)} \left(\|G(F(y)) - y\|_1\right).\end{aligned} \tag{3}$$

Moreover, the CycleGAN generator has two vehicle identity mapping regularizers [21] that are employed as real images on the target domain and fed as input for the source generator. This loss is called identity loss ($\mathcal{L}_{identity}$) where η_3 and η_4 are the hyper-parameters:

$$\begin{aligned}\mathcal{L}_{Identity}(G, F) = \eta_3 * \mathbb{E}_{y\sim p_{\text{data}}(y)} \left(\|G(y) - y\|_1\right) \\ +\eta_4 * \mathbb{E}_{x\sim p_{\text{data}}(x)} \left(\|F(x) - x\|_1\right).\end{aligned} \tag{4}$$

The total CycleGAN loss ($\mathcal{L}_{CycleGAN}$) is the summation of the adversarial loss, cycle consistency loss, and identity loss where λ_a, λ_b, and λ_c are the hyper-parameters:

$$\begin{aligned}\mathcal{L}_{CycleGAN}(G, F, D_x, D_y) = \lambda_a * \mathcal{L}_{GAN}(G, F, D_x, D_y) \\ +\lambda_b * \mathcal{L}_{Cycle}(G, F) + \lambda_c * \mathcal{L}_{Identity}(G, F).\end{aligned} \tag{5}$$

2. Contrastive Unpaired Image Translation (CUT)

The CUT network consists of the adversarial loss and PatchNCE loss where the PatchNCE loss maximizes the mutual information $I(X, Y) = H(X) - H(X|Y)$. Therefore, the CUT network minimizes the $H(X|Y)$ which is similar to the cycle consistency goal [14]. The total CUT objective can be depicted as:

$$\begin{aligned}\mathcal{L}_{CUT}(G, D, X, Y) = \lambda_g * \mathcal{L}_{GAN}(G, D, X, Y) \\ +\lambda_X * \mathcal{L}_{PatchNCE}(G, H, X) + \lambda_Y * \mathcal{L}_{PatchNCE}(G, H, Y),\end{aligned} \tag{6}$$

where λ_g, λ_X, and λ_Y are the hyper-parameters.

The PachNCE objective can be denoted as:

$$\mathcal{L}_{\substack{Patch\\NCE}}(X, Y) = -\sum_{i=1}^{N} log[\frac{e^{(x_i.y_i/\tau)}}{e^{(x_i.y_i/\tau)} + \sum_{\substack{j=1\\j\neq i}}^{N} e^{(x_i.y_j/\tau)}}], \tag{7}$$

where τ is a temperature parameter and X, Y are the generated target domain and ground truth images, respectively, $X = [x_1, x_2, ..., x_N]$ and $Y = [y_1, y_2, ..., y_N]$ are encoded feature vectors from $1st, 4th, 8th, 12th$ and $16th$ layers of the encoder, and that is passed through a two-layer MLP network [14], [16]. In the Eq. 6, the first component is the adversarial loss, the second component is the PatchNCE loss for domain translation and the third component is the identity loss. The CUT network considers N class classification where anchor provides the same contrastive force to the $N-1$ negative samples that is too strict. To overcome this issue modulated contrastive NCE loss has been proposed [16].

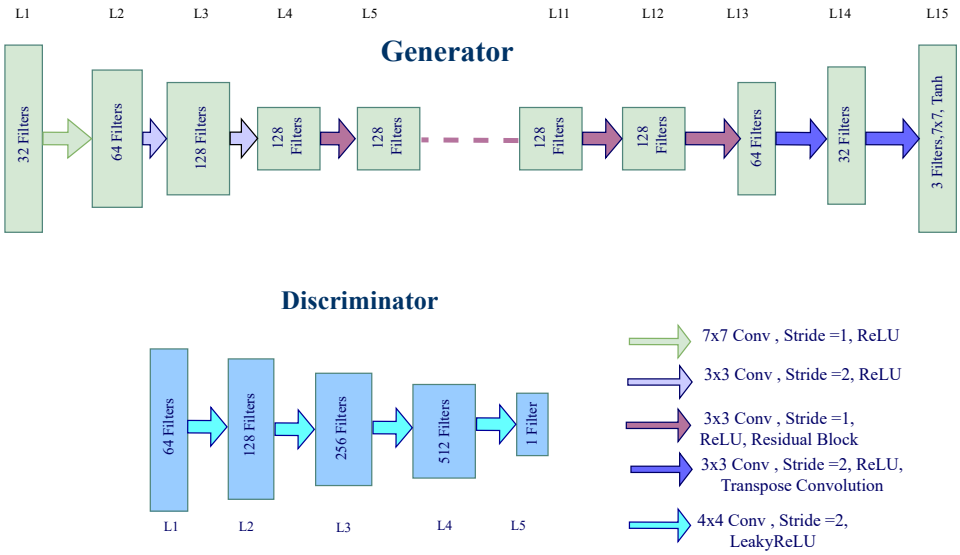


Fig. 2: Block diagram of the CycleGAN & CUT's generator and discriminator [65].

3. *Modulated Contrast-based CUT*

In the contrastive learning literature hardness of negative samples has been adequately investigated [24], [17], [23]. Hard negatives can facilitate to learn the data representation faster [24] in the contrastive learning. Moreover, in the unpaired image translation literature, the hardness of the negative patches is defined by considering their similarity to the query, as proposed in [16]. The similarity between a negative sample y_j and an anchor x_i is defined by the hard negative weighting as shown in Eq. (8):

$$w_{ij} = \frac{e^{(x_i.y_i/\beta)}}{\sum_{j=1}^{N} e^{(x_i.y_j/\beta)}}, \tag{8}$$

where β is the weighting temperature parameter. The modulate contrast NCE loss objective employs re-weighing strategies by enforcing the constraint defined by Eq. (9):

$$\sum_{i=1}^{N} w_{ij} = 1, \sum_{j=1}^{N} w_{ij} = 1; i, j \in [1, N]. \tag{9}$$

Equation (10) is the optimal transport [66] for this algorithm which also considers the Equation (9) as constraint:

$$\min_{w_{ij}, i,j \in [1,N]} [\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} w_{ij}.e^{(x_i.y_i/\tau)}], \tag{10}$$

$$\min_{T} \langle C, T \rangle \;\; s.t \;\; \langle T \overrightarrow{1} \rangle = 1, \; \langle T^T \overrightarrow{1} \rangle = 1, \tag{11}$$

where $\langle C, T \rangle$ is the inner product of the cost matrix (C) and transport plan (T). In the hard mining strategy, the cost matrix is $e^{((1-x_i.y_i)/\beta)}$ when $i \neq j$; if $i = j$ then $C_{ij} = \infty$. The Sinkhorn [67], [68] algorithm is applied to Eq. (11) for calculating the optimal transport plan. Furthermore, similar to the PatchNCE and MoNCE methods, the modulated contrast-based loss also uses multiple layers (1st, 4th, 8th, 12th, and 16th) of the encoder features for contrastive learning. The phenomena of modulated and vanilla contrast are depicted in Figure 3 and Figure 4.

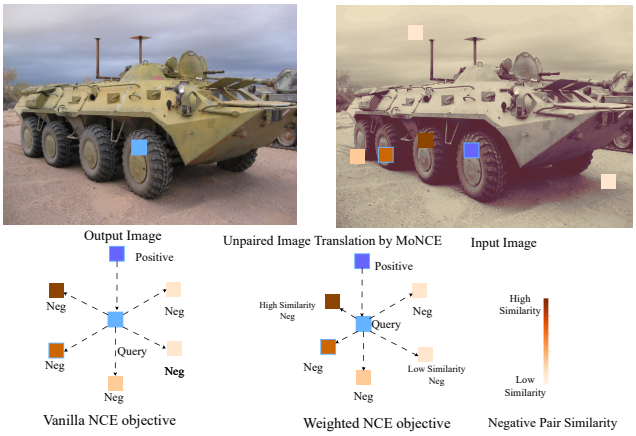


Fig. 3: Vanilla and weighted contrastive objective [16].

