electronics

MDPI

Article

A Triple Adversary Network Driven by Hybrid High-Order Attention for Domain Adaptation

Meng Wang * and Jiawei Fu

The Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650093, China; Fu912718544@tom.com

* Correspondence: wangmeng@kust.edu.cn.com

Received: 7 November 2020; Accepted: 9 December 2020; Published: 11 December 2020



Abstract: How to bridge the knowledge gap between the annotated source domain and the unlabeled target domain is a basic challenge to domain adaptation. The existing approaches can relieve this gap by feature alignments across domains; however, aligning non-transferable features may lead to negative shift confusing the knowledge learning on target domains. In this paper, a triple adversary network is proposed on the basis of a high-order attention, hopefully to solve the problem. The proposed architecture focuses on the detailed feature alignment by a hybrid high-order attention using a fast iteration algorithm. In addition, an orthogonal loss of two complementary modules is applied to constrain the mutual exclusion of foreground and background features. Finally, a triple adversarial strategy is introduced to further improve the training convergence for the composed architectures. Numeric experiments on datasets of Digits, Office-31 and Office-home illuminate that the proposed network can effectively improve the state-of-art domain adaptations with superior transferring performance.

Keywords: domain adaptation; high-order attention; triple adversarial strategy; orthogonal loss

1. Introduction

Supervised learning has achieved great success in many applications by utilizing fully annotated training data, such as image recognition [1–3], speech recognition [4,5], etc. In these scenarios, a practical difficulty is the manual collection of huge amounts of training data and annotations. To solve this problem, existing solutions usually resort to the rich knowledge of the easily labeled datasets, namely, the source domain, to promote effective adaption learning in those domains with scarce labels, namely, the target domains, also known as domain adaptation (DA). Generally, DA can be categorized into supervised adaptation and unsupervised adaption. The former assumes that a small amount of labeled target data can be collected for training [6–8], and the latter focuses on the cases involving fully unlabeled target examples [3,9–11]. Though the latter case is more common and significant progress has been made on it in recent years [12,13], challenges still remain in several specific aspects.

The unsupervised domain adaptation (UDA) specifically focuses on how to reduce the domain shift between the fully labeled source domain and the unlabeled target domain, also known as domain inconsistency, which is caused by many practical factors, such as the vision variations of captured angles, lighting quality, and image resolutions varying in different scenes [14,15]. To improve it, deep domain confusion (DDC) [13] has been studied to represent the domain invariant knowledge by introducing a adaption layer and a maximum mean difference (MMD)-based domain confusion loss. In addition, the deep adaptation network (DAN) [12] integrates task-specific layers into the kernel Hilbert space to enhance the transfer ability between domain representations; and geodesic flow kernel (GFK) [16] uses Kullback–Leibler (KL) divergence to estimate domain differences, and a limited

number of subspaces are integrated to discover new representations. However, since the transferring maps of domain representations are generally too complex to be efficiently obtained, especially in these deep models, various models integrating adversarial strategies have been widely tried to align the unknown distributions of these data domains.

In recent years, adversarial training strategies have been introduced to generate domain invariant features, having greatly improved the performance of unsupervised domain adaptation (UDA) models [7]. The studies of the domain adversarial neural network (DANN) [17] suggest that the features suitable for domain transferring should be discriminative and domain invariant; thus, domain classifications are integrated at the tops of the basic representation layers to guide the generations of domain invariant features. The adversarial discriminative domain adaptation (ADDA) [18] firstly learns the representation of the source domain, and then maps the target data to the same space in light of a domain adversarial loss. The conditional domain adversarial networks (CDAN) [19] extend the adversarial transferring to discriminative information conveyed in the classifier outputs to enable finer-grained alignment of multi-modal structures. In the cycle-consistent conditional adversarial transfer networks (3CANT) model [20], two feature translators and one corresponding cycle-consistent loss are integrated into conditional adversarial networks for domain alignment. Although the adversarial domain adaptation (ADA) has been significantly improved, the detailed properties on different image regions are not considered. Obviously, different regions of images cannot be transferred equally. Some regions, such as the background, can be aligned across domains between the feature spaces, but may not contribute much to the discriminative domain information. In addition, some images that are significantly different across domains in the feature space should not be forcibly aligned across domains, or this negative transfer might easily involve irrelevant knowledge. For these reasons, some scholars have suggested to introduce an attention mechanism to adversarial adaptation, such as self-attention generative adversarial networks (SAGAN) [21] using self-attention in generative adversarial networks to build remote and multi-level dependency across image regions. In transferable attention for domain adaptation (TADA) [22], multiple region-level domain discriminators are applied to formulate transferable local attention, and a single image-level domain discriminator is adopted to obtain transferable global attention to emphasize transferable images. Although TADA considers the difference in the transfer of regional features to efficiently align them across domains, these approaches yielding spatial and channel attention mainly utilize the discrimination masks based on low-order feature distribution, so they are only efficient in capturing rough spatial representations.

Although the above approaches have greatly improved the transfer performance of UDA, three issues still need further studying: (1) how to formulate more effectively adversarial strategies to improve the local parameter convergence of complex adversarial transfer networks; (2) presently, most of the attention applied to transferring networks belongs to low-level representations; thus, its ability to capture locally salient details is limited; (3) as for the adversarial domain adaption, the compulsory transferring of background regions might confuse the salient spatial representations of foregrounds, leading to negative transfer.

From the perspective of distribution matching, the UDA approaches often align features based on maximum mean difference (MMD) [13] and correlation comparison (CORAL) [9]. They are designed to match the first-order (mean) and second-order (covariance) statistics of different distributions, but for real-world applications such as image recognition, these deep features are generally complex non-Gaussian distributions [23,24], unable to be fully characterized by the first or second-order statistics. Inspired by it, this paper focuses on the solution of using high-order statistics for domain feature matching. Since high-order statistics can approximate more complex non-Gaussian distributions, the attention based on high-order moment is expected to achieve comprehensive domain alignment. The main contributions can be summarized as follows:

(1) A hybrid high-order triple-adversary network (HTAN) is proposed to achieve the detailed feature alignments of domain adaption, gradually focusing on both spatial and channel features to approximate high-order representation distributions.

(2) The proposed architecture is further driven by reverse phase modules and a orthogonal loss to constrain the positive interaction of foreground and background features, thereby ultimately solving the domain inconsistency in the feature transfers.
(3) Moreover, a triple-player-based adversarial learning strategy is introduced into the proposed network to improve the iterative convergence of complex network parameters. Numerical experimental results have verified that on the datasets including MNIST [25], USPS [26], SVHN [27], Office-31 [1] and Office-home [28], the proposed network has superior performances compared with other state-of-art benchmarks.

The remaining contents of this paper are arranged as follows. Section 2 reviews the related work, including adversarial domain adaptation and attention mechanism. The proposed domain adversarial architecture is detailed in Section 3. In the next section are the experimental settings and the comparison results and discussions. The last section summaries this paper.

2. Results

2.1. Unsupervised Domain Adversarial Adaptation

Recent research methods on domain adaptation mainly focus on the following aspects. One is to utilize metric methods to measure the shift across domains, such as maximum mean discrepancy (MMD) [10,13], second-order statistics correlation alignment CORAL [3,9] and center moment discrepancy (CMD) [29]. Wasserstein distance-based discriminator is adopted in these methods to bring the two distributions closer. These methods explicitly minimize domain distribution discrepancy to exploit transferable domain features. According to the above formulas, the existing domain adversarial networks [17,18] have been verified to achieve excellent performances on the transferring scenes where the distributions of the source and the target domain are complex and short of prior knowledge. Generally, the adversarial approaches can solve the domain-shift difficulties by globally matching example features across these domains. However not all spatial representations should be transferred for domain adaptation, and the negative transfers might arise because of the confused knowledge transferred to the target domains.

Adversarial learning is another way to convey domain information. Specifically, the domain-adversarial neural network (DANN) [17] first leverages adversarial learning between the domain classifier and feature generator to learn domain-invariant representations by adding a simple gradient reversal layer (GRL). Further, to address the mode collapse issue, multi-adversarial domain adaptation (MADA) [30] presents a multi-adversarial domain adaptation approach with the help of multiple domain classifiers. The adversarial discriminative domain adaptation (ADDA) [18] uses label learning in the source domain to distinguish representations, and then uses asymmetric mapping (without weight sharing) learned by standard generative adversarial network (GAN) loss to map the target data onto a separate code in the same space. Designed in cycle-consistent adversarial domain adaptation (CyCADA) [31], cyclic consistency loss strengthens the consistency of structure and semantics during adversarial domain adaptation (ADA). MADA [30] captures multi-modal structures to achieve fine-grained alignment of different data distributions based on multiple domain identifiers. Co-regularized domain alignment (Co-DA) [32] constructs a number of different feature spaces, and aligns the source and target distributions in each feature space. In contrast, compared with the existing DA method based on adversarial learning, the key improvement of the proposed model is the ability to add hybrid high-order attention to jointly capture the knowledge of complex non-Gaussian distribution fine features and distinguish structure. This helps to achieve satisfactory performance when the field gap is large.

