HGNet: Hybrid Generative Network for Zero-shot Domain Adaptation

Haifeng Xia[1] and Zhengming Ding[2]

[1] Department of ECE, Indiana University-Purdue University Indianapolis
[2] Department of CIT, Indiana University-Purdue University Indianapolis
{haifxia,zd2}@iu.edu

Abstract. Domain Adaptation as an important tool aims to explore a generalized model trained on well-annotated source knowledge to address learning issue on target domain with insufficient or even no annotation. Current approaches typically incorporate data from source and target domains for training stage to deal with domain shift. However, most domain adaptation tasks generally suffer from the problem that measuring the domain shift tends to be impossible when target data is inaccessible. In this paper, we propose a novel algorithm, *Hybrid Generative Network* (HGNet) for Zero-shot Domain Adaptation, which embeds an adaptive feature separation (AFS) module into generative architecture. Specifically, AFS module can adaptively distinguish classification-relevant features from classification-irrelevant ones to learn domain-invariant and discriminative representations when task-relevant target instances are invisible. To learn high-quality feature representation, we also develop hybrid generative strategy to ensure the uniqueness of feature separation and completeness of semantic information. Extensive experimental results on several benchmarks illustrate that our method achieves more promising results than state-of-the-art approaches.

Keywords: Deep Learning · Domain Adaptation· Generative Learning

1 Introduction

Computer vision community always suffers from insufficient annotation issue, which dramatically obstructs the practical applications of most techniques. However, domain adaptation provides an alternative strategy to handle with such a problem [24, 9, 28]. Concretely, it attempts to borrow knowledge from well-annotated modality (source domain) to solve classification task on target domain without any label information [30, 32, 35]. Although various domains share the high-level semantic information, their data distributions contain significant discrepancy defined as domain shift [10, 34, 13]. For example, due to light condition or occlusions, visual instances involving the same object are different from each other [4]. As a result, the previously-trained model generally tends to be fragile when evaluated on target domain.

Domain adaptation (DA) as a solution to learn domain-invariant knowledge attracts great interest [7, 2, 20, 22]. To learn transferable information, it assumes

that instances of target modality are available [23, 17, 5]. Under such an assumption, recent works mainly explore two approaches: discrepancy measurement [16] and domain adversarial confusion [15, 35]. Specifically, the first strategy aims to define novel statistic indicators like maximum mean discrepancy (MMD) [7] promoting the consistency of distribution. While methods based on domain adversarial confusion expect to transform data of source and target domain into the similar hidden space by using adversarial relationship between generator and discriminator. They actually have achieved promising improvement in distinctive tasks. In real-world scenarios, however, the assumption which they depend on is infeasible due to the absence of target domain. The general situation is defined as *zero-shot domain adaptation* (ZSDA) [21], which is also known as *missing modality transfer learning* [8]. For instance, to protect privacy of patient, hospital fails to share medical records to train the model, even though they expect to apply the trained model for their work, where these documents represent target domain. In this sense, the current DA methods are more likely to be invalid since the guidance of target datatset becomes invisible.

The awkward situation inspires [19] to proposes domain-invariant component analysis (DICA) by using multiple source domains with identical label space to build a generalized model for unseen target recognition. However, they hardly collect sufficient source domains to observe the information of unseen target modality. To solve this problem, the intuitive motivation is to introduce auxiliary task-irrelevant dataset (TIR), which also includes two same modalities with the task-relevant one (TR) [8]. Alternatively, [21] develops the first deep model for zero-shot domain adaptation which firstly attempts to achieve the feature alignment on task-irrelevant datasets and then allows source modalities in TR and TIR to share the same network. Moreover, the generalization of neural network facilitates the consistency of cross-domain distribution on task-relevant dataset. Albeit the training manner enables model to generate domain-invariant representation, features tend to be less discriminative without the guidance of annotation when training model on task-irrelevant inputs, leading to the decrease of recognition. Meanwhile, due to the huge achievement of generative adversarial model in abundant practical scenarios, it is appropriate to utilize this manner to synthesis missing modality and directly perform domain adaptation in TR datasets [27] named CocoGAN. However, the drawbacks of generative adversarial network is that there exists bias between generated instances and real samples, since synthesised images only try to approximate the real distribution. Thus, estimating the influence of bias on the final classification task tend to be very difficult. On the other hand, we naturally post a question about CocoGAN: "Is the explicit generation of missing target dataset necessary for learning domain-invariant feature?".

To answer this question, we rethink Zero-shot Domain Adaptation from feature separation and propose Hybrid Generative Network (HGNet), which not only synthesises domain-invariant feature but also effectively facilitates high-level representation to be more discriminative. Specifically, the whole network architecture mainly consists of four components: feature extractor, adaptive feature

separation module, hybrid generator and classifier. Input signals of TR and TIR datasets firstly pass through feature extractor and are transformed into shallow convolutional units. For the second step, feature separation module adaptively selects several channels to form classification-relevant high-level feature, while others are considered as classification-irrelevant information. In the final stage, on one hand, we apply the supervision of annotation to learn more discriminative units. On the other hand, hybrid generator will integrate object context and domain information belonging to various datasets to reconstruct input data. Extensive experimental performances illustrate that the hybrid strategy guarantees the uniqueness of feature separation as well as the completeness of semantic information. The contributions of our method are summarized in three folds:

- From the perspective of feature separation, we introduce a novel strategy named Hybrid Generative Network (HGNet) to fight off ZSDA more effectively. The proposed feature separation module guided by annotation explores global information from shallow convolutional layers to extract more discriminative and domain-invariant units.
- To perform high-quality feature separation, we develop hybrid generation module assisting model to capture association between task-relevant (TR) and task-irrelevant (TIR) datasets. The benefit of such a relationship is to utilize cross-domain knowledge learned from TIR to eliminate domain shift on TR datasets.
- We assess our model on several visual cross-domain tasks, and HGNet outperforms competitive approaches by large margin in most cases, illustrating the effectiveness on solving ZSDA challenge. We further conduct extensive empirical study to demonstrate the function of hybrid generation.