4. *Query Selected Attention-based CUT*

The CUT network selects queries from random locations of generated images, but this approach may not always serve the intended purpose of domain translation. Because, all the locations of the image patches do not contain similar importance of domain-related information. The AttnGAN [69] guided the generator of GAN to translate the relevant area of the images by using a learnable foregrounded mask. The F-LSeSim [18] method a variant of CUT uses self-similarity into the local region of the images and employ the NCE loss. All the afore mentioned algorithms can not properly depict the domain-relevant patches. The QS-Attn [15] overcomes this issue by using the entropy-based attention for query

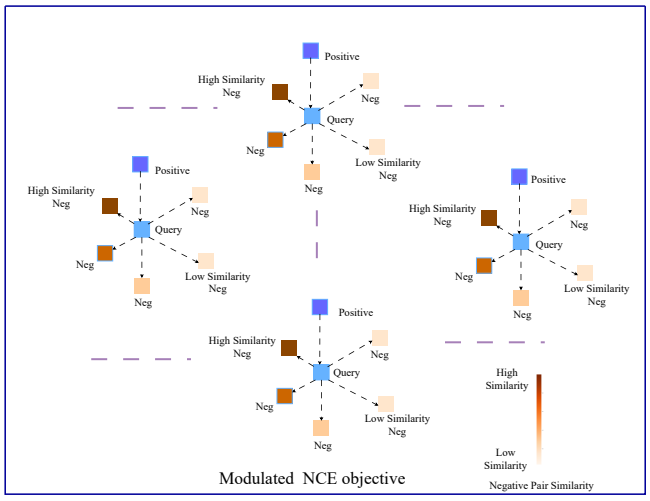


Fig. 4: Modulated contrastive objective [16].

selection and incorporates it into the CUT network. For this purpose, the QS-Attn reshape the source domain images $F_x \in \mathbb{R}^{H\times W\times C}$ into a two dimensional matrix $Q \in \mathbb{R}^{HW\times C}$ and its transpose $K \in \mathbb{R}^{C\times HW}$. The Q and K are multiplied and imposed softmax into its row then it produced attention matrix $A_g \in \mathbb{R}^{HW\times HW}$. The relevant region is selected by the entropy Eq. (12):

$$H_g(i) = -\sum_{j=1}^{HW}(A_g(i,j)\log A_g(i,j), \tag{12}$$

where i, j denote the indices of query (Q) and key (K) metrics. For selecting relevant query the entropy matrix H_g is sorted by ascending order and the smallest M row are chosen. This calculation is performed only into the features of source domain images. The final size of the global QS-Attn matrix is $A_{QS} \in \mathbb{R}^{N\times HW}$. The QS-Attn-based CUT network is illustrated in the Figure 5.

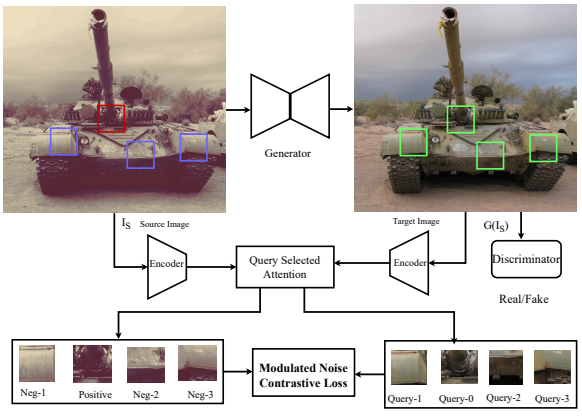


Fig. 5: The query selected attention-based CUT architecture [15].

C. Generation of Synthetic Patches in CUT Networks

The harder negative samples in the contrastive loss function is crucial for quicker and better learning [24]. Additionally, mixing these harder negative examples to generate synthetic hard negatives has been shown to improve performance and lead to faster convergence in contrastive learning [23]. Moreover, mixing can be performed at feature or pixel levels [70], [71]. To improve the generalization and robustness of deep neural networks, various techniques have been proposed, including manifold mixup [70], cutmix [72], noisy mixup [25], and puzzle mixup [73]. More recently, a new method called noisy feature mixup (NFM) [25] has been introduced, which has been found to outperform the other methods in terms of both robustness and the smoothness of the decision boundary of the neural network model [25].

1. *Noisy Feature Mixup*

NFM combines the features of DNN and injects additive and multiplicative noise to improve generalization. If noise injection is not used, NFM can be described as a form of the manifold mixup method. In this paper, we only consider the feature-level NMF and not the label mixup. The equation for NFM is presented by 13:

$$\tilde{g}_k = (\mathbb{1} + \sigma_{mult}\xi_k^{mult}) \odot M_\lambda(g_k(x), g_k(x')) + \sigma_{add}\xi_k^{add} \tag{13}$$

where $\mathrm{M}_\lambda(a,b) = \lambda * a + (1-\lambda) * b$ and $\lambda \sim Beta(\alpha,\beta)$, where ξ_k^{add} and ξ_k^{mult} are $\mathbb{R}^{d_k}$ valued independent random variables that model the additive and multiplicative noise respectively, and σ_{add}, $\sigma_{mult} \geq 0$ are the specific noise levels.

D. Source and Target Domain Classifiers

The transductive network shown in Figure 2 has a source and a target domain ATR classifier. In our experiment, the architecture of both classifiers is ResNet-18 [74]. The skip connection in the ResNet-18 architecture makes it possible to build a deeper network and helps to backpropagate loss more efficiently. In the proposed method, we optimize the cross-entropy loss ($\mathcal{L}_{CE}$) of the classifiers:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{10} y_{o,c}\log(p_{o,c}), \tag{14}$$

where $y_{o,c}$ is the true probability and $p_{o,c}$ is the predicted probability of the target chip in the training data, the class number is denoted by c.

E. Transductive CycleGAN

Our proposed transductive CycleGAN is depicted in Figure 1, consists of a CycleGAN and two residual network-based CNN classifiers. The architecture of the transductive CycleGAN generators and discriminators is

illustrated in Figure 2. The transductive CycleGAN generator consists of fractionally-strided convolutions, regular convolutions, and residual blocks.

In this experiment, we consider the MWIR and VIS domains ATR vehicles in the DSIAC datasets [35] as the source and target domain data, respectively. Moreover, in the VAIS dataset [75], the source domain is regarded as the VIS domain, whereas the target domain is considered to be the IR domain. Conversely, in the FLIR ATR dataset [76], [77], the source and target domains are represented by mid-wave and long-wave IR images, respectively. Furthermore, the source domain classifier is assumed to be well-trained (called the optimal MWIR classifier) and its weights are fixed. During the training, the source domain classifier's weights are initialized to the target domain classifier's weights which are gradually trained using the transductive transfer learning process. The proposed network's weights are obtained by minimizing the adversarial loss, cycle consistency loss, identity loss of the CycleGAN, and the categorical cross-entropy loss of the source and target domain ATR classifiers. It should be pointed out that the source domain classifier loss is also backpropagated to the transductive network, but its weights are not updated during the training of the transductive network. The total loss of the transductive CycleGAN model is $\mathcal{L}_{Total}$, given by:

$$\mathcal{L}_{Total} = \mathcal{L}_{CycleGAN} + \lambda_{CE} * \mathcal{L}_{CE-Source} + \lambda_{CE} * \mathcal{L}_{CE-Target}, \tag{15}$$

where $\mathcal{L}_{CE-Source}$ and $\mathcal{L}_{CE-Target}$ are the cross-entropy losses for the source and target domain classifiers, respectively.

F. Hybrid CUT (H-CUT) Network

Our proposed H-CUT network consists of a contrastive unpaired image-to-image (I2I) translation network [14] that optimizes the modulated NCE loss [16]. Furthermore, the H-CUT network has a query selection attention module [15] which gives more attention to the relevant anchors and Noisy Feature Mixup (NFM) module [25] for generating synthetic negative patches. The components of the H-CUT network are depicted in Figure 6. In this network, both the original and synthetic patches undergo an optimal transport plan, which modulates the hardness of negative patches. Finally, the H-CUT network has three components which are i) query selection attention, ii) noisy patch generation, and iii) MoNCE module. The proposed H-CUT network is the building block for the contrastive learning-based TTL network.

G. Transductive CUT Network

The proposed transductive CUT network consists of two H-CUT networks for bidirectional image translation between source and target domains. The reconstructed source images are used to compute the categorical-cross-entropy loss by the source domain classifier, that is backpropagated to guide the network to produce class-preserving image generation. Additionally, this TTL network constructs a target domain classifier that will annotate unlabeled target domain target chips. All components of this network are depicted in Figure 7. In the H-CUT TTL network, we will investigate the importance of bijection mapping by cycle consistency; therefore, in Figure 7, cycle consistency is included.