2.2. Attention Mechanism in Deep Architectures

Recently, the attention mechanism has made significant progress in various tasks, such as speech recognition [33] and domain adaptation [22]. It can be divided into two categories: spatial attention and

channel attention. For the first category, the size of feature maps is $C \times H \times W$. It can be considered as a $H \times W$ image and the representation of every pixel is C dimensional. The attention model can learn to re-weigh every pixel. This model is used in [34] to refine spatial attention. For the second category, more attention is paid to each channel of a feature map, viewed as a process of selecting semantic attributes. Its most famous application is SeNet [35] which is the foundation of the imageNet large scale visual recognition challenge's (ILSVRC) top classification submission in 2017. The convolutional block attention module (CBAM) [36] combines spatial attention with channel attention to independently perfect convolution features. Although there are few studies on attention adaptation, it is worth noting that in [22], the domain-adapted transferable attention TADA was proposed. It focuses on two complementary, transferable local and global attentions. There are few studies on the combination of channel attention and spatial attention in the field of confrontation adaptation. Therefore, in our study, channel attention and spatial attention are combined in the field of confrontation adaptation to study their learning of different levels of feature information from different perspectives.

2.3. High-Order Statistics for Spatial Representations

In the study of features based on deep learning statistics, statistics above the first order [37–39] can successfully be used to represent the significant details of an image. In the field of image and video recognition particularly, the second-order convolutional neural networks (SO-CNNs) [40] extract the covariance second-order statistics matrix from the convolution activation to construct the covariance descriptor unit's second-order convolutional neural network (CNN). Bilinear CNN (B-CNN) [39] calculates the outer products of the convolutional description vector and combines them to obtain the image descriptor. On the surface, second-order statistics, covariance and Gaussian descriptors show better performances than those of descriptors using zero-order or first-order statistics. For example, when the feature distribution is non-Gaussian, the use of the second-order or lower statistical information may not be enough [39]. Therefore, many researchers have turned to the exploration of the higher-order information. For example, the video recognition task [41] combines an effective dot product attention mechanism with temporal reasoning to dynamically discover high-order object interactions. The convolution neural network (CNN) is fully parameterized by the high-order tensor [42] to jointly capture the complete structure of the neural network. These high-order statistical representations can capture more discriminative information than the first-order and obtain promising improvements. Therefore, this paper attempts to combine high-order statistics with spatial attention for the first time to continue to explore high-order moment tensors for comprehensive domain alignment research.

3. The Proposed Domain Adaption Architecture

In this section, a hybrid architecture is presented to achieve detailed feature transfers for the domain adaption. Given n_s labeled examples from a source domain $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ and n_t unlabeled examples from a target domain $\mathcal{D}_t = \{(x_j^t)\}_{j=1}^{n_t}$, $P(x_s, y_s)$ and $Q(x_t, y_t)$ are joint distributions of the source and target domains respectively. The i.i.d. assumption is violated, as $P \neq Q$ by assuming that the two domains have an identical number of categories. This paper aims to formulate a hybrid adversarial transferring architecture, able to be pre-trained on $\mathcal{D}_s$ and then generalized well to $\mathcal{D}_t$. The overall procedure is illuminated in Figure 1.

At present, adversarial domain adaptation [17,19] has been verified as one of the basic transfer schemes to align the representations on the two domains following different probability distributions. According to these solutions, the features generated through insufficient training might still deceive the domain discriminator, implying that the intrinsic mechanism of feature transferring and discriminating cannot be fully understood by the adversarial learning. Thus, attention-driven modules have been integrated into generative adversarial frameworks, as in self-attention generative adversarial networks (SAGANs) [21], which are helpful to the long-range and multi-level dependencies of modeling across image regions. On this basis, an extended attention mechanism can be formulated by mixed channel

and spatial masks with high-order representations in order to perform precisely spatial alignments, and meanwhile, a triple-game adversarial strategy based on spatial orthogonal losses is integrated to enhance the super-parameter convergence during training.

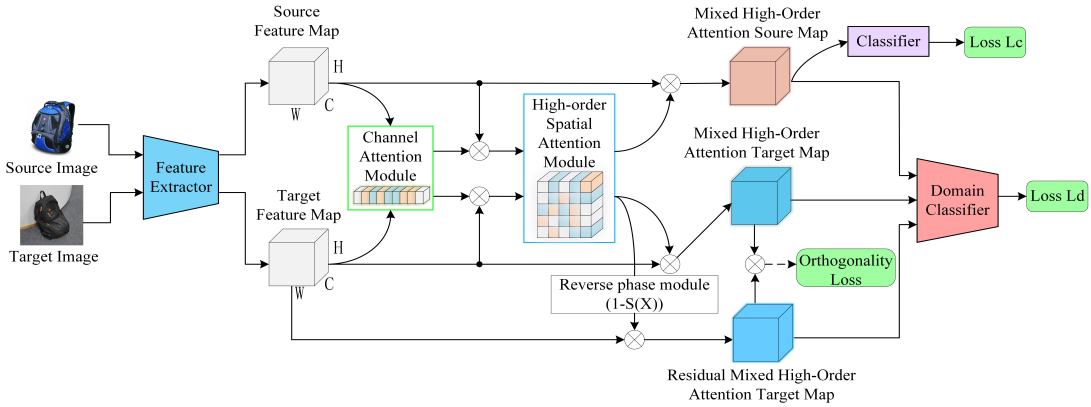


Figure 1. The proposed hybrid high-order triple-adversary network (HTAN) consists of six modules, including feature extractor G, label predictor C, domain discriminator D, complementary module $\bar{G}$ with orthogonal loss L_{orth}, channel attention module F_c and high-order spatial attention module F_s.

3.1. The Mixed High-Order Attention for Feature Alignments

Given a feature map $F \in \mathbb{R}^{C\times H\times W}$ on intermediate layers as the input of transferring modules, the module of mixed attention sequentially infers a 1D channel attention map $M_c \in \mathbb{R}^{C\times 1\times 1}$ and a 2D spatial attention map $M_s \in \mathbb{R}^{1\times H\times W}$ as

$$\begin{cases} F_c = F \otimes M_c(F) \\ F^* = F \otimes M_s(F_c) \end{cases} \tag{1}$$

where F^* stands for the final output of the mixed channel and spatial attention, F_c for the output of the channel attention and $\otimes$ for the convolutional operation. To exploit the inter-channel relationship of features, the spatial dimension of the feature F is firstly aggregated by two different spatial descriptors $AvgPool(F)$ and $MaxPool(F)$ respectively. Both the descriptors are merged using element-wise summation, and then forwarded to a shared layer as multi-layer perceptron (MLP) with sigmoid activation to generate the channel attention mask M_c. The channel attention map can be denoted as

$$M_c(F) = Sigmoid(MLP(AvgPool(F) + MaxPool(F))) \tag{2}$$

In terms of the perception of feature maps, the channel attention is applied globally, while the spatial attention often works locally. However, since these spatial masks can only be represented by the low-order statistics, which are inefficient for accurately capturing the spatial semantic details, a high-order spatial attention $M_s(F) \in \mathbb{R}^{1\times H\times W}$ and detailed high-order statistics are adopted for the feature alignments. Firstly, a linear polynomial predictor is defined on top of the high-order statistics of $f \in \mathbb{R}^C$, denoting a local descriptor at a specific spatial location of f as

$$m(f) = \sum_{r=1}^{R} \langle w^r, \otimes_r f \rangle \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner-product of two equally-sized tensors, R the number of order, $\otimes_r x$ the r-th order outer-product of f that comprises each the degree-r monomial in f and w^r the r-th order tensor to be learned that contains the weights of degree-r variable combinations in f. Suppose that

when $r > 1$, w_r can be approximated by D_r rank-1 tensors by tensor decomposition; then Equation (3) can be expanded as

$$\begin{aligned} m(f) &= \langle w^1, f\rangle + \sum_{r=2}^{R} \langle w^r, \otimes_r f\rangle \\ &= \langle w^1, f\rangle + \sum_{r=2}^{R} \langle \sum_{d=1}^{D^r} \alpha^{r,d} u_1^{r,d} \otimes \cdots \otimes u_r^{r,d}, \otimes_r f\rangle \end{aligned} \tag{4}$$