2 Related Work

Domain adaptation (DA) has attracted great interest as it addresses limited annotation problem [25]. And recent works attempt to apply DA strategy in computer vision like image classification [12, 18], object segmentation [29, 36, 26] and image caption [3]. However, they generally suffer from a primary challenge defined as domain shift deriving from the difference of distribution across domains. To mitigate such an issue, current proposed approaches are divided into two branches: dissimilarity measurement using statistic indicators to align distribution [16, 11] and domain adversarial confusion [15, 35, 30] adopting adversarial manner to generate cross-domain features in the same latent space. Although these methods effectively learn domain-invariant representation, they significantly depend on the existence of samples from target domain. As a result, the situation where we fails to have access to the target modality dramatically obstructs the practical application of these techniques, which triggers another hot research topic named zero-shot domain adaptation (ZSDA) [27] also known as missing modality transfer learning [8]. The novel problem assumes that we just are given task-relevant source domain and auxiliary datasets including task-irrelevant source and target domains.

For the existing methods to solve ZSDA, they firstly attempt to utilize task-irrelevant samples to eliminate cross-domain discrepancy and then they transform samples of task-relevant source and target domain into the same hidden space [21]. In addition, with the advance of generative adversarial network in recent year, [27] proposes conditional coupled GAN (CoCoGAN) to generate task-relevant paired samples in the first step and train classifier on synthesised dataset. Different from them, we rethink ZSDA from the perspective of feature separation selecting more discriminative feature as domain-invariant feature, which effectively promote the generalization of model. And to obtain high-quality feature separation, we propose hybrid generative strategy ensuring the uniqueness of feature and the completeness of semantic information.

3 The Proposed Method

3.1 Preliminaries and Motivation

Zero-shot Domain Adaptation aims to exploit all accessible data to learn robust and generalized model used to deal with classification issue on target domain. Concretely, we are given well-annotated task-relevant source dataset $\mathcal{D}^{r,s} = \{(\mathbf{X}_i^{r,s}, Y_i^{r,s})\}_{i=1}^{n}$, where $\mathbf{X}_i^{r,s}$ and $Y_i^{r,s}$ separately denote i-th visual instance and its corresponding label. In addition, we also have access to task-irrelevant cross-domain paired datasets $\mathcal{D}^{ir,s} = \{(\mathbf{X}_i^{ir,s}, Y_i^{ir,s})\}_{i=1}^{m}$ and $\mathcal{D}^{ir,t} = \{(\mathbf{X}_i^{ir,t}, Y_i^{ir,t}\}_{i=1}^{m}$. Although $\mathbf{X}_i^{ir,s}$ and $\mathbf{X}_i^{ir,t}$ lie in various domains (source and target), they belong to the same category i.e., $Y_i^{ir,s} = Y_i^{ir,t}$. To this end, it is impossible for model to capture any knowledge of task-relevant target dataset $\mathcal{D}^{r,t} = \{\mathbf{X}_i^{r,t}\}_{i=1}^{n}$ only available in the test stage. The current scenario mainly involves two challenges: 1) **Generation of domain-invariant representation:** The absence of $\mathcal{D}^{r,t}$ results in huge difficulty of directly measuring cross-domain discrepancy between $\mathcal{D}^{r,s}$ and $\mathcal{D}^{r,t}$; 2) **Fusion of various datasets:** Tremendous difference among $\mathcal{D}^{r,s}$, $\mathcal{D}^{ir,s}$ and $\mathcal{D}^{ir,t}$ dramatically interferes their connection.

To capture domain shift between $\mathcal{D}^{r,s}$ and $\mathcal{D}^{r,t}$, the intuitive idea [27] is to firstly synthesize missing modality $\mathcal{D}^{r,t}$ and then transform them into the similar latent space, which arises a question: "*Is the explicit generation of missing target dataset necessary for learning domain-invariant feature*?" To answer this question, we rethink and explore the extraction of domain-invariant representation from the perspective of feature separation. Specifically, the intrinsic knowledge of input data generally is stored in high-level semantic representation via feature extractor. However, these semantic information is not equally necessary in terms of classification task. Admittedly, partial abstract representations record abundant essential content as visual style or background in object image, but they are drastically various across domains. We consider these representations as classification-irrelevant features, which are undesirable in domain adaptation. On the other hand, the remaining part defined as classification-relevant feature has positive influence on our final object classification task. Considering the previous approaches about domain-invariant feature learning, it is irrational or