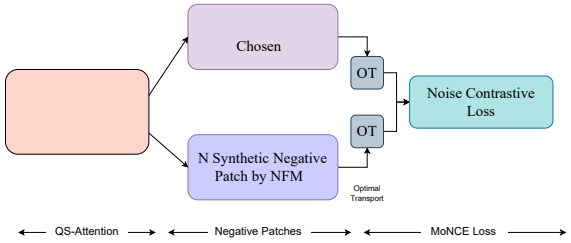


Fig. 6: Block diagram of the QS-Attn + NFM + MoNCE module in the H-CUT architecture.

$$\mathcal{L}_{TTL-CUT} = \mathcal{L}_{CUT_1} + \mathcal{L}_{CUT_2} + \lambda_c * \mathcal{L}_{cycle-consistency_1} + \lambda_c * \mathcal{L}_{cycle-consistency_2} + \lambda_{CE} * \mathcal{L}_{CE-Source} + \lambda_{CE} * \mathcal{L}_{CE-Target}, \tag{16}$$

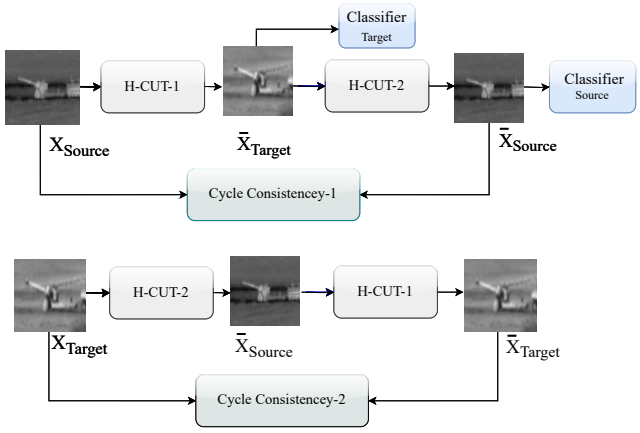


Fig. 7: Block diagram of the H-CUT based TTL network.

IV. Experiments

A. Dataset

In this section, the details of two ATR vehicle and one ship dataset's general information and data prepossessing will be discussed.

1. *DSIAC dataset*

We implement our transductive transfer learning approach using the publicly available DSIAC dataset [35]. This dataset is collected by the US Army Night Vision

and Electronic Sensors Directorate (NVESD). The DSIAC dataset contains ten vehicles in the visible and MWIR domain. Among these classes, two of them are civilian vehicles ('Pickup', 'Sport vehicle'), one is an artillery piece ('D20'), and the rest of the seven are military vehicles ('2S3', 'BTR70', 'BRDM2', 'BMP2', 'MT-LB', 'T72', 'ZSU23'). The distance between target vehicles and camera varies from 1 to 5 kilometers. The dataset contains 189 video sequences in the MWIR domain and 97 video sequences in the visible domain. Each video sequence contains 1800 video frames. Generally, the image size in the DSIAC dataset is 640x480 pixels. In this work, we detect and crop the ATR vehicles (target chips) from the DSIAC dataset using the information from the Meta-UDA [34]. All the target vehicles at different ranges are projected into a canonical distance (e.g., 2 kilometers) using bi-cubic interpolation. The final target chip size is 68x68x3.

2. Visible and Infrared Spectrums (VAIS) Dataset

The VAIS [75] dataset is the world's first publicly available maritime dataset. This dataset consists of 1,623 visible and 1,242 infrared domain images of maritime vehicles. This dataset has a total 1,088 corresponding pair images. Moreover, this dataset has six classes which are 'merchant ships', 'medium passenger ships','sailing ships', 'small boats', 'tugboats', and 'medium other ships'. The distribution of the dataset is depicted in Table I. In this experiment, we resize all the targets of the VAIS dataset similar to the DSIAC dataset.

TABLE I: Distribution of classes in the VAIS dataset.

Class Name	Sub-class	No of IR target	No of VIS target
Merchant	Barges	33	35
	Cargo	147	149
Medium passenger Ship	Ferries	40	42
	Tour boats	89	98
Sailing Ship	Sail up	159	278
	Sail down	140	134
Small Boat	Speed Boat	132	243
	Jet-skis	23	58
	Small pleasure Boat	75	82
	Large pleasure Boats	49	65
Tug Boat		99	57
Medium other ship	Fishing	28	33
	Medium other	124	152
Total		1138	1416

3. FLIR ATR Dataset

For evaluation of the proposed TTL method, we use 12-bit gray-scale MW and LW FLIR input images that are captured by an experimental laboratory infrared sensor [76], [77], [78]. Furthermore, a quantum well infrared photodetector focal plane array is used for the LW sensor. In contrast, an indium antimonide (InSb) focal plane array is employed for the MW sensor. Each input image size is 304x504 pixels. The target chips are cropped into 40x75 pixels and then reshape into 68x68x3 pixels. This dataset consists of 461 paired MW and LW images with 572 target chips. Those target chips are captured between 1 to 4 kilometers distance range with different poses. The dataset has seven classes of ATR vehicles which are 'M60A3', 'PICKUP', 'HMMWV', 'M2', 'M35', 'M113', and 'UNKNOWN'.

B. Training Details

Our first transductive network consists of a CycleGAN and two CNN-based classifiers. The learning rate of the CycleGAN network is set to 0.0002 for the first 50 epochs; after that, the learning rate is reduced to 0.0001 for the rest of the 50 epochs. For fast convergence of the network, we have initialized the weights of the CycleGAN generator weight from the weights of summer to winter translation on the Yosemite dataset [79]. The hyper-parameters of Equations 3, 4, and 5 are set to $\eta_1 = \eta_2 = 10$, $\eta_3 = \eta_4 = 5$, and $\lambda_a = \lambda_b = \lambda_c = 1$, respectively. For the initial 20 epochs of training, we set $\lambda_{CE} = 0.5$, after that it becomes 2.5 in Eq. (15). To optimize the weights of the generators and discriminators, we used the Adam optimizer [80] with $\beta 1 = 0.5$ and $\beta 2 = 0.999$. To stabilize the GAN training, we update the discriminator loss five times less than the generator loss. For training the classifiers, we also used the Adam optimizer with a learning rate of 0.0005 and $\beta 1 = 0.9$, $\beta 2 = 0.999$ for 40 epochs. We finetune the target domain classifier for 10 epochs to get the result from 1% and 10% labeled data as shown in Table II. The other variants of TTL hyper-parameter are the same as the transductive CycleGAN. In Eq. (6), we set $\lambda_X = \lambda_Y = 1$ for the PatchNCE and MoNCE loss calculation. This experiment was conducted on an NVIDIA RTX-8000 GPU using the PyTorch framework [81] with a batch size of 160.

C. Ablation Study

In this section, we will investigate the effect of each module in the TTL network. Therefore, we compare the performance of different variants of the TTL network. The performance comparison is extensively conducted using the DSIAC dataset.

The effect of Cycle Consistency. In the CUT-TTL method, there are two CUT networks for two-way translation. We eliminate the cycle consistency loss and analyze the annotation performance in the DSIAC dataset. After that, we added cycle consistency along with PatchNCE and MoNCE, then trained the whole TTL network. The confusion matrices of contrastive TTL without cycle consistency are illustrated in Figure 9a, Figure 9b, with cycle consistency illustrated in Figure 9c and Figure 9d. The performance of the TTL network is increased substantially after adding cycle consistency loss. In the MoNCE-based TTL, cycle consistency increases the average accuracy by

TABLE II: Performance of different variants of the TTL network on the DSIAC dataset.