where $u_1^{r,d} \in \mathbb{R}^C, ..., u_r^{r,d} \in \mathbb{R}^C$ represent vectors, $\otimes$ is the outer-product and $\alpha^{r,d}$ is the weight for d-th rank-1 tensor. Then, according to the tensor algebra, the above formula is reformulated as

$$\begin{aligned} m(f) &= \langle w^1, f\rangle + \sum_{r=2}^{R} \sum_{d=1}^{D^r} \alpha^{r,d} \prod_{s=1}^{r} \langle u_s^{r,d}, f\rangle \\ &= \langle w^1, f\rangle + \sum_{r=2}^{R} \sum_{d=1}^{D^r} \alpha^{r,d} z^{r,d} \\ &= \langle w^1, f\rangle + \sum_{r=2}^{R} \langle \alpha^r, \mathrm{z}^r\rangle \end{aligned} \tag{5}$$

where $\alpha^r = [\alpha^{r,1}, \cdots, \alpha^{r,D^r}]^T$ is the weight vector, $\mathrm{z}^r = [z^{r,1}, \cdots, z^{r,D^r}]^T$ with $z^{r,d} = \prod_{s=1}^{r} \langle u_s^{r,d}, f\rangle$.

For convenience in the calculation of high-order statistics, let $D^r = D$, where $r = 1, 2, ..., R$, and then z^r can be obtained with a fast iteration equation

$$\begin{aligned} \mathrm{z}^r &= \mathrm{z}_1^r \odot \mathrm{z}_2^r \odot \cdots \odot \mathrm{z}_r^r \\ &= \mathrm{z}^{r-1} \odot \mathrm{z}_r^r; \quad r = 2, 3, ..., R \end{aligned} \tag{6}$$

indicating that z_r^r should be multiplied by the previous order to obtain the current order statistics. Then, according to this fast iteration, Equation (5) can be rewritten as

$$\begin{aligned} m(f) &= \langle w^1, f\rangle + \sum_{r=2}^{R} \langle \alpha^r, \mathrm{z}^{r-1} \odot \mathrm{z}_r^r\rangle \\ &= \langle w^1, f\rangle + \sum_{r=2}^{R} \langle \alpha^r, \mathrm{z}^{r-1} \odot \langle \mathrm{u}_r^r, f\rangle\rangle \end{aligned} \tag{7}$$

The above equation contains two terms, so for clarity, it is formulated into a more general case. Suppose that w^1 can be approximated by the multiplication of two matrixes $v^1 \in \mathbb{R}^{1\times D^1}$ and $\alpha^1 \in \mathbb{R}^{D^1\times 1}$, i.e., $w^1 = v^1 \times \alpha^1$; then an overall equation is obtained as

$$\langle w^1, f\rangle + \sum_{r=2}^{R} \langle \alpha^r, \mathrm{z}^r\rangle = \sum_{r=1}^{R} \langle \alpha^r, \mathrm{z}^r\rangle \tag{8}$$

In Equation (7), since $m(f)$ is capable of modeling and using the high-order statistics of the local descriptor f, the high-order attention mask can be obtained by performing Sigmoid function Sigmoid function on Equation (7) as

$$M_s(F) = Sigmoid([m(f_{ij})]); f_{ij} \in F \tag{9}$$

where $M_s(f_{ij}) \in \mathbb{R}^C$ and the value of each element in $M_s(f_{ij})$ is within the interval [0, 1]. For the high-order spatial attention module shown in Figure 2. To this end, the spatial attention mechanism is modeled by combining complex high-order statistics to capture more complex and advanced information for precise parts, so that the feature extractor generates more high-level information transferability to distinguish fine features.

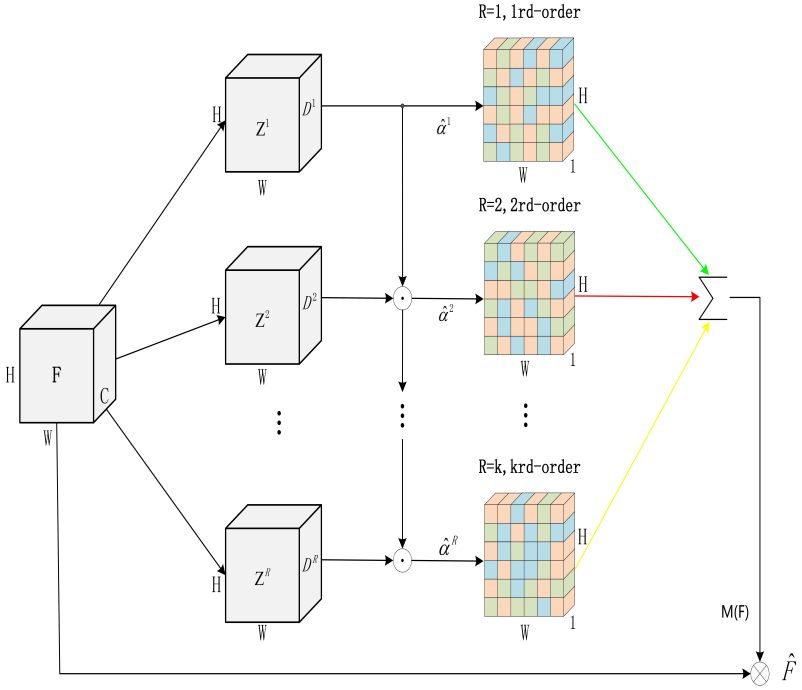


Figure 2. Illustration of the high-order spatial attention module.

$M(f) \in \mathbb{R}^C$ and the value of each element in $M(f)$ is in the interval [0, 1]. To this end, we are committed to modeling the spatial attention mechanism by combining complex high-order statistics in order to capture more complex and advanced information between precise parts. Furthermore, the adopted order statistics have the time complexity $O(R)$ as a faster calculation model, compared with $O(\frac{R^2}{2})$ for the previous version [23], so that the feature extractor generates high-level transferable information and distinguishes fine features more efficiently.

3.2. The Orthogonal Loss in Reverse Phase Modules

The previous studies [43] indicate that orthogonality helps the optimization of deep neural networks (DNN) by preventing explosion or vanishing of back-propagated gradients. Rodrguez et al. [44] proposed an orthogonal regulation to enforce feature orthogonality locally based on cosine similarities of filters. Jia et al. [45] proposed the algorithms of orthogonal deep neural networks (OrthDNNs) to meet the recent interest in spectrally regularized deep learning methods. A cheap orthogonal constraint was proposed based on parameterizations from exponential maps [46]. To overcome the redundancy in improving feature diversity, orthogonality regularization [47] was proposed. Presently, few studies focus on the effectiveness of background discrimination in unsupervised confrontation adaptive tasks. Thus, in our research, the proposed attention-based adaption framework is adopted to further integrate the accuracy influences of background and foreground on the overall migration effect.

To obtain accurate spatial features of foregrounds in vision scenes and eliminate the backgrounds or other interfering factors, a pair of complementary modules with orthogonal loss are integrated. The hybrid high-order attention focuses more on the important target features of the image, while the complementary attention generated by the complementary module can capture the background and other factors. It is hoped that the features of the two are orthogonal in the image space, so that the overlapped features are approximate to zero. Thus, an orthogonal loss is applied to constraining the size of the two spatially overlapped features as follows.

$$L_{\text{orth}} = \sum_{i=1}^{n_t} < G(x_i), \bar{G}(x_i) >, \quad x_i \in \mathcal{D}^t \tag{10}$$

It is used to calculate the sum of outer-products between the hybrid high-order attention $G(x_i) = F(x_i) \otimes M_s$ and complementary attention $\bar{G}(x_i) = F(x_i) \otimes (1 - M_s)$ to constrain insignificant background features and other factors in the images. This idea can further help the model enhance the cross-domain transfers with distinguishing fine spatial features, and solve the drawback that the domain classifier may still be deceived in extracting insignificant features.