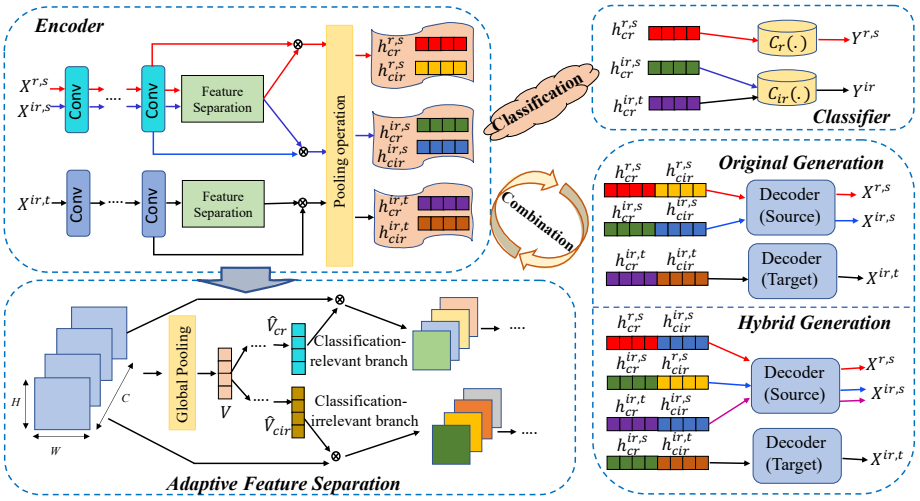


Fig. 1: Overview of the proposed HGNet, which mainly includes four components: encoder, decoder, classifier and adaptive feature separation module. The encoder firstly aims to extract convolutional features, and then the adaptive feature separation module attempts to learn classification-relevant and classification-irrelevant units. On one hand, we utilize label information to guarantee the effect of feature separation. On the other hand, we explore two reconstruction manners to promote the completeness of semantic information and the uniqueness of learned feature.

even counterproductive to incorporate all information into the same representation. Therefore, we achieve two primary conclusions: 1) Feature separation is important to distinguish domain-invariant feature out of classification-irrelevant features instead of generating missing dataset $\mathcal{D}^{r,t}$; and 2) we should only explore discriminative information on the the selected classification-relevant representations. According to these discussions, we propose our adaptive feature separation module embedded into auto-encoder framework.

Due to feature separation, classification-irrelevant representations of instances from $\mathcal{D}^{r,s}$ and $\mathcal{D}^{ir,s}$ should preserve high-similarity. Such relationship is also applied to $\mathcal{D}^{r,t}$ and $\mathcal{D}^{ir,t}$. Cross-domain paired datasets $\mathcal{D}^{ir,s}$ and $\mathcal{D}^{ir,t}$ tend to be transformed into the same hidden space with respect to classification-relevant feature, which is also suitable for $\mathcal{D}^{r,s}$ and $\mathcal{D}^{r,t}$. Based on these above analyses, we develop hybrid reconstruction strategy to build the connection among various datasets and promote the performance of domain adaptation.

3.2 Adaptive Feature Separation

To effectively learn domain-invariant hidden units, we propose adaptive feature separation module, which is capable of distinguishing classification-relevant features from classification-irrelevant ones. As a result, the mechanism tends to

describe same instance from two completely distinctive semantic views. To be specific, a branch of this module guided by discriminative information (annotation) aims to generate classification-relevant features, while the other branch will store other semantic contents. Moreover, auto-encoder framework combines them to reconstruct the input signal, which indeed guarantees the completeness of information and the difference between these two types of feature. From this property, we explore automatic feature selection from channel level.

Additionally, due to the generalization of deep neural network on feature learning, $\mathcal{D}^{r,s}$ and $\mathcal{D}^{ir,s}$ belonging to the same modality should share the network architecture and corresponding parameters. For $\mathcal{D}^{ir,t}$, the difference between source and target domain inspires us to adopt a distinctive network framework sharing parameters in higher network layers with the network for source domain. As shown in Figure 1, two various encoders involving convolutional operation convert the input signals $\mathbf{X}^{r,s}$, $\mathbf{X}^{ir,s}$ and $\mathbf{X}^{ir,t}$ into abstract representations $\mathbf{F}^{r,s}$, $\mathbf{F}^{ir,s}$, $\mathbf{F}^{ir,t} \in \mathbf{R}^{W\times H\times C}$, where W, H separately denote the width and height of each tensor, and C is the number of channel in tensor. At this time, the extracted features incorporate all semantic information of input data.

To learn domain-invariant features, we implement convolutional transformation to generate classification-relevant feature $\mathbf{F} \rightarrow \hat{\mathbf{F}}_{cr} \in \mathbf{R}^{W\times H\times C}$ and classification-irrelevant one $\mathbf{F} \rightarrow \hat{\mathbf{F}}_{cir} \in \mathbf{R}^{W\times H\times C}$, where $\mathbf{F}$ is selected from $\{\mathbf{F}^{r,s}, \mathbf{F}^{ir,s}, \mathbf{F}^{ir,t}\}$. The first transformation $\mathbf{F} \rightarrow \hat{\mathbf{F}}_{cr}$ performs a positive activation on convolutional layer via the guidance of label information to capture more discriminative information while gradually eliminating classification-irrelevant semantic content preserved in $\hat{\mathbf{F}}_{cir}$ with negative activation. Concretely, we firstly operate global average pooling technique on shallow convolutional feature $\mathbf{F}$ to obtain the information increment of each channel defined by $\mathcal{V} \in \mathbf{R}^{1\times 1\times C}$. Intuitively, each element $v_i \in \mathcal{V}$ roughly reflects content and style of the corresponding channel. To observe the connection across channels and separate features, we first adopt two distinctive non-linear manners to compress $\mathcal{V}$ to $\widetilde{\mathcal{V}}_{cr}$ and $\widetilde{\mathcal{V}}_{cir} \in \mathbf{R}^{1\times 1\times \frac{C}{\gamma}}$, where γ is a ratio controlling the scale of dimension-reduction and then utilize various full-connection layers to obtain new channel-wise statistics $\hat{\mathcal{V}}_{cr}$ and $\hat{\mathcal{V}}_{cir} \in \mathbf{R}^{1\times 1\times C}$. After the activation operation, $\hat{\mathcal{V}}_{cr}$ ideally promotes performance of several channels recording extensive discriminative information, while $\hat{\mathcal{V}}_{cir}$ enhances representation of others. Based on the above explanation, convolutional conversion can be formulated as:

$$\hat{\mathcal{V}}_{cr} = \sigma\Big(\mathbf{W}_{cr}\delta\big(g_{cr}(\mathcal{V})\big)\Big), \qquad \hat{\mathcal{V}}_{cir} = \sigma\Big(\mathbf{W}_{cir}\delta\big(g_{cir}(\mathcal{V})\big)\Big), \tag{1}$$

where $\mathbf{W}_{cr}$, $\mathbf{W}_{cir} \in \mathbf{R}^{C\times \frac{C}{\gamma}}$, $\sigma(\cdot)$ and $\delta(\cdot)$ represent **Sigmoid** and **ReLU** activation functions, $g_{cr}(\cdot)$ and $g_{cir}(\cdot)$ refer to the non-linear dimension-reduction operations. To achieve the feature separation based on classification-task, we conduct channel-wise multiplication ($\otimes$) between original convolutional features $\mathbf{F}$ and learned channel-wise indicators $\hat{\mathcal{V}}_{cr}$, $\hat{\mathcal{V}}_{cir}$ as the following:

$$\hat{\mathbf{F}}_{cr} = \hat{\mathcal{V}}_{cr} \otimes \mathbf{F} = \{\hat{v}_{cr,i} \cdot \mathbf{F}_i\}_{i=1}^{C}, \qquad \hat{\mathbf{F}}_{cir} = \hat{\mathcal{V}}_{cir} \otimes \mathbf{F} = \{\hat{v}_{cir,i} \cdot \mathbf{F}_i\}_{i=1}^{C}. \tag{2}$$

To guide feature separation on convolutional layer, we enforce $\hat{\mathbf{F}}_{cr}$ and $\hat{\mathbf{F}}_{cir}$ to pass through a series of operations including **Pooling**, **FC**, **ReLU** and **FC** to synthesize high-level semantic features h_{cr} and $h_{cir} \in \mathbf{R}^{d\times 1}$, where d is the dimension of feature. The learned representation h_{cr} as domain-invariant feature should be fed into the corresponding classifier to promote its discriminative ability. Considering that h_{cir} is required to preserve classification-irrelevant information, the concatenation of h_{cr} and h_{cir} will be taken as input for decoder including several deconvolutional layers [33] to achieve the reconstruction about the input data, i.e., $\hat{\mathbf{X}} = \mathcal{G}(h_{cr}, h_{cir})$, where $\mathcal{G}$ denotes neural network of decoder. Therefore, the objective function of adaptive feature separation module is written as:

$$\begin{aligned} \min_{\Theta} \quad & \mathcal{L}_c(\mathbf{C}(h_{cr}), Y) + \|\mathbf{X} - \mathcal{G}(h_{cr}, h_{cir})\|_F^2 \\ & h_{cr} \in \{h_{cr}^{r,s}, h_{cr}^{ir,s}, h_{cr}^{ir,t}\}, \quad Y \in \{Y^{r,s}, Y^{ir,s}, Y^{ir,t}\} \\ & h_{cir} \in \{h_{cir}^{r,s}, h_{cir}^{ir,s}, h_{cir}^{ir,t}\}, \quad \mathbf{X} \in \{\mathbf{X}^{r,s}, \mathbf{X}^{ir,s}, \mathbf{X}^{ir,t}\}, \end{aligned} \tag{3}$$

where Θ refers to all parameters of model, $\mathbf{C} = \{\mathbf{C}^r, \mathbf{C}^{ir}\}$ represents classifier ($h_{cr}^{ir,s}$ and $h_{cr}^{ir,t}$ share classifier C^{ir}, while classifier C^r is target for $h_{cr}^{r,s}$), $\mathcal{L}_c(\cdot)$ means cross-entropy loss and $\mathcal{G}$ consists of two types: $\mathcal{G}_s$ shared by source domain and $\mathcal{G}_t$ used by target domain. Note that the application of objective function requires the consistence of superscript.

3.3 Hybrid Generation

The benefit of adaptive feature separation is to extract more discriminative domain-invariant feature with the guidance of label information. To further eliminate domain shift, we propose hybrid reconstruction strategy capturing the connection across various datasets. In other words, we explore the feature alignment between $\mathcal{D}^{ir,s}$ and $\mathcal{D}^{ir,t}$ as well as the consistence of modality over $\mathcal{D}^{r,s}$ and $\mathcal{D}^{ir,s}$ to reduce cross-domain discrepancy of $\mathcal{D}^{r,s}$ and unavailable $\mathcal{D}^{r,t}$.