Method Name	Accuracy (%) (No label data)	Accuracy (1% label data)	Accuracy (10% label data)
CycleGAN TTL	71.56	80.24	94.86
QS-Attn+PatchNCE TTL	64.87	72.18	87.00
PatchNCE+cycle-consistency TTL	65.23	81.78	92.87
QS-Attn+MoNCE+cycle-consistency TTL	76.22	88.40	94.40
QS-Attn+MoNCE+NMF TTL	70.09	84.74	93.98
QS-Attn+MoNCE+cycle-consistency+NMF TTL	76.61	88.49	97.13
SimCLR [3]	—	75.69	95.15
BYOL [82]	—	78.62	94.43
SwAV [83]	—	75.42	96.02
B-Twins [84]	—	75.50	94.17
100% labeled data			
ResNet-18 (supervised)	99.28		

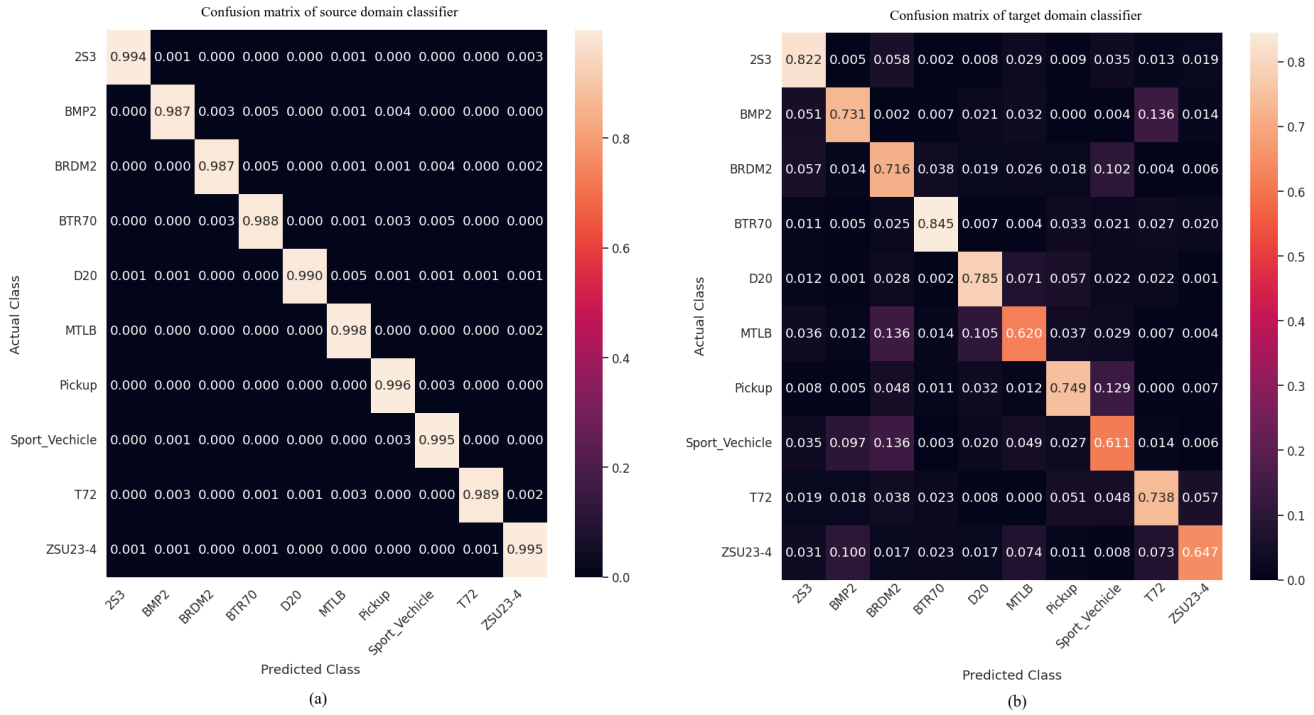


Fig. 8: Confusion matrices of (a) source domain classifier and (b) target domain classifier (no labeled data) using the transductive CycleGAN network.

TABLE III: Performance of the TTL network on the VAIS dataset.

Method Name	Average Accuracy (%)
No labeled data	
CycleGAN TTL	53.60
QS-Attn+MoNCE+cycle-consistency TTL	55.41
QS-Attn+MoNCE+cycle-consistency+NMF TTL	54.50
100% labeled data	
ResNet-18 (supervised)	86.04

TABLE IV: Performance of the TTL network on the FLIR ATR dataset.

Method Name	Average Accuracy (%)
No labeled data	
CycleGAN TTL	78.95
QS-Attn+MoNCE+cycle-consistency TTL	80.00
QS-Attn+MoNCE+cycle-consistency+NMF TTL	78.95
100% labeled data	
ResNet-18 (supervised)	93.68

TABLE V: FID score of different variants of the TTL network on the DSIAC dataset.

TTL Method Name	FID score
CycleGAN	216.109,
QS-Attn+PatchNCE	209.372
PatchNCE+cycle-consistency	190.316
QS-Attn+MoNCE+NMF	182.686
QS-Attn+PatchNCE+cycle-consistency	171.485
QS-Attn+MoNCE+cycle-consistency	94.989
QS-Attn+MoNCE+cycle-consistency+NMF	87.957

6.52% for the DSIAC dataset. Therefore, in the CUT-TTL network, the cycle consistency is an essential component.

The effect of MoNCE. In contrastive TTL, we also investigate the benefits of the MoNCE loss over PatchNCE loss. Using the DSIAC dataset, when we implemented MoNCE loss by replacing PatchNCE loss, that imporved the annotation performance. These results are depicted in Table II. It is because MoNCE is able to transfer domain more accurately compared to the PatchNCE-based CUT.

The effect of QS-Attention. The QS-Attention is intentionally used in the CUT network for query selection to perform a more accurate domain translation. In the transductive CUT network, we investigated necessity of QS-Attention in the DSIAC dataset and investigate the FID score improvement compared to no query selection. The QS-Attn block improves the FID score to 18.831 with the PatchNCE and cycle consistency based TTL as shown in Table V. This indicates that the query selection improves the image translation capabilities when used in the TTL network.

The effect of Synthetic Patches. We insert synthetic patches in TTL network and observe the performance improvement. The performance of NFM based synthetic patches, and no synthetic patch are depicted in the Table II, Table III, and Table IV. From there, we observe that negative patches improve the TTL network performance by a small amount (0.39%) for the DSIAC dataset; however, it decreases the performance slightly for the VAIS and FLIR ATR datasets. Inserting the NMF module in the TTL network, the classification performance of the TTL network improves in four classes and degrades in the other six classes for the DSIAC dataset, that scenario is illustrated in the confusion matrices of Figure 9c and Figure 9d. Moreover, Table V shows that the NMF module of the TTL network improves the FID[85] score ($94.989 \rightarrow 87.957$) for the DSIAC dataset.

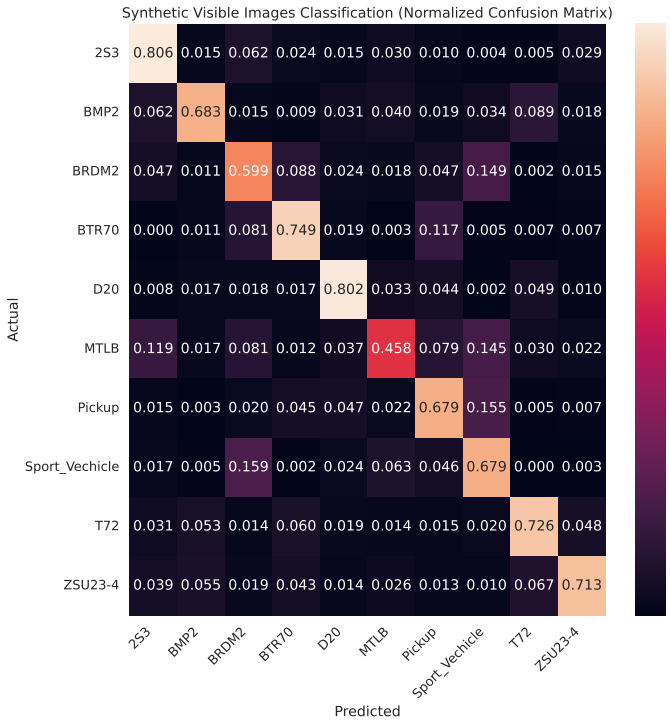
D. Numerical Evaluation

We evaluate the performance of the source and target domain classifiers using the DSIAC dataset. The training, testing, and validation datasets are constructed by randomly dividing the DSIAC dataset into 70:15:15 ratio, respectively. The confusion matrix of the source domain classifier is illustrated in Figure 8(a). The confusion matrix provides the normalized performance of the classifier. The average accuracy of the source domain classifier is 99.16%. The performance of the source domain classifier with targets at different distances is investigated in Table VI. From this table, we can conclude that the classification performance of the source domain classifier is almost consistent when capturing images at distances between one kilometer to four kilometers. However, the source classifier performance is degraded when the target capturing distance is beyond four kilometers. Because the long-distance target chips are low in quality as compared to short-distance ones.