3.3. The Training Based on Triple Adversarial Strategy

According to the adversarial-based domain adaption, the distribution shift between the source and the target domains is reduced by generating globally transferable features [17,30]. The existing adversarial strategy can be regarded as a two-player game, where the first player is the domain discriminator D, who distinguishes the source domain from the target domain, and the second player is the feature generator G simultaneously trained to confuse the discrimination results of D. To obtain domain-invariant features F, the trainable parameters θ_g, θ_d and θ_c are optimized by minimizing the losses of the three modules G, D and C alternately. Then, the objective of domain adversarial network [17] can be denoted by

$$V_0(\theta_g, \theta_c, \theta_d) = \sum_{x_i \in \mathcal{D}_s} L_y(C(G(x_i)), y_i) - \alpha \sum_{x_i \in (\mathcal{D}_s \cup \mathcal{D}_t)} L_d(D(G(x_i)), d_i) \tag{11}$$

where the losses L_y and L_d can be assigned as cross-entropy, and α is a trade-off coefficient between the two objectives formulating the feature generation during training. After training convergence, parameters $\hat{\theta}_g$, $\hat{\theta}_d$ and $\hat{\theta}_c$ will deliver a saddle point of Equation (11). However, the challenge lies in that local convergence of the model often arises even after the two players have got training balance, particularly when the extended attention modules with more trainable parameters are integrated in this game.

In this section, an expanded adversarial procedure involving three games is proposed to relieve the local convergence. The extended formula can be illuminated as

$$\begin{aligned} V_0^*(\theta_f, \theta_c, \theta_d) = &\sum_{x_i \in \mathcal{D}_s} L_y(C(G(x_i)), y_i) - \alpha \sum_{x_i \in (\mathcal{D}_s \cup \mathcal{D}_t)} L_d(D(G(x_i)), d_i) \\ &+ \beta \sum_{x_i \in \mathcal{D}_t} L_d(D(\bar{G}(x_i)), d_i) \end{aligned} \tag{12}$$

where the modules $G/\bar{G}$, D and C are the three given players formulating the feature alignments during training, and α and β are the trade-off coefficients. After training convergence, the parameters $\hat{\theta}_g$, $\hat{\theta}_d$ and $\hat{\theta}_c$ will deliver a saddle point of Equation (12) alternately by adversarial iterations as follows:

$$\begin{cases} \hat{\theta}_g, \hat{\theta}_c = \arg\min\limits_{\theta_g, \theta_c} V_0^*(\theta_g, \theta_d, \theta_c) \\ \hat{\theta}_d = \arg\max\limits_{\theta_d} V_0^*(\theta_g, \theta_d, \theta_c) \end{cases} \tag{13}$$

According to the above equations, the training with triple adversarial strategy can be detailed as following stages. Firstly, the adversarial game is applied to G, D and C. The features from the source domain are fed into two branches, namely, the binary domain discriminator D and the label classifier C to predict input labels in a supervised manner. The features generated from the target domain are only fed into the domain discriminator D to promote an adversarial training across the domains. The parameters of these modules are alternately frozen to update the other module in the adversarial iterations. The second adversarial game is performed between D and $\bar{G}$. The feature extractor G trained in the first game focuses on the foregrounds, while the backgrounds are captured by the complementary module $\bar{G}$ as a game player to accelerate the convergence of the feature generation.

To ensure that the discriminator receives transferable information to divide the domain samples, the hybrid high-order attention is also updated by this stage to generate accuracy spatial alignments.

Finally, the updating of both G and $\bar{G}$ should be achieved by a supplementary game. These two complementary modules respectively align the foreground and the background details to support the subsequent domain discrimination. Thus, the discrimination modules D and C are frozen to accelerate the feature generations with salient spatial details in this game. In short, the role of this stage can be considered as both the cooperation and competition between the two complementary modules. In the triple adversarial training, this three-player game can learn more transferable and distinguishable fine spatial representations, and prevent the module parameters from falling into local convergence in an end-to-end optimization of complex architectures.

4. Results and Analysis

Labeled source images and unlabeled target images were used for training, and then tests were conducted on the remaining data. The proposed hybrid high-order triple-adversary network (HTAN) was evaluated with state-of-the-art approaches on three standard unsupervised domain adaptation datasets: Digits [27], Office-31 [1] and Office-home [28].

4.1. Experimental Settings

The comparisons are among the domain adaptation methods, followed by evaluations comparing the proposed HTAN and the latest domain adaptive methods, including the traditional shallow migration method (TCA) [48]; GFK [16]; and the difference-based methods—DDC [13], DAN [12], RTN [11], JAN [10], D-CORAL [9] and JDDA [49]. Finally, comparisons are made among the confrontation-based methods, DANN [17], SE [50], MADA [30], CyCADA [31], GTA [51], ADDA [18], CDAN [19], CAN [52], BSP [53], 3CATN [20] and TADA [22].

Specifically, three digits datasets are investigatedm including MNIST [25], USPS [26] and SVHN [27]. These datasets each contain digits of 10 classes ranging from 0 to 9. In particular, MNIST and USPS contain 28×28 and 16×16 gray images respectively, and SVHN consists of 32×32 color images, which might contain more than one digit in each image. The evaluation protocol with four transfer tasks is adopted as USPS to MNIST, MNIST to USPS and SVHN to MNIST. Office-31 is the most widely used dataset for visual domain adaptation, with 4652 images and 31 categories collected from three distinct domains: Amazon (A), Webcam (W) and DSLR (D). All methods are evaluated on six transfer tasks A $\rightarrow$ W, D $\rightarrow$ W, W $\rightarrow$ D, A $\rightarrow$ D, D $\rightarrow$ A and W $\rightarrow$ A. Office-Home is a better organized but more difficult dataset than Office-31, for it consists of 15,500 images with 65 object classes in office and home settings, forming four extremely dissimilar domains: Art (Ar) with 2427 paintings, sketches or artistic depiction images; Clipart (Cl) with 4365 images; Product (Pr) containing 4439 images; and Real-World (Rw) with 4357 regularly captured images. All 12 transfer tasks are performed on this dataset: Ar $\rightarrow$ Cl, Ar $\rightarrow$ Pr, Ar $\rightarrow$ Rw, Cl $\rightarrow$ Ar, Cl $\rightarrow$ Pr, Cl $\rightarrow$ Rw, Pr $\rightarrow$ Ar, Pr $\rightarrow$ Cl, Pr $\rightarrow$ Rw, Rw $\rightarrow$ Ar, Rw $\rightarrow$ Cl, Rw $\rightarrow$ Pr. Figure 3 shows some sample images of the three data sets.

For Office-31 and Office-Home datasets, deep adaptation methods are implemented on the basis of the PyTorch framework of residual neural networks (ResNet-50) [54], and the ResNet-50 model is also used on the ImageNet [55] dataset. For pre-training, back-propagation is applied to fine-tune the model with labeled training domain samples and completely unlabeled test domain samples. The average grade is used to assess the statistical significance of different methods. To train the model, the batch size of all experiments is set as 64. In order to optimize the network, the learning rate is set as 0.001 and the momentum as 0.9 mini-batch stochastic gradient descent (SGD). To guarantee fair comparison, all of the methods are re-implemented, and each method is trained five times with the average taken as the final result. In the comparisons, the standard protocol for unsupervised domain adaptation is observed, with labeled source data and unlabeled target data applied to all migration tasks, such as [26].



Figure 3. Sample images from three experimental datasets are illustrated. From the top line to the bottom line, these partial pictures were formulated on the original sets of MNIST, SVHN, Office-31 and Office-home.

In order to distinguish the transferable hybrid high-order attention and the transferable complementary orthogonal attention module, and the individual contributions of the three-player game training method, HTAN is used to represent the hybrid high-order attention used to counter the adaptive DANN model, and HTAN-k to represent the k-th order hybrid high-order attention model. Lor + (HTAN-k) is used to represent the k-th order hybrid high-order attention and complementary orthogonal attention. The use of Lor + Tri + (HTAN-k) means that in the combination of the k-th order of mixed high-order attention with complementary orthogonal attention, the three-player game training method is used for the entire module.

To guarantee fair comparison, all of the methods are re-implemented, and each method is trained five times with the average taken as the final result.

4.2. Results on Dataset Digits

According to our experimental scheme, in MNIST $\rightarrow$ USPS and USPS $\rightarrow$ MNIST experiments, the size of USPS is adjusted to 28 $\times$ 28 pixels, and MNIST adopts the original size of 28 $\times$ 28 pixels. In the SVHN $\rightarrow$ MNIST experiment, the SVHN dataset contains images with colored background, multiple numbers and extremely fuzzy numbers. MNIST is composed of binary black and white handwritten digital images, meaning significant domain difference between the two datasets. As the image size of MNIST is much smaller than that of SVHN, the size of MNIST used is 32 $\times$ 32 of SVHN with three channels. The proposed method shows competitive performances in all three migration tasks. The results are recorded in Table 1.