According to Section 3.2, any input signals passing through corresponding encoder and adaptive feature separation module will be transformed into classification-relevant features and classification-irrelevant ones. Due to the paired relationship between $\mathbf{X}^{ir,s}$ and $\mathbf{X}^{ir,t}$, it is reasonable to assume that there exists high similarity between $h_{cr}^{ir,s}$ and $h_{cr}^{ir,t}$ (i.e. $h_{cr}^{ir,s} \equiv h_{cr}^{ir,t}$) derived from corresponding input data. In terms of such equivalent property, we can assert the decoder $\mathcal{G}_t$ performed on ($h_{cr}^{ir,s}$, $h_{cir}^{ir,t}$) and ($h_{cr}^{ir,t}$, $h_{cir}^{ir,t}$) tend to generate the same result, which is formulated as:

$$\mathcal{G}_t(h_{cr}^{ir,s}, h_{cir}^{ir,t}) \equiv \mathbf{X}^{ir,t} \equiv \mathcal{G}_t(h_{cr}^{ir,t}, h_{cir}^{ir,t}). \tag{4}$$

With respect to the decoder of source domain $\mathcal{G}_s$, we can similarly draw the conclusion as:

$$\mathcal{G}_s(h_{cr}^{ir,s}, h_{cir}^{ir,s}) \equiv \mathbf{X}^{ir,s} \equiv \mathcal{G}_s(h_{cr}^{ir,t}, h_{cir}^{ir,s}). \tag{5}$$

To this end, the loss function of hybrid reconstruction and feature alignment is defined:

$$\begin{aligned}\mathcal{L}_{hr}^{ir} = \lambda\|h_{cr}^{ir,s} - h_{cr}^{ir,t}\|_F^2 + \|\mathcal{G}_t(h_{cr}^{ir,s}, h_{cir}^{ir,t}) - \mathbf{X}^{ir,t}\|_F^2 \\ + \|\mathcal{G}_s(h_{cr}^{ir,s}, h_{cir}^{ir,s}) - \mathbf{X}^{ir,s}\|_F^2,\end{aligned} \tag{6}$$

where λ is the hyper-parameter controlling the reconstruction and feature alignment. The first term in Eq. (6) not only achieves distribution alignment over task-irrelevant datasets, but also gradually eliminates the difference of models on feature learning. Under such condition, even though target dataset $\mathcal{D}^{r,t}$ is unavailable for training stage, the similarity of model effectively facilitates the consistency of feature representation across $\mathcal{D}^{r,s}$ and $\mathcal{D}^{r,t}$. Meanwhile, hybrid reconstruction loss plays an essential role in achieving the goal of feature separation, which aims to preserve abundant meaningful and discriminative feature in classification-relevant representation via the last two terms.

From Figure 1, we observe that classification-irrelevant units derived from $\mathbf{X}^{r,s}$ and $\mathbf{X}^{ir,s}$ ideally should maintain high correlation, since their corresponding input signals belong to the same modality. However, $h_{cr}^{r,s}$ and $h_{cr}^{ir,s}$ tend to describe distinctive objects of images. The expected association between $\mathcal{D}^{r,s}$ and $\mathcal{D}^{ir,s}$ is expressed as:

$$\mathcal{G}_s(h_{cr}^{r,s}, h_{cir}^{r,s}) \equiv \mathbf{X}^{r,s} \approx \mathcal{G}_s(h_{cr}^{r,s}, h_{cir}^{ir,s}). \tag{7}$$

$$\mathcal{G}_s(h_{cr}^{ir,s}, h_{cir}^{ir,s}) \equiv \mathbf{X}^{ir,s} \approx \mathcal{G}_s(h_{cr}^{ir,s}, h_{cir}^{r,s}). \tag{8}$$

Therefore, we explore hybrid generation to satisfy such a requirement and reformulate our objective function as:

$$\mathcal{L}_{hr}^{s} = \|\mathcal{G}_s(h_{cr}^{r,s}, h_{cir}^{ir,s}) - \mathbf{X}^{r,s}\|_F^2 + \|\mathcal{G}_s(h_{cr}^{ir,s}, h_{cir}^{r,s}) - \mathbf{X}^{ir,s}\|_F^2. \tag{9}$$

Remarks: If we have access to the missing target modality $\mathbf{X}^{r,t}$, the constraint of Eq. (9) enables the model to capture relationships: $\mathcal{G}_t(h_{cr}^{r,t}, h_{cir}^{ir,t}) \approx \mathbf{X}^{r,t} \equiv \mathcal{G}_t(h_{cr}^{r,t}, h_{cir}^{r,t})$ and $\mathcal{G}_t(h_{cr}^{ir,t}, h_{cir}^{r,t}) \approx \mathbf{X}^{ir,t} \equiv \mathcal{G}_t(h_{cr}^{ir,t}, h_{cir}^{ir,t})$. Moreover, under the supervision of Eq. (9), we also achieve the conclusion $\mathcal{G}_t(h_{cr}^{r,s}, h_{cir}^{r,t}) \approx \mathbf{X}^{r,t} \equiv \mathcal{G}_t(h_{cr}^{r,t}, h_{cir}^{r,t})$ and $\mathcal{G}_s(h_{cr}^{r,t}, h_{cir}^{r,s}) \approx \mathbf{X}^{r,s} \equiv \mathcal{G}_s(h_{cr}^{r,s}, h_{cir}^{r,s})$. Through such mediate manner, the model finally achieves domain adaptation across $\mathcal{D}^{r,s}$ and $\mathcal{D}^{r,t}$.