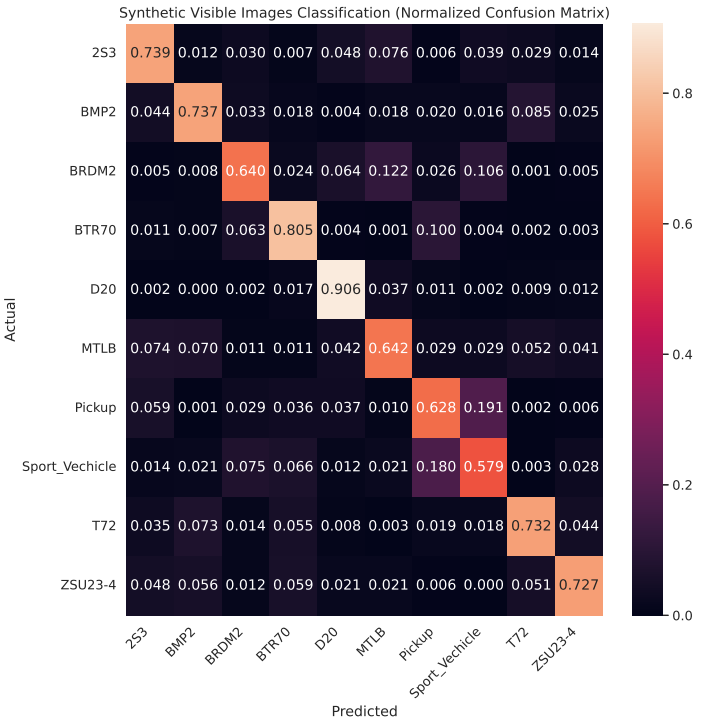
TABLE VI: Performance of the source domain classifier at different target distances.

Target Distance (m)	Accuracy(%)	Number of Samples
1000	99.47	4,885
1500	99.72	4,713
2000	99.78	4,895
2500	99.54	4,830
3000	99.45	4,558
3500	99.76	4,193
4000	99.11	4,506
4500	97.83	4,196
5000	93.55	4,106

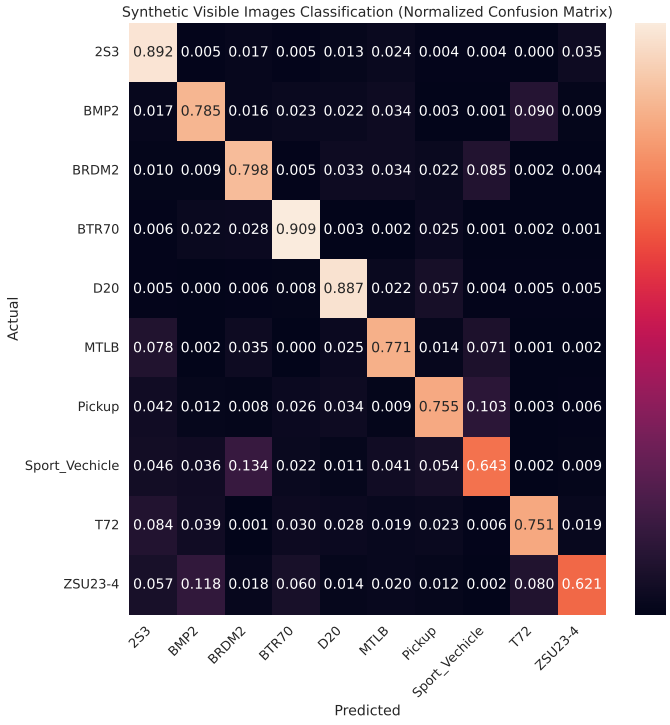
We construct a target domain classifier in this experiment using the transductive CycleGAN network. The confusion matrix of the target domain classifier is depicted in Figure 8(b). The average accuracy of the target domain classifier is 71.6% for the visible domain images in the DSIAC dataset. The confusion matrix of the target domain classifier illustrates that the classification performances of the 'MTLB', 'Sport Vehicle', and 'ZSU23-4' are lower than other classes. Moreover, we investigate the target domain classifier's performance with a fraction of the labeled data. The accuracy of the target domain classifier is 80.24%, 90.79%, and 94.86%, with 1%, 5%, and 10% of the labeled target domain dataset, respectively. The performance of the visible domain classifier with 1% and 10% labeled data are depicted in Table II. The synthetic images generated by the transductive CycleGAN are also illustrated in Figure 11. The classification performance of different variants of the CUT-based TTL with no labeled and partially labeled is depicted in Table II. This table provides the need for the modulating contrast and query selection in the CUT network along with the cycle-consistency in the TTL network. Furthermore, the table depicts that the classification performance of the best TTL method is 76.61% that is around 22% less than the supervised CNN classifier in the DSIAC dataset. The proposed TTL network and state of the arts semi-supervised networks performance comparison is conducted using the DSIAC dataset. In the semi-supervised setting, the Table II depicted that proposed TTL network able to annotate more accurately than SimCLR [3], BYOL [82], SwAV [83], and B-Twins [84] networks.



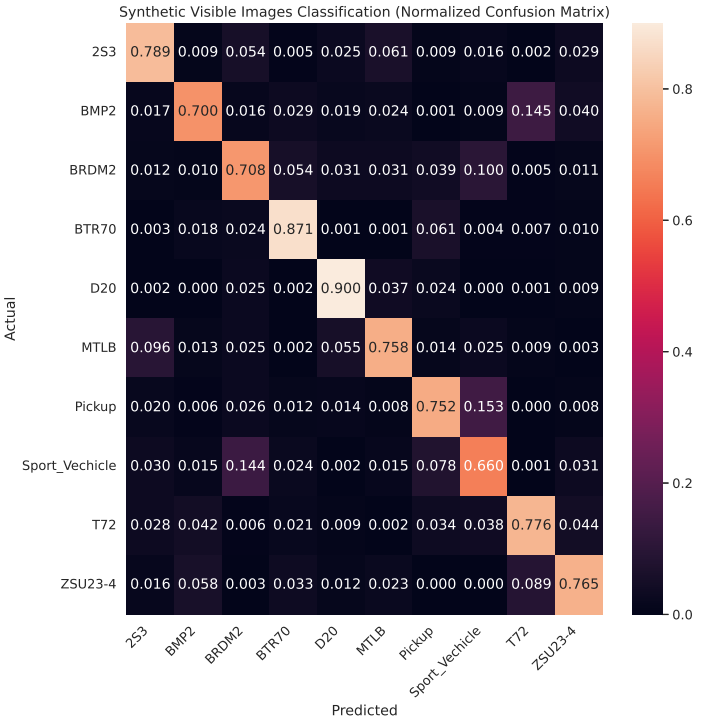
(a) Confusion Matrix (QS-Attn+PatchNCE based TTL netwrok (no cycle consistency)).



(b) Confusion Matrix (QS-Attn+MoNCE+NFM based TTL network (no cycle consistency)).

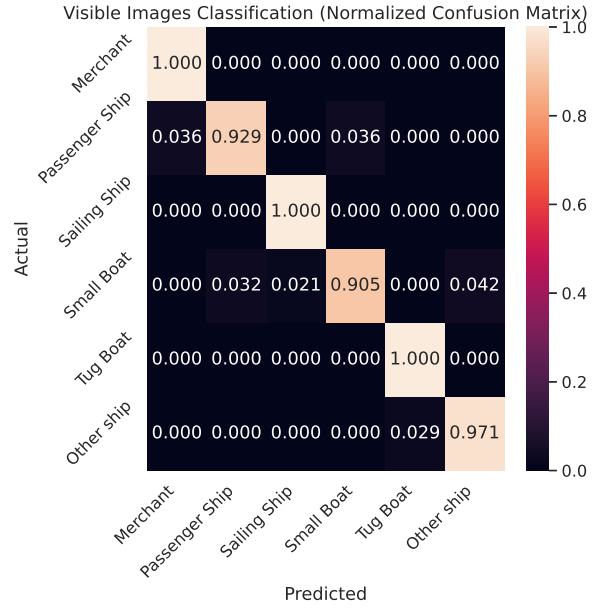


(c) Confusion Matrix (QS-Attn+MoNCE+cycle consistency based TTL network).

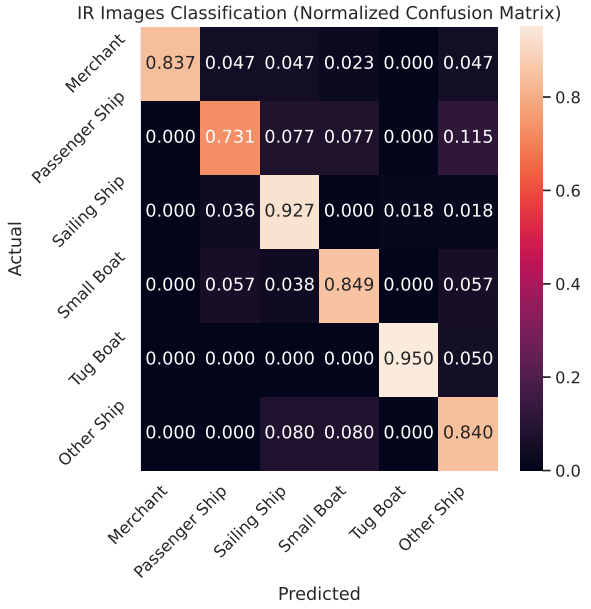


(d) Confusion Matrix (QS-Attn+MoNCE+NFM+cycle consistency based TTL network).