In the experiment, the performances of our method on MNIST $\rightarrow$ USPS, USPS $\rightarrow$ MNIST and SVHN $\rightarrow$ MNIST were significantly better than those of the basic model DANN, by 14.9%, 9% and 19.7%, respectively. The results show that our method outperforms the latest method 3CATN. The accuracy improved by 0.9%, 0.4% and 1.1% on MNIST $\rightarrow$ USPS, USPS $\rightarrow$ MNIST and SVHN $\rightarrow$ MNIST respectively. The average increase in accuracy, compared with the state-of-the-art 3CATN method, reached 1.2%. The HTAN model extracts the fine features of key targets in the foreground of the image, using orthogonal loss to restrict each other to eliminate insignificant background, and uses the three to confront. The training method further learns more refined feature representations that can be transferred and distinguished.

Table 1. Digits Adaptation. Four datasets' benchmarks in domain adaptation tasks. The result of each is represented by accuracy(%).

Methods	MNIST→USPS	USPS→MNIST	SVHN→MNITS	Average
DAN [12]	79.12	75.8	71.9	75.6
DDC [13]	84.8	72.3	71.1	76.06
D-CORAL [9]	89.3	91.5	59.6	80.13
DANN [17]	82.1	89.7	73.9	81.9
ADDA [18]	89.4	90.1	76.0	85.16
GTA [51]	95.3	90.8	92.4	92.83
CyCADA [31]	95.6	96.5	90.4	94.16
CDAN [19]	95.6	98.0	89.2	94.26
BSP+CDAN [53]	95.0	98.1	92.1	95.06
3CATN [20]	96.1	98.3	92.5	95.63
HTAN	91.2	93.9	88.2	91.1
HTAN-6	94.7	96.8	91.4	94.3
Lor + (HTAN-6)	96.0	97.8	92.5	95.43
Lor + Tri + (HTAN-6)	97.0	98.7	93.6	96.43

From the comparison results, some conclusions can be drawn. First of all, from the perspective of migration tasks between digital datasets, DANN, 3CTAN and Lor + Tri + (HTAN-6) based on confrontation have better performances than those of DDC, Dan and d-coral based on difference. Secondly, on average, the proposed method performs the best among the three migration schemes. The results show that our method is most competitive. This also proves that the domain adaptation method with mixed high-order attention in the digital dataset experiment can obtain a better adaptive effect than other latest domain adaptation methods, such as 3CATN. Finally, the sound proof of our method shows that it is beneficial for the UDA model to take into account both domain confrontation and attention.

The purpose of the confusion matrixes shown in Figure 4 is to intuitively illustrate the effectiveness of our method. From Figure 4a, it can be seen that most of the samples of number category "8" are incorrectly predicted as "3", so misclassification is likely in some cases, especially in testing similar numbers between 7 and 2, also 9 and 4, when they reveal the huge differences between domains. With the increase of the order, the effect of HTAN-6 is improved, especially on "3" and "4". In some cases, they are likely to be misclassified, while Lor + (HTAN-6) reduces the differences of background edges under the constraint of orthogonal loss. In contrast, the use of Lor + Tri + (HTAN-6) shows more accurate predictions on the diagonal, proving that the proposed method can effectively alleviate regional and categorical differences.

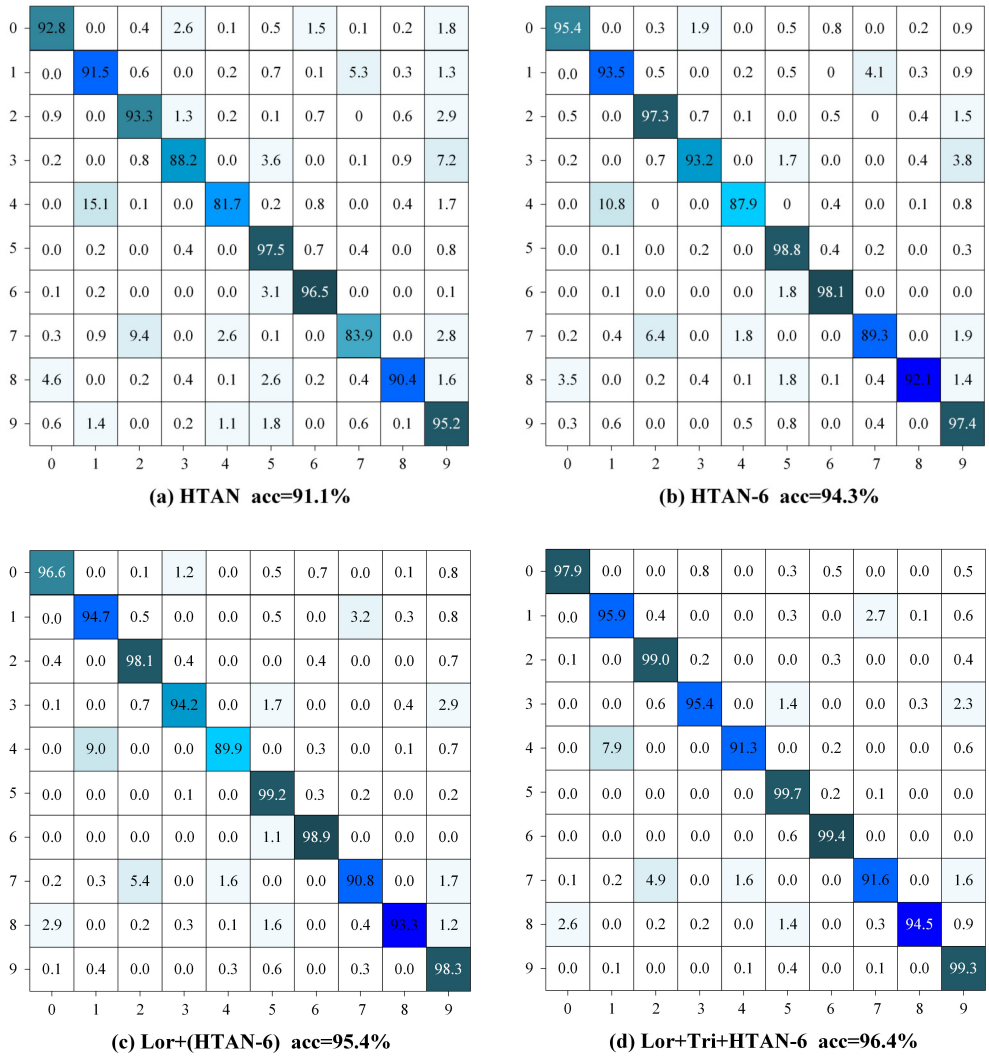


Figure 4. Confusion matrices for the digit adaptation (SVHN → MNIST).

4.3. Results on Dataset Office-31

Six transmission tasks were performed in the context of domain adaptation: A → W, D → W, D → A, W → A, W → D and A → D. The results are recorded in Table 2. From the migration tasks between office-31 datasets, it can be seen that DDC, DAN and JDDA are better than traditional shallow migration methods TCA and GFK. However, DANN, 3CTAN and lor+Tri(HTAN-6) have better performances than DDC, DAN and JDDA. The calculated average accuracy rate of two different tasks is 89.63%, higher than the average accuracy rate of 0.73% of the latest 3CATN model. It is worth noting that the method proposed by us obtained two migration tasks A → W and A → D. Based on the latest results, it should be noted that our method performs better than the basic model DANN on all migration tasks. In particular, the HTAN method achieves higher classification accuracy on some difficult migration tasks: A → W and A → D. Among these tasks, in terms of shooting angle and object attributes—color, etc.—the domain difference between the source domain and the target domain is significantly larger. HTAN can improve the adaptation tasks with larger domain differences, such as A → W, A → D, D → A and W → A, and achieve comparable classification accuracy on adaptation tasks with small domain differences.

Table 2. Accuracy (%) on the Office-31 dataset for unsupervised domain adaptation (ResNet-50 as base network).