3.4 Training and Inference

Given accessible datatsets $\mathcal{D}^{r,s}$, $\mathcal{D}^{ir,s}$ and $\mathcal{D}^{ir,st}$, we firstly perform initial feature separation within each dataset. And then hybrid reconstruction as an important component captures delicate association across all datasets to gradually reduce cross-domain discrepancy between $\mathcal{D}^{r,s}$ and missing target dataset $\mathcal{D}^{r,t}$. Finally, we utilize the feature extractor of target domain to learn feature of $\mathbf{X}^{r,t}$ and apply classifier $\mathbf{C}_s(\cdot)$ to perform classification task. Therefore, the overall process is summarized as three steps:

Step A: Input data including $\mathbf{X}^{r,s}$, $\mathbf{X}^{ir,s}$ and $\mathbf{X}^{ir,t}$ first is fed into the encoder to learn convolutional features. And then we perform adaptive feature separation on convolutional layers to obtain classification-relevant unit h_{cr} and h_{cir}. Finally, the concatenation of h_{cr} and h_{cir} is exploited to reconstruct input signal. During learning stage, we explore objective function (3) to update model.
Step B: To achieve the expected feature separation and domain adaptation, we should integrate hybrid reconstruction and the guidance of label information into a unified loss function as:

$$\begin{aligned} \min_{\Theta} \quad & \mathcal{L}_{hr}^{ir} + \mathcal{L}_{hr}^{s} + \mathcal{L}_{c}\big(\mathbf{C}(h_{cr}), Y\big) \\ & h_{cr} \in \{h_{cr}^{r,s}, h_{cr}^{ir,s}, h_{cr}^{ir,t}\}, Y \in \{Y^{r,s}, Y^{ir,s}, Y^{ir,t}\}, \end{aligned} \tag{10}$$

where $\mathbf{C}$ consists of $\mathbf{C}_s$ classifier used by $h_{cr}^{r,s}$ and $\mathbf{C}_t$ classifier shared by $h_{cr}^{ir,s}$ and $h_{cr}^{ir,t}$. We train the network according to Eq. (10) until convergence.
Step C: During inference stage, instances $X^{r,t}$ will be passed through the encoder used by $X^{ir,t}$ to obtain high-level feature $h_{cr}^{r,t}$. Eventually, we utilize classifier $\mathbf{C}_s$ to predict the annotation of $h_{cr}^{r,t}$.

4 Experiments

4.1 Datasets and Comparisons

We perform experiments on three popular benchmarks involving MNIST [14], Fashion MNIST [31] and EMNIST [6] to verify the effectiveness of our method. For the convenience and clarity, we utilize dataset IDs D_M, D_F and D_E to refer to them. In addition, there exists three techniques to transform each gray-scale image into the corresponding negative, color and edge images.

MNIST (D_M) dataset is developed to identify handwritten digit image. The dataset includes 70,000 gray-scale images, where 60,000 training instances and 10,000 testing images. Each visual instance with same size 28×28 only represents one of ten digits from 0 to 9.

Fashion MNIST (D_F) dataset includes abundant fashion trappings images. Experts in fashion field artificially divide them into ten categories: *T-shirt*, *trouser*, *pullover*, *dress*, *coat*, *sandals*, *shirt*, *sneaker*, *bag*, and *ankle boot*. The dataset has the same sample scale with MNIST, i.e 60,000 training instances and 10,000 testing samples. The image size of each sample is also 28×28.

EMNIST (D_E) dataset different from MNIST records extensive handwritten alphabets images. The uppercase and lowercase letters are merged into a balanced dataset with 26 categories. The image size of each sample is 28×28. Moreover, it involves 124,800 images for training and 20,800 images for testing.

Modality Transformation: All instances in the above mentioned datasets are gray-scale images and we define this modality as *G-domain*. To perform domain adaptation, We firstly follow the operations in [27] to convert all original data into negative image (*N-domain*) by using $\mathbf{X}_n = 255 - \mathbf{X}$, $\mathbf{X} \in \mathbf{R}^{m \times n \times 1}$ where m and n are the spatial dimensions of image. Moreover, we apply canny

Table 1: Classification Accuracy (%) of our method and three baselines for domain adaptation from gray-scale modality (*G-domain*) to color modality (*C-domain*). The best result in each column is in bold.

RT	MNIST (D_M)		Fashion-Mnist		EMNIST (D_E)	
IRT	D_F	D_E	D_M	D_E	D_M	D_F
ZDDA [21]	73.2	94.8	51.6	65.3	71.2	47.0
CoGAN [27]	68.3	74.7	39.7	55.8	46.7	41.8
CoCoGAN [27]	78.1	**95.6**	56.8	66.8	**75.0**	54.8
HGNet	**85.3**	95.0	**64.5**	**71.1**	71.3	**57.9**

detector to create edge images $\mathbf{X}_e$ (*E-domain*). Finally, in terms of color version, we randomly extract several patches ($\mathbf{P} \in \mathbf{R}^{m \times n}$) from the BSDS500 dataset [1] and then blend them with images $\mathbf{X}$ to form color images $\mathbf{X_c}$ (*C-domain*).