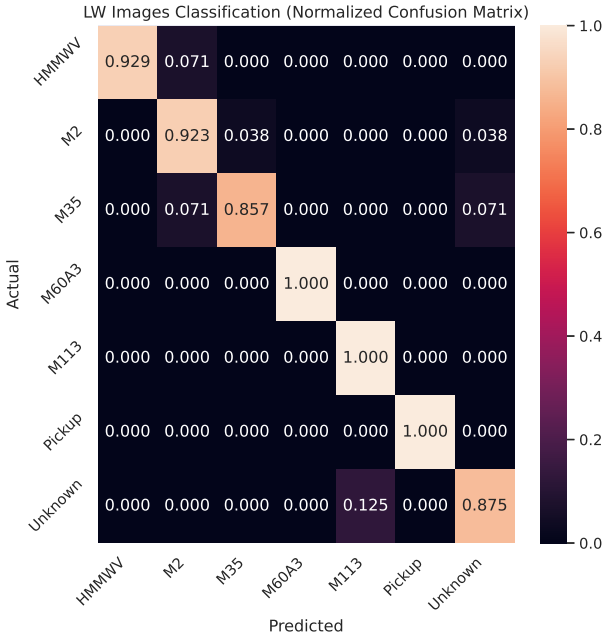
Fig. 9: Confusion matrices of the target domain classifiers by the TTL netwrok using the DSIAC dataset (no label data)).



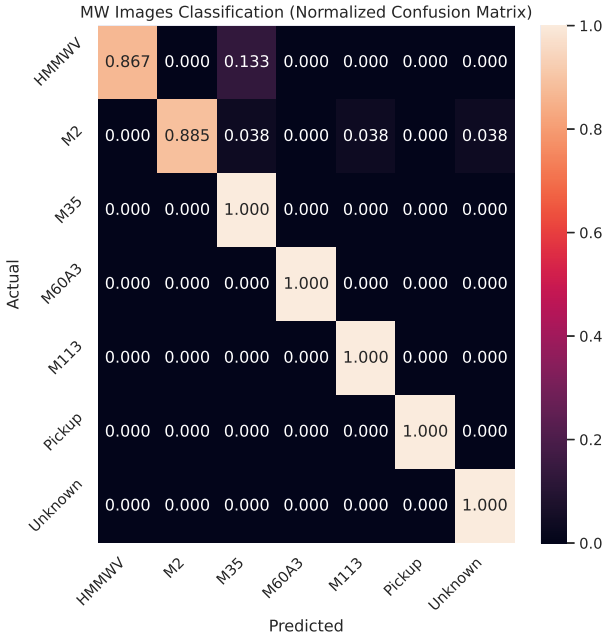
(a) VIS domain classifier using VAIS dataset (100% labeled data).



(b) IR domain classifier using VAIS dataset (100% labeled data).



(c) LW domain classifier using FLIR ATR dataset (100% labeled data).



(d) MW domain classifier using FLIR ATR dataset (100% labeled data).

Fig. 10: Confusion matrices of the source and target domain classifiers using the VAIS and FLIR ATR datasets.

Moreover, we investigate the classification performance of visible (source) and infrared (target) images for the VAIS ship target dataset [75]. This dataset is divided into train, test, and validation using 70:20:10 ratio, respectively. The confusion matrices of the supervised VIS and IR classifiers for the VAIS dataset are depicted in Figure 10a and Figure 10b, respectively. The average accuracy of visible and IR domain classifiers is 96.73% and 86.04%, respectively in supervised settings. Table III provides that the different variants of the TTL network can successfully annotate target domain images in this dataset. The table elicits that the QS-Attn-MoNCE-cycle-consistency-based TTL is the best performer to annotate IR images in the VAIS dataset. Moreover, the noisy feature mixup module degrades the annotation performance by a small amount in this dataset.

Also, we analyze mid-wave (source domain) and long-wave (target domain) ATR target classification performance for FLIR ATR dataset[76]. The train, test and validation ratio is the same as the VAIS dataset. The source

and target domain confusion matrices are illustrated in Figure 10c and Figure 10d. The classification performance of source and target domain classifier is 95.95%, and 93.68%, respectively, in the supervised setting. The proposed semi-supervised TTL method can annotate LW targets at 80.00% accuracy. Integrating the NMF module into CUT-TTL network degrades the classification performance by 1.05% in this dataset. The performance of different variants of the TTL network and supervised classifier in the target domain is depicted in Table IV. This table illustrates that the combination of QS-Attn, MoNCE, and cycle consistency-based TTL network performs the best for the FLIR ATR dataset.

E. Visual Evaluation

The original and synthetic ATR vehicle images of the DSIAC dataset are presented in Figure 11. From there, we can deduce that QS-Attn-MoNCE-cycle and QS-Attn-MoNCE-NMF-cycle TTL produced high-quality synthetic images compared to without i) query selection ii) cycle consistency, and iii) modulated contrast NCE-based TTL. Furthermore, the Fréchet Inception Distance (FID) [85] score of different variants of the TTL network by changing unpaired I2I network are shown in Table V. This table depicts that the combination of MoNCE, QS-Attn, noisy feature mixup, and cycle consistency-based TTL provides the best FID score (87.96) and highest quality synthetic images among the other variants of TTL using the DSIAC dataset.

F. Discussion

The classification performances of different variants of the TTL network using DSIAC, FLIR ATR datasets are better than the VAIS dataset. It is maybe that the VAIS dataset is noisy, and an information gap stays between visible and infrared domain.

V. Conclusion

This paper proposes a CycleGAN-based and a contrastive learning-based transductive transfer learning network to annotate unlabeled ATR images. The modulated contrastive learning with cycle consistency-based TTL outperforms the CycleGAN-based TTL. We have also investigated the usefulness of integrating the MoNCE loss, QS-attention, cycle consistency, and noisy feature mixup modules in the proposed TTL network. Aggregating these modules is essential for better performance of the TTL network. Proposed TTL networks can successfully annotate the unlabeled images in the DSIAC, VAIS, and FLIR ATR datasets. Our proposed method outperforms the state-of-the-art semi-supervised learning algorithms for the annotation performance in partial label data. Finally, using our proposed TTL network it overcomes the laborious and cumbersome manual annotation task.

ACKNOWLEDGMENT

This material is based upon work supported in part by the U. S. Army Research Laboratory and the U. S. Army Research Office under contract number: W911NF2210117.

REFERENCES

[1] N. M. Nasrabadi, "DeepTarget: An automatic target recognition using deep convolutional neural networks," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 6, pp. 2687–2697, 2019.

[2] P. Y. Simard, D. Steinkraus, J. C. Platt *et al.*, "Best practices for convolutional neural networks applied to visual document analysis." in *Icdar*, vol. 3, no. 2003. Edinburgh, 2003.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[4] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.

[5] J. Chen, L. Du, H. He, and Y. Guo, "Convolutional factor analysis model with application to radar automatic target recognition," *Pattern Recognition*, vol. 87, pp. 140–156, 2019.

[6] K. El-Darymli, E. W. Gill, P. Mcguire, D. Power, and C. Moloney, "Automatic target recognition in synthetic aperture radar imagery: A state-of-the-art review," *IEEE access*, vol. 4, pp. 6014–6058, 2016.

[7] J. M. Topple and J. A. Fawcett, "MiNet: Efficient deep learning automatic target recognition for small autonomous vehicles," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 6, pp. 1014–1018, 2020.

[8] S. Dang, Z. Cao, Z. Cui, Y. Pi, and N. Liu, "Open set incremental learning for automatic target recognition," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4445–4456, 2019.

[9] W. Shi, Y. Gong, C. Ding, Z. M. Tao, and N. Zheng, "Transductive semi-supervised deep learning using min-max features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 299–315.

[10] M. Rohrbach, S. Ebert, and B. Schiele, "Transfer learning in a transductive setting," *Advances in neural information processing systems*, vol. 26, 2013.