Methods	A→W	D→W	W→D	A→D	D→A	W→A	Average
ResNet-50 [54]	68.4	96.7	99.3	68.9	62.5	60.7	76.1
GFK [16]	72.8	95.0	98.2	74.5	63.4	61.0	77.5
TCA [48]	72.7	96.7	99.6	74.1	61.7	60.9	77.6
DDC [13]	75.6	96.0	98.2	76.5	62.2	61.5	78.3
JDDA [49]	82.6	95.2	99.7	79.8	57.4	66.7	80.2
DAN [12]	80.5	97.1	99.6	78.6	63.6	62.8	80.4
DANN [17]	82.0	96.9	99.1	79.7	68.2	67.4	82.2
ADDA [18]	86.2	96.2	98.4	77.8	69.5	68.9	82.9
MADA [30]	90.0	97.4	99.6	87.8	70.3	66.4	85.2
GTA [51]	89.5	97.9	99.8	87.7	72.8	71.4	86.5
iCAN [52]	92.5	98.8	100	90.1	72.1	69.9	87.2
CDAN [19]	94.1	98.6	100	92.9	71.0	69.3	87.7
BSP+DANN [53]	93.0	98.0	100	90.0	71.9	73.0	87.7
BSP+CDAN [53]	93.3	98.2	100	93.0	73.6	72.6	88.5
3CATN [20]	95.3	99.3	100	94.1	73.1	71.5	88.9
HTAN	91.1	97.5	99.7	89.2	70.1	69.5	86.18
HTAN-6	94.1	98.5	100	92.4	72.5	72.9	88.4
Lor + (HTAN-6)	95.2	98.9	100	93.3	73.3	74.0	89.12
Lor + Tri + (HTAN-6)	95.8	99.1	100	94.3	73.9	74.7	89.63

From the results, the following conclusions can be drawn. First of all, DANN can train an additional domain classifier to minimize the difference, making its performance about 9.5% better than that of the standard deep neural network RTN. This improvement also shows that adversarial learning is useful for minimizing the difference between the source data and the target data. Secondly, the proposed method is significantly better than a series of other metric-based methods, such as DDC, DAN and CORAL. They all help the model's fully connected layer distribution. The minimization of difference indicates that blindly aligning the content and background parts may have a negative impact on the final result. Thirdly, combining high-order mixed attention with adversarial learning can improve the average accuracy rate by 6.2%. Finally, a new training method is added to HTAN-6. The rise of average accuracy by 0.51% indicates that our model is superior to the baseline method in reducing domain differences.

Figure 5 shows the confirmation of the convergence performances of ResNet-50, DAN and HTAN. From it, it can be seen that the proposed HTAN-6 enjoys faster convergence than DAN, while the performance of Lor + (HTAN-6) is better than that of HTAN-6. It is worth noting that the performance of Lor + (HTAN-6) has similarly stable converge performance as that of Lor + Tri + (HTAN-6) at the beginning of adversarial training, while Lor + Tri + (HTAN-6) remarkably outperforms Lor + (HTAN-6) in the whole procedure of convergence. Thus, as the training progresses, more fine-grained features are gradually learned between source domain and target domain, and the performance of Lor + Tri + (HTAN-6) becomes better than other approaches. The above findings confirm that our model can achieve minimum test error smoothly and quickly, resulting in better domain transfer.

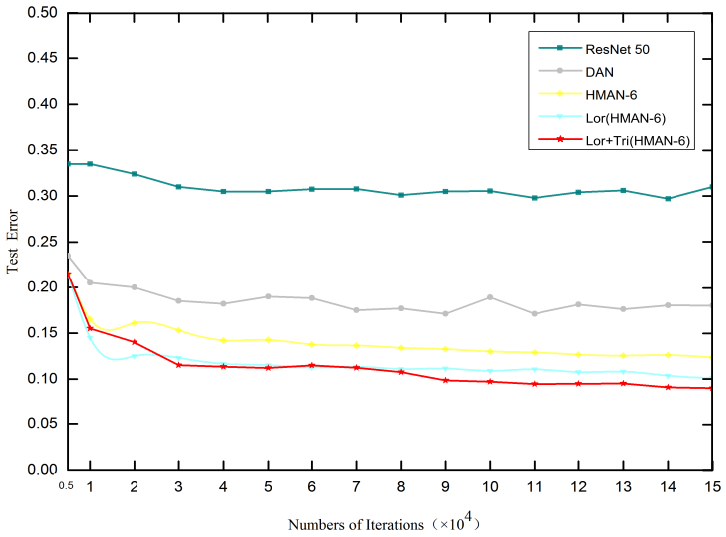


Figure 5. Convergence performance on the transfer task of A → W by ResNet-50, DAN, HTAN-6, Lor + (HTAN-6) and Lor + Tri + (HTAN-6).

4.4. Results on Dataset Office-Home

In the context of domain adaptation, 12 migration tasks were performed for four domains on the Office-Home dataset. The results recorded are listed in Table 3. The previous best average accuracy rate of TADA was 67.6%, while the average accuracy rate of the proposed HTAN was 68.66%. That accuracy is the new best performance. The average accuracy rate achieved by our method was 11.06% higher than the baseline method DANN. It is encouraging that HTAN adapts to tasks in some difficult areas (for example: Cl → Pr, meaning that HTAN can learn more transferable features for effective transfer learning.

From the results, the following important observations can be made. On the one hand, combining high-order hybrid attention with adversarial learning can increase the average accuracy rate by 3.63%, proving that our method can accurately select foreground alignment and retain fine feature information. On the other hand, a new three-part confrontation training method was added to the high-order hybrid confrontation model HTAN-6. In doing so, the average accuracy was increased by 0.57% on the original basis, showing that our model reduces the domain and that its ability of differentiation is better than that of the baseline adversarial approach.

Table 3. Accuracy (%) on the Office-home dataset for unsupervised domain adaptation (ResNet-50 as base network).

Methods	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Average
ResNet-50 [54]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [12]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
JAN [10]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
DANN [17]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
SE [50]	48.8	61.8	72.8	54.1	63.2	65.1	50.6	49.2	72.3	66.1	55.9	78.7	61.5
CDAN [19]	50.6	65.9	73.4	55.7	62.7	64.2	51.8	49.1	74.5	68.2	56.9	80.7	62.8
BSP+DANN [53]	51.4	68.3	75.9	56.0	67.8	68.8	57.0	49.6	75.8	70.4	57.1	80.6	64.9
BSP+CDAN [53]	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	59.3	81.9	66.3
TADA [22]	53.1	72.3	77.2	59.1	71.2	72.1	59.7	53.1	78.4	72.4	**60.0**	82.9	67.6
HTAN	51.2	68.1	75.4	55.9	68.7	69.5	59.2	50.5	75.9	70.3	56.6	79.1	65.03
HTAN-6	53.6	70.7	77.4	58.3	70.6	71.9	61.1	52.2	77.6	72.7	58.4	81.2	67.14
Lor + (HTAN-6)	54.5	72.1	77.5	59.1	71.4	72.9	62.4	53.2	79.0	73.6	59.1	82.3	68.09
Lor + Tri + (HTAN-6)	55.2	73.1	77.9	59.9	72.2	73.4	62.3	53.9	79.3	74.6	59.7	83.1	68.66

4.5. Effectiveness Verification

Ablation study: From Table 4, we analyze the individual contributions of hybrid high-order attention, orthogonal loss and adversarial training. When we use the DANN basic model without high-level attention, we get a certain improvement, but the recognition rate is not very obvious. It shows that the model did not learn the complex multimodal structure distribution. Secondly, the use of orthogonal loss and the three adversarial training can improve HTAN to a certain extent, indicating that they reduce the difference of background edges and learn domain invariant features under the constraint of orthogonal loss. By fair comparison, the results of the proposed Lor + Tri + (HTAN-6) uniformly outperform those of the other variants among these ablation experiments, which certifies its remarkable effect in matching the features and adversarial domain adaptation across domains.

Table 4. Accuracy(%) of different methods on Digits dataset (ResNet).

Methods	MNIST→USPS	USPS→MNIST	SVHN→MNITS	Average
DANN [17]	82.1	89.7	73.9	81.9
Lor+(DANN)	83.4	90.7	75.1	83.07
Lor+Tri+(DANN)	84.3	91.5	76.1	83.97
HTAN	91.2	93.9	88.2	91.1
Lor+(HTAN)	92.3	94.8	89.1	92.07
Lor + (HTAN-6)	96.0	97.8	92.5	95.43
Lor+Tri+(HTAN)	94.1	95.9	90.4	93.47
Lor + Tri + (HTAN-6)	97.0	98.7	93.6	96.43

Visualization: We visualize the network activations from the feature extractors of HTAN, HTAN-6, Lor + (HTAN-6) and Lor + Tri + (HTAN-6) on the transfer task A→ W by T-SNE [56] in Figure 6. By HTAN features, the source domain and target domain are not well aligned. From left HTAN to right Lor + Tri + (HTAN-6), the source and target domains are made more and more indistinguishable. For feature representations of Lor + Tri + (HTAN-6), the source and target domains are perfectly aligned while different classes are well discriminated. All of the above observations can demonstrate the advantages of the proposed method. The proposed method can learn more domaininvariant features with the hybrid high-order attention mechanism and triple adversarial learning, which is proved intuitively.