Comparisons: To evaluate the performance of our method, we select three baselines as competed methods which are currently the only works exploring the application of deep learning on zero-shot domain adaptation problem. The first compared approach is *ZDDA* [21], which propose sensor fusion to solve domain shift. Moreover, [27] utilizes two models named *CoGAN* and *CoCoGAN* to address ZSDA issue, which are considered as two various approaches.

4.2 Implementation Details

The network architecture of our method mainly includes three components: encoder, decoder and classifier. Although source and target utilize various networks, they have the same network structure. Thus, we take the branch of source domain as an example to illustrate the specific implementation. With respect to the encoder, we adopt three convolutional layers with stride 2 to extract channel-level feature and apply **ReLU** to activate the output of the first two layers. Symmetrically, the decoder has three deconvolutional layers with stride 2 to recover hidden representation to input data. There are two classifiers used in our proposed method and they both have two full-connection layers followed by **Softmax** function.

4.3 Experimental Results

In order to validate the effectiveness of our method, we create five different zero-shot domain adaption settings. We firstly consider gray-scale images as source domain and the other three domains will be target domain. Thus, there are three domain adaptation tasks: *G-domain* $\rightarrow$ *N-domain*, *G-domain* $\rightarrow$ *E-domain* and *G-domain* $\rightarrow$ *C-domain*. In addition, we also attempt to transfer knowledge from color domain or negative domain to gray domain, i.e., *C-domain* $\rightarrow$ *G-domain* and *N-domain* $\rightarrow$ *G-domain*.

Table 2: Classification Accuracy (%) of our method and three baselines for two domain adaptation tasks :*N-domain* $\rightarrow$ *G-domain* and *G-domain* $\rightarrow$ *N-domain*. The best result in each column is in bold.

Task	*N-domain*$\rightarrow$*G-domain*						*G-domain*$\rightarrow$*N-domain*			
RT	D_M		D_F		D_E		D_M		D_F	D_E
IRT	D_F	D_E	D_M	D_E	D_M	D_F	D_F	D_E	D_E	D_F
ZDDA [21]	78.5	87.6	56.6	67.1	67.7	45.5	77.9	90.5	62.7	53.4
CoGAN [27]	66.1	76.3	49.9	58.7	53.0	32.5	62.7	72.8	51.2	39.1
CoCoGAN [27]	80.1	93.6	63.4	72.8	**78.8**	58.4	80.3	93.1	69.3	56.5
HGNet	**87.5**	**95.0**	**64.6**	**75.1**	78.0	**67.9**	**83.7**	**95.7**	**71.7**	**62.3**

According to descriptions of dataset, we know these three datasets involves three completely distinctive objects: digits, trappings and letters. When selecting one of them as task-relevant dataset, we can consider others as task-irrelevant datasets which assist model to capture cross-domain discrepancy and promote classification accuracy on missing target modality ($D^{r,t}$). Firstly, we attempt to transfer knowledge from gray-scale modality (*G-domain*) to color modality (*C-color*). Compared with gray-scale image, original RGB image generally involve three color channels, which dramatically increase the difficult in achieving domain adaptation. Experimental performances are summarized in Table 1. In terms of these results, our proposed method (HGNet) obtains the best classification accuracy in three datasets. And there exist significant differences between HGNet and CoCoGAN achieving the second best performance. Specifically, our proposed approach surpasses CoCoGAN by 7.7% when Fashion-MNIST and MNIST separately are task-relevant and task-irrelevant datasets. On the one hand, the empirical results provide convincing answer (No) to the question in Section 3.1: is the generation of missing target dataset necessary for learning domain-invariant feature. On the other hand, it illustrates that hybrid generative manner guarantees the uniqueness of feature separation and the application of it enable model to learn more discriminative domain-invariant feature.

For the second step, we conduct transformation between gray-scale modality (*G-domain*) and negative modality (*N-domain*) and summary the corresponding performances in Table 2. From these experimental results, we can obtain three conclusions. First of all, the proposed algorithm (HGNet) achieves more promising performances than other baselines in most cases. Specifically, when separately selecting D_E and D_F as task-relevant and task-irrelevant datasets, our approach outperforms CoCoGAN by 5.8% on the domain adaptation task (*G-domain* $\rightarrow$ *N-domain*). Secondly, we notice that classification accuracy of all mentioned methods on Fashion-MNIST (task-relevant datatset) is lower than that on other two datasets. The main reason for this derives from that trappings images are more complex than digits and letters images. However, HGNet still improve 1%$\sim$3% when compared with the second best result obtained by

Table 3: Classification Accuracy (%) of our method and three baselines for two domain adaptation tasks :*G-domain* $\rightarrow$ *E-domain* and *C-domain* $\rightarrow$ *G-domain*. The best result in each column is in bold.