[11] O. Sener, H. O. Song, A. Saxena, and S. Savarese, "Unsupervised transductive domain adaptation," *arXiv preprint arXiv:1602.03534*, 2016.

[12] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Transactions on Multimedia*, 2021.

[13] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[14] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *European conference on computer vision*. Springer, 2020, pp. 319–345.

[15] X. Hu, X. Zhou, Q. Huang, Z. Shi, L. Sun, and Q. Li, "QS-Attn: Query-selected attention for contrastive learning in I2I translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 291–18 300.

[16] F. Zhan, J. Zhang, Y. Yu, R. Wu, and S. Lu, "Modulated contrast for versatile image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 280–18 290.

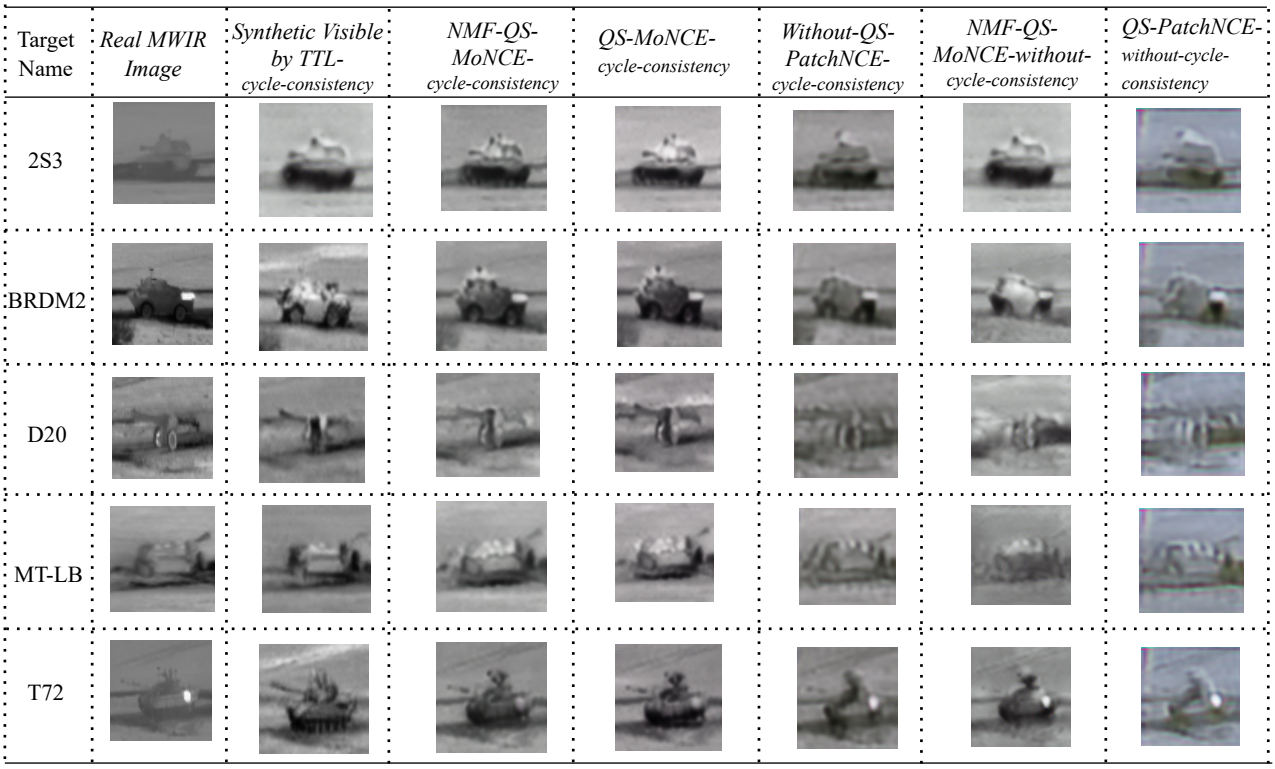


Fig. 11: Generated images by the proposed TTL framework using the DSIAC dataset.

[17] W. Wang, W. Zhou, J. Bao, D. Chen, and H. Li, "Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 020–14 029.

[18] C. Zheng, T.-J. Cham, and J. Cai, "The spatially-correlative loss for various image translation tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 407–16 417.

[19] J. Han, M. Shoeiby, L. Petersson, and M. A. Armin, "Dual contrastive learning for unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 746–755.

[20] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[22] T. Zhang, L. Li, S. Cao, T. Pu, and Z. Peng, "Attention-guided pyramid context networks for detecting infrared small target under complex background," *IEEE Transactions on Aerospace and Electronic Systems*, 2023.

[23] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 798–21 809, 2020.

[24] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," *arXiv preprint arXiv:2010.04592*, 2020.

[25] S. H. Lim, N. B. Erichson, F. Utrera, W. Xu, and M. W. Mahoney, "Noisy feature mixup," *arXiv preprint arXiv:2110.02180*, 2021.

[26] S. M. Sami, N. M. Nasrabadi, and R. Rao, "Deep transductive transfer learning for automatic target recognition," in *Automatic Target Recognition XXXIII*. SPIE, 2023.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[28] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[30] C. Zheng, X. Jiang, and X. Liu, "Semi-supervised SAR ATR via multi-discriminator generative adversarial network," *IEEE Sensors Journal*, vol. 19, no. 17, pp. 7525–7533, 2019.

[31] F. Zhang, C. Hu, Q. Yin, W. Li, H. Li, and W. Hong, "SAR target recognition using the multi-aspect-aware bidirectional LSTM recurrent neural networks," *arXiv preprint arXiv:1707.09875*, 2017.

[32] S. Deng, L. Du, C. Li, J. Ding, and H. Liu, "SAR automatic target recognition based on Euclidean distance restricted autoencoder," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 7, pp. 3323–3333, 2017.

[33] V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Sparsity-motivated automatic target recognition," *Applied optics*, vol. 50, no. 10, pp. 1425–1433, 2011.

[34] V. Vs, D. Poster, S. You, S. Hu, and V. M. Patel, "Meta-UDA: Unsupervised domain adaptive thermal object detection using meta-learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1412–1423.

[35] Us army night vision and electronic sensors directorate (NVESD). [Online]. Available: https://dsiac.org/databases/atr-algorithm-development-image-database/

[36] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE*

Geoscience and remote sensing letters, vol. 13, no. 3, pp. 364–368, 2016.

[37] J. Wang, T. Zheng, P. Lei, and X. Bai, "Ground target classification in noisy SAR images using convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 11, pp. 4180–4192, 2018.

[38] L. Wang, X. Bai, R. Xue, and F. Zhou, "Few-shot SAR automatic target recognition based on Conv-BiLSTM prototypical network," *Neurocomputing*, vol. 443, pp. 235–246, 2021.

[39] L. Wang, X. Bai, C. Gong, and F. Zhou, "Hybrid inference network for few-shot SAR automatic target recognition," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 11, pp. 9257–9269, 2021.

[40] K. Fu, T. Zhang, Y. Zhang, Z. Wang, and X. Sun, "Few-shot SAR target classification via metalearning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[41] R. M. Marcacini, R. G. Rossi, I. P. Matsuno, and S. O. Rezende, "Cross-domain aspect extraction for sentiment analysis: A transductive learning approach," *Decis. Support Syst.*, vol. 114, pp. 70–80, Oct. 2018.

[42] Y. He, J. Yuan, and L. Li, "Enhancing RNN based OCR by transductive transfer learning from text to images," *AAAI*, vol. 32, no. 1, Apr. 2018.

[43] Y. Zong, W. Zheng, X. Huang, K. Yan, J. Yan, and T. Zhang, "Emotion recognition in the wild via sparse transductive transfer linear discriminant analysis," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 163–172, Jun. 2016.

[44] K. Yan, W. Zheng, T. Zhang, Y. Zong, C. Tang, C. Lu, and Z. Cui, "Cross-domain facial expression recognition based on transductive deep transfer learning," *IEEE Access*, vol. 7, pp. 108 906–108 915, 2019.

[45] N. Sameera and M. Shashi, "Deep transductive transfer learning framework for zero-day attack detection," *ICT Express*, vol. 6, no. 4, pp. 361–367, Dec. 2020.

[46] J. Kobylarz, J. J. Bird, D. R. Faria, E. P. Ribeiro, and A. Ekárt, "Thumbs up, thumbs down: non-verbal human-robot interaction through real-time EMG classification via inductive and supervised transductive transfer learning," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 12, pp. 6021–6031, Dec. 2020.