4.6. High-Order Performance Analysis and Convergence Performance

Effects of HTAN: Firstly, denote the order of HTAN with "HTAN-k", and the orders are R = $\{1, 2, 3, ..., k\}$. Then quantify HTAN, followed by a quantitative comparison of HTAN. The results are shown in Figure 6. It can be seen from this table that the proposed HTAN can significantly improve the recognition rate of the MNIST → USPS migration task. Specifically, as the order of HTAN rises, the recognition rate is further improved. For example, in the migration of MNIST → USPS, when the HTAN order rises from order 1 to order 6, the HTAN performance increases from 91.2% to 94.7%. This phenomenon shows that the use of a hybrid high-order attention model helps to capture the interactions of some complex and high-order statistical information. In order to further eliminate the background features, orthogonal modules were added to limit the background features. It can be seen from the figure that with the increase of the order, the migration recognition rate of Lor+(HTAN-k) model also increases from 92.2% to 96%, showing that our model can better concentrate on the fine features of the foreground. Therefore, a domain confrontation adaptive training method was added based on three-player games. Lor+Tri+(HTAN-k) model improves the recognition rate further on the basis of the first two. The performance of Lor+Tri+(HTAN-k) on three benchmarks is much better than

all baseline models, proving that our method has a better ability to express fine feature vectors and is effective for invariant features in the learning domain. However, when the order of the Lor + Tri + (HTAN-k) model is further increased, such as R = 7, the performance improvement remains almost unchanged; data not reported here.

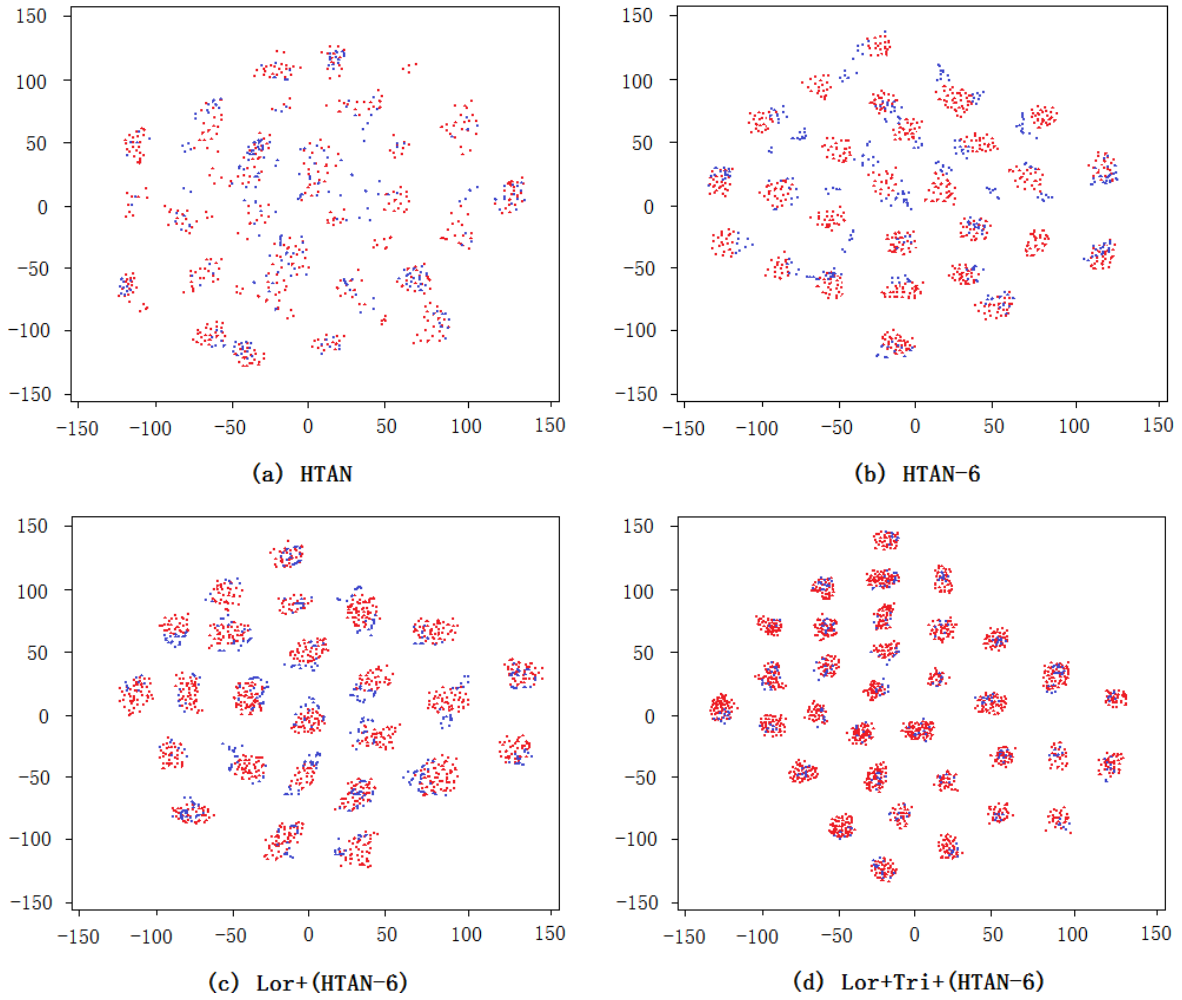
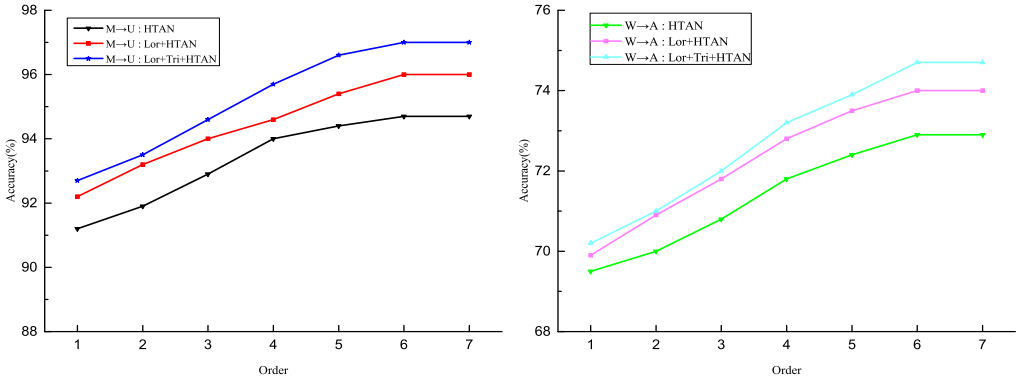


Figure 6. Visualization of images and feature distributions of source and target domains using T-SNE; the red and blue digits represent source Amazon (A) dataset and target Webcam (W) dataset respectively. (**a**) HTAN, (**b**) HTAN-6, (**c**) Lor + (HTAN-6) and (**d**) Lor + Tri + (HTAN-6) (red: A; blue: W).

Parameter sensitivity: From Figure 7 we can see that when the hyperparameter order is greater than 6, the performance of our hyperparameter order remains almost unchanged. From the analysis of convergence performance, the above findings confirm that our model can smoothly and quickly achieve the smallest test error, thereby achieving better domain migration. We recommend using the HTAN-6 model to generate multiple modules with different orders so that diverse and complementary high-order information can be used explicitly, thereby encouraging the richness of learning functions and preventing the learning of partial/biased visual information. When the order is 6, the model learns the best significant difference feature. Additionally, we use adversarial learning to generate diversified high-order attention maps with the order of constrained hyperparameters, which will not lead to a decrease in benchmark performance.



(**a**) The order change of MNIST→USPS task. (**b**) The order change of W→A task.

Figure 7. The influences of different orders on the migration task.

5. Conclusions

This paper concerns the proposition of a hybrid high-order triple-adversary network (HTAN), a new type of confrontation learning method with a hybrid high-order attention mechanism. It is different from the previous methods, which only match the low-order feature representation across domains, often leading to negative feature transfers. The proposed network uses a hybrid high-order attention mechanism to weight the extracted features, effectively eliminating the influences of non-transferable features. By considering the transitivity of different regions or images, the complex multi-modal structural information is further developed to achieve more precise feature matching. In addition, the orthogonal loss is integrated to further constrain the background features, then the triple-adversary strategy is adopted to improve the training convergence of the hybrid network. In the experiments, comprehensive evaluations on three benchmark datasets verified the superior performance of the proposed network, compared with the related adaptive models. Based on the paper, further studies might work in several areas. First, the task-specific decision boundary between categories should be focused on avoiding generated confusing features close to the category boundary; secondly, relieving the problem that the time cost excessively depends on the model order; and finally, specific challenging fields should be explored, such as cross-domain pedestrian re-recognition, in order to introduce more distinctive high-level attentions.

Author Contributions: Conceptualization, M.W. and J.F.; software, M.W. and J.F.; formal analysis, M.W.; investigation, M.W.; resources, M.W.; data curation, M.W.; writing—original draft preparation, M.W. and J.F.; writing—review and editing, M.W.; validation, J.F.; writing—original draft preparation, J.F.; methodology, J.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DAN	Deep Adaptation Network
DDC	Domain Confusion
DANN	Domain Adversarial Neural Network
JAN	Joint Adaptation Networks
MADA	Multi-Adversarial Domain Adaptation
CAN	Collaborative and Adversarial Network
ADDA	Discriminative Domain Adaptation ADDA
CDAN	Conditional Domain Adversarial Network

BSP	Batch Spectral Penalization
3CATN	Cycle-Consistent Conditional Adversarial Transfer networks
TADA	Transferable Attention For Domain Adaptation

References

1. Saenko, K.; Kulis, B.; Fritz, M.; Darrell, T. Adapting Visual Category Models to New Domains. In Proceedings of the European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; pp. 213–226.
2. Long, M.; Wang, J.; Ding, G.; Sun, J.; Yu, P.S. Transfer feature learning with joint distribution adaptation. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2200–2207.
3. Sun, B.; Feng, J.; Saenko, K. Return of frustratingly easy domain adaptation. *arXiv* **2015**, arXiv:1511.05547.
4. Chen, M.; Xu, Z.; Weinberger, K.; Sha, F. Marginalized denoising autoencoders for domain adaptation. *arXiv* **2012**, arXiv:1206.4683
5. Glorot, X.; Bordes, A.; Bengio, Y. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In Proceedings of the 28th International Conference on International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011.
6. Zhang, L.; Zuo, W.; Zhang, D. LSDT: Latent sparse domain transfer learning for visual adaptation. *IEEE Trans. Image Process.* **2016**, *25*, 1177–1191. [CrossRef] [PubMed]
7. Yan, Y.; Li, W.; Ng, M.K.; Tan, M.; Wu, H.; Min, H.; Wu, Q. Learning Discriminative Correlation Subspace for Heterogeneous Domain Adaptation. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 3252–3258.
8. Yan, Y.; Li, W.; Wu, H.; Min, H.; Tan, M.; Wu, Q. Semi-Supervised Optimal Transport for Heterogeneous Domain Adaptation. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 2969–2975.
9. Sun, B.; Saenko, K. Deep Coral: Correlation Alignment for Deep Domain Adaptation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 443–450.
10. Long, M.; Zhu, H.; Wang, J.; Jordan, M.I. Deep Transfer Learning with Joint Adaptation Networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 2208–2217.
11. Long, M.; Zhu, H.; Wang, J.; Jordan, M.I. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2016; pp. 36–144.
12. Long, M.; Cao, Y.; Wang, J.; Jordan, M. Learning transferable features with deep adaptation networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–7 August 2015; pp. 97–105.
13. Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; Darrell, T. Deep domain confusion: Maximizing for domain invariance. *arXiv* **2014**, arXiv:1412.3474.
14. Ben-David, S.; Blitzer, J.; Crammer, K.; Pereira, F. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2007; pp. 137–144.
15. Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; Vaughan, J.W. A theory of learning from different domains. *Mach. Learn.* **2010**, *79*, 151–175. [CrossRef]
16. Gong, B.; Shi, Y.; Sha, F.; Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2066–2073.
17. Ganin, Y.; Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–7 August 2015; pp. 1180–1189.
18. Tzeng, E.; Hoffman, J.; Saenko, K.; Darrell, T. Adversarial discriminative domain adaptation. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 7167–7176.
19. Long, M.; Cao, Z.; Wang, J.; Jordan, M.I. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2018; pp. 1640–1650.

20. Li, J.; Chen, E.; Ding, Z.; Zhu, L.; Lu, K.; Huang, Z. Cycle-consistent conditional adversarial transfer networks. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 747–755.
21. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Beach, CA, USA, 10–15 June 2019; pp. 7354–7363.
22. Wang, X.; Li, L.; Ye, W.; Long, M.; Wang, J. Transferable attention for domain adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, UAS, 27 January–1 February 2019; Volume 33, pp. 5345–5352.
23. Cai, S.; Zuo, W.; Zhang, L. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 511–520.
24. Wang, Q.; Li, P.; Zhang, L. G2DeNet: Global Gaussian distribution embedding network and its application to visual recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2730–2739.
25. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
26. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 2096–2030.
27. Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A.Y. Reading Digits in Natural Images with Unsupervised Feature Learning. 2011. Available online: https://storage.googleapis.com/pub-tools-public-publication-data/pdf/37648.pdf (accessed on 10 December 2020).
28. Venkateswara, H.; Eusebio, J.; Chakraborty, S.; Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5018–5027.
29. Shen, J.; Qu, Y.; Zhang, W.; Yu, Y. Wasserstein distance guided representation learning for domain adaptation. *arXiv* **2017**, arXiv:1707.01217.
30. Pei, Z.; Cao, Z.; Long, M.; Wang, J. Multi-adversarial domain adaptation. *arXiv* **2018**, arXiv:1809.02176.
31. Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.Y.; Isola, P.; Saenko, K.; Efros, A.; Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1989–1998.
32. Kumar, A.; Sattigeri, P.; Wadhawan, K.; Karlinsky, L.; Feris, R.; Freeman, B.; Wornell, G. Co-regularized alignment for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2018; pp. 9345–9356.
33. Chorowski, J.K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2015; pp. 577–585.
34. Show, A.; Xu, K. Tell: Neural image caption generation with visual attention. *arXiv* **2015**, *23*, Pre-Print.
35. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
36. Woo, S.; Park, J.; Lee, J.Y.; So Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
37. Ionescu, C.; Vantzos, O.; Sminchisescu, C. Matrix backpropagation for deep networks with structured layers. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2965–2973.
38. Li, P.; Xie, J.; Wang, Q.; Zuo, W. Is second-order information helpful for large-scale visual recognition? In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2070–2078.
39. Lin, T.Y.; RoyChowdhury, A.; Maji, S. Bilinear cnn models for fine-grained visual recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1449–1457.
40. Yu, K.; Salzmann, M. Second-order convolutional neural networks. *arXiv* **2017**, arXiv:1703.06817.

41. Ma, C.Y.; Kadav, A.; Melvin, I.; Kira, Z.; AlRegib, G.; Peter Graf, H. Attend and interact: Higher-order object interactions for video understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6790–6800.
42. Kossaifi, J.; Bulat, A.; Tzimiropoulos, G.; Pantic, M. T-net: Parametrizing fully convolutional nets with a single high-order tensor. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7822–7831.
43. Xie, D.; Xiong, J.; Pu, S. All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6176–6185.
44. Rodríguez, P.; Gonzalez, J.; Cucurull, G.; Gonfaus, J.M.; Roca, X. Regularizing cnns with locally constrained decorrelations. *arXiv* **2016**, arXiv:1611.01967.
45. Jia, K.; Li, S.; Wen, Y.; Liu, T.; Tao, D. Orthogonal deep neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, 1. [CrossRef] [PubMed]
46. Lezcano-Casado, M.; Martínez-Rubio, D. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. *arXiv* **2019**, arXiv:1901.08428.
47. Chen, Y.; Jin, X.; Feng, J.; Yan, S. Training group orthogonal neural networks with privileged information. *arXiv* **2017**, arXiv:1701.06772.
48. Pan, S.J.; Tsang, I.W.; Kwok, J.T.; Yang, Q. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* **2010**, *22*, 199–210. [CrossRef] [PubMed]
49. Chen, C.; Chen, Z.; Jiang, B.; Jin, X. Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 3296–3303.
50. French, G.; Mackiewicz, M.; Fisher, M. Self-ensembling for visual domain adaptation. *arXiv* **2017**, arXiv:1706.05208.
51. Sankaranarayanan, S.; Balaji, Y.; Castillo, C.D.; Chellappa, R. Generate to adapt: Aligning domains using generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8503–8512.
52. Zhang, W.; Ouyang, W.; Li, W.; Xu, D. Collaborative and adversarial network for unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3801–3809.
53. Chen, X.; Wang, S.; Long, M.; Wang, J. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In Proceedings of the International Conference on Machine Learning, Beach, CA, USA, 10–15 June 2019; pp. 1081–1090.
54. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
55. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
56. Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Darrell, T. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. 2013. Available online: https://arxiv.org/pdf/1310.1531.pdf (accessed on 10 December 2020).

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).