[47] Z. Deng, P. Xu, L. Xie, K.-S. Choi, and S. Wang, "Transductive Joint-Knowledge-Transfer TSK FS for recognition of epileptic EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 8, pp. 1481–1494, Aug. 2018.

[48] C. Yang, Z. Deng, K.-S. Choi, and S. Wang, "Takagi–Sugeno–Kang transfer learning fuzzy logic system for the adaptive recognition of epileptic electroencephalogram signals," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 5, pp. 1079–1094, Oct. 2016.

[49] W. Fu, B. Xue, M. Zhang, and X. Gao, "Transductive transfer learning in genetic programming for document classification," in *Simulated Evolution and Learning*. Springer International Publishing, 2017, pp. 556–568.

[50] A. Moreo, A. Esuli, and F. Sebastiani, "Lost in transduction: Transductive transfer learning in text classification," *ACM Trans. Knowl. Discov. Data*, vol. 16, no. 1, pp. 1–21, Jul. 2021.

[51] Y. Luo, Z. Zhang, L. Zhang, J. Han, J. Cao, and J. Zhang, "Developing high-resolution crop maps for major crops in the european union based on transductive transfer learning and limited ground data," *Remote Sensing*, vol. 14, p. 1809, Apr. 2022.

[52] L. V. Utkin and M. A. Ryabinin, "A deep forest for transductive transfer learning by using a consensus measure," ser. Communications in computer and information science. Springer International Publishing, 2018, pp. 194–208.

[53] H. Hoyez, C. Schockaert, J. Rambach, B. Mirbach, and D. Stricker, "Unsupervised image-to-image translation: A review," *Sensors*, vol. 22, no. 21, p. 8540, 2022.

[54] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189.

[55] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, and M.-H. Yang, "DRIT++: Diverse image-to-image translation via disentangled representations," *International Journal of Computer Vision*, vol. 128, no. 10, pp. 2402–2417, 2020.

[56] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in neural information processing systems*, vol. 30, 2017.

[57] S. Benaim and L. Wolf, "One-sided unsupervised domain mapping," *Advances in neural information processing systems*, vol. 30, 2017.

[58] M. Amodio and S. Krishnaswamy, "TraVeLGAN: Image-to-image translation by transformation vector learning," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2019, pp. 8983–8992.

[59] H. Fu, M. Gong, C. Wang, K. Batmanghelich, K. Zhang, and D. Tao, "Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2427–2436.

[60] H. Sahak, D. Watson, C. Saharia, and D. Fleet, "Denoising diffusion probabilistic models for robust image super-resolution in the wild," *arXiv preprint arXiv:2302.07864*, 2023.

[61] M. Zhao, F. Bao, C. Li, and J. Zhu, "EGSDE: unpaired image-to-image translation via energy-guided stochastic differential equations," *arXiv preprint arXiv:2207.06635*, 2022.

[62] M. Özbey, S. U. Dar, H. A. Bedel, O. Dalmaz, Ş. Özturk, A. Güngör, and T. Çukur, "Unsupervised medical image translation with adversarial diffusion models," *arXiv preprint arXiv:2207.08208*, 2022.

[63] H. Sasaki, C. G. Willcocks, and T. P. Breckon, "UNIT-DDPM: Unpaired image translation with denoising diffusion probabilistic models," *arXiv preprint arXiv:2104.05358*, 2021.

[64] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[65] M. B. Ali, I. Y.-H. Gu, M. S. Berger, J. Pallud, D. Southwell, G. Widhalm, A. Roux, T. G. Vecchio, and A. S. Jakola, "Domain mapping and deep learning from multiple MRI clinical datasets for prediction of molecular subtypes in low grade gliomas," *Brain sciences*, vol. 10, no. 7, p. 463, 2020.

[66] G. Peyré, M. Cuturi *et al.*, "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[67] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, 2013.

[68] J. Altschuler, J. Niles-Weed, and P. Rigollet, "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration," *Advances in neural information processing systems*, vol. 30, 2017.

[69] X. Chen, C. Xu, X. Yang, and D. Tao, "Attention-GAN for object transfiguration in wild images," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 164–180.

[70] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *International conference on machine learning*. PMLR, 2019, pp. 6438–6447.

[71] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, A. Solin, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," *Neural Networks*, vol. 145, pp. 90–106, 2022.

[72] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[73] J.-H. Kim, W. Choo, and H. O. Song, "Puzzle mix: Exploiting saliency and local statistics for optimal mixup," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5275–5285.

[74] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[75] M. M. Zhang, J. Choi, K. Daniilidis, M. T. Wolf, and C. Kanan, "VAIS: A dataset for recognizing maritime imagery in the visible and infrared spectrums," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 10–16.
[76] A. Mehmood and N. M. Nasrabadi, "Anomaly detection for longwave FLIR imagery using kernel Wavelet-RX," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 1385–1388.
[77] ——, "Wavelet-RX anomaly detection for dual-band forward-looking infrared imagery," *Applied optics*, vol. 49, no. 24, pp. 4621–4632, 2010.
[78] ——, "Kernel Wavelet-Reed–Xiaoli: an anomaly detection for forward-looking infrared imagery," *Applied optics*, vol. 50, no. 17, pp. 2744–2751, 2011.
[79] Cyclegan summer-winter image translation[pytorch]. [Online]. Available: http://www.kaggle.com/code/balraj98/cyclegan-summer-winter-image-translation-nvesd/data/
[80] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[81] A. e. a. Paszke, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.
[82] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.
[83] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.
[84] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 310–12 320.
[85] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

Shoaib Meraj Sami received the B.Sc. and the M.Sc. degree in Electrical and Electronic Engineering degree from the Bangladesh University of Engineering and Technology, Bangladesh, in 2016 and 2021, respectively. He also worked as an engineer at Nuclear Power Plant Company Bangladesh Limited and Chittagong Port Authority. He is currently pursuing the Ph.D. degree with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV, USA. His main focus is on the development of machine learning and deep learning algorithms and their applications with computer vision, biometrics, and automatic target recognition.

Nasser M. Nasrabadi (Fellow, IEEE) received the B.Sc. (Eng.) and Ph.D. degrees in electrical engineering from the Imperial College of Science and Technology, University of London, London, U.K., in 1980 and 1984, respectively. In 1984, he was at IBM, U.K., as a Senior Programmer. From 1985 to 1986, he was at the Philips Research Laboratory, New York, NY, USA, as a member of the Technical Staff. From 1986 to 1991, he was an Assistant Professor at the Department of Electrical Engineering, Worcester Polytechnic Institute, Worcester, MA, USA. From 1991 to 1996, he was an Associate Professor at the Department of Electrical and Computer Engineering, State University of New York at Buffalo, Buffalo, NY, USA. From 1996 to 2015, he was a Senior Research Scientist at the U.S. Army Research Laboratory. Since 2015, he has been a Professor with the Lane Department of Computer Science and Electrical Engineering. His current research interests include image processing, computer vision, biometrics, statistical machine learning theory, sparsity, robotics, neural networks, and image processing. He is a fellow of the International Society for Optical Engineers (SPIE). He has served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Raghuveer M. Rao (Fellow, IEEE) received the M.E. degree in electrical communication engineering from the Indian Institute of Science, Malleswaram, Bangalore, India, and the Ph.D. degree in electrical engineering from the University of Connecticut, Storrs,CT, USA, in 1981 and 1984, respectively. He was a Member of Technical Staff at AMD Inc., Santa Clara, CA, USA, from 1985 to 1987. He joined the Rochester Institute of Technology, Rochester, NY, USA, in December 1987, where, at the time of leaving in 2008, he was a Professor of electrical engineering and imaging science. He is currently the Chief of the Intelligent Perception Branch, U.S. DEVCOM Army Research Laboratory (ARL), Adelphi, MD, USA, where he manages researchers and programs in computer vision and scene understanding. He has held visiting appointments at the Indian Institute of Science, the Air Force Research Laboratory, Dahlgren, VA, USA, the Naval Surface Warfare Center, Rome, NY, USA, and Princeton University, Princeton, NJ, USA. His research contributions cover multiple areas, such as signal processing with higher order statistics, wavelet transforms and scale-invariant systems, statistical self-similarity, and their applications to modeling in communication and image texture synthesis. His recent work is focused on machine learning methods for scene understanding on maneuvering platforms. Dr. Rao is an elected fellow of the International Society for Optical Engineers (SPIE). He is an ABET program evaluator for electrical engineering and has served on the Technical Committee for the IEEE Signal Processing Society and SPIE. He has also served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, and the Journal of Electronic Imaging.