Task	*G-domain*$\rightarrow$*E-domain*				*C-domain*$\rightarrow$*G-domain*			
RT	MNIST (D_M)		EMNIST (D_E)		MNIST (D_M)		Fashion (D_F)	
IRT	D_F	D_E	D_M	D_F	D_F	D_E	D_M	D_E
ZDDA [21]	72.5	93.2	73.6	50.7	67.4	87.6	55.1	59.5
CoGAN [27]	67.1	81.5	63.6	51.9	54.7	63.5	43.4	51.6
CoCoGAN [27]	79.6	95.4	77.9	58.6	73.2	94.7	61.1	**70.2**
HGNet	**86.5**	**96.1**	**81.1**	**59.5**	**78.9**	**95.0**	**65.9**	68.5

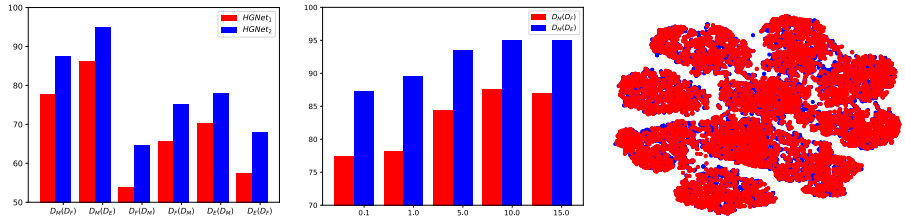
CoCoGAN. Finally, although these two transformation (*G-domain* $\rightarrow$ *N-domain* and *N-domain* $\rightarrow$ *G-domain*) are mutually inverse operations, classification accuracy of most approaches on *G-domain* $\rightarrow$ *N-domain* are better than their performances on *N-domain* $\rightarrow$ *G-domain*. But the results of HGNet on these two transformations are competitive, which means our method has much better generalization.

In the final experiment, we explore *G-domain*$\rightarrow$*E-domain* and *C-domain*$\rightarrow$*G-domain* to further verify the effectiveness of HGNet. Results are reported in Table 3. The performance of HGNet is better than others in most cases. Interestingly, we find that although there exists high similarity between D_M and D_E, it difficult for most methods to achieve great transformation on D_E with the assistance of D_M. Different from them, our method fully utilizes association across all available datasets to reduce cross-domain discrepancy, leading to the improvement on classification accuracy to 81.1%.

4.4 Ablation Study

Effect of Hybrid Strategy: According to the discussion about hybrid reconstruction, we know that this part enable the proposed model to further guarantee the uniqueness of feature separation and promote generalization across various domains by using association of all given datasets. In order to clearly observe the effect of hybrid reconstruction, we firstly attempt to remove this part from our method to form another competed method named as HGNet_1, while the overall version of our method is denoted as HGNet_2. The goal of experiments in this section is to achieve the transformation from *N-domain* to *G-domain* and Figure 2 (a) lists results, where the expression *A(B)* means A is task-relevant dataset while B represents task-irrelevant one.

As seen in Figure 2 (a), the absence of hybrid reconstruction suffers from significant negative influence on the classification accuracy. HGNet_2 outperforms HGNet_1 by 10% for $D_F(D_M)$, illustrating that hybrid strategy not only effec-



(a) (b) (c)

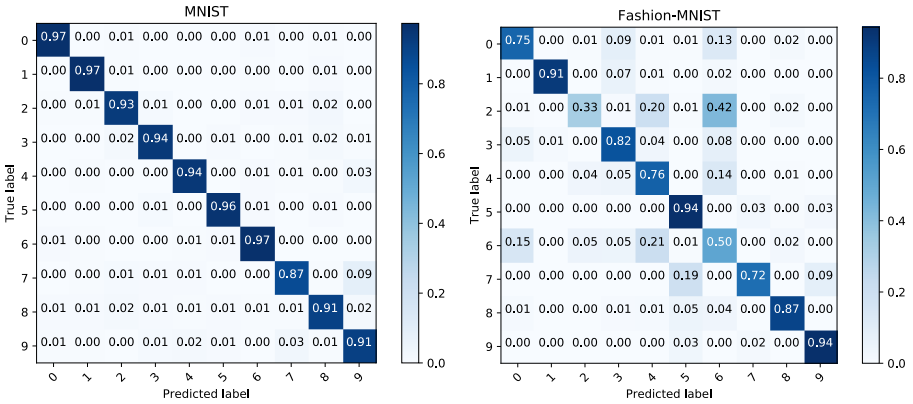
Fig. 2: Experiments are performed on adaptation from *N-domain* to *G-domain*. And the expression *A(B)* means A is the task-relevant dataset while B represents the task-irrelevant one. (a) We denote our proposed method without hybrid reconstruction as $HGNet_1$ and the overall version as $HGNet_2$. (b) We select λ from $\{0.1, 1.0, 5.0, 10.0, 15.0\}$ and observe the classification accuracy. (c) When D_E is the task-irrelevant datasets, we show the feature visualization on MNIST.

tively generates more discriminative feature representation but also captures more cross-domain information from all available data to reduce domain shift.

Additionally, we present the generated images in Figure 3 via hybrid generation to verify its ability performing transformation between source and target domains. In terms of the visualization, we find that hybrid strategy captures cross-domain discrepancy. Specifically, in the first two rows, images synthesised by $\mathcal{G}(h_{cr}^{ir,s}, h_{cir}^{ir,t})$ actually integrate main objects from $X^{ir,s}$ and the corresponding modality style (*N-domain*) from $X^{ir,t}$. It means that our proposed method achieves high-quality separation of semantic information, which assists model to learn domain-invariant feature and promote classification accuracy.

Fig. 3: Visualization of hybrid generation. The first three columns represents the inputs: $X^{r,s}, X^{ir,s}$ and $X^{ir,t}$, while the last four columns are hybrid generative visual signals: $\mathcal{G}_s(h_{cr}^{r,s}, h_{cir}^{ir,s})$, $\mathcal{G}_s(h_{cr}^{ir,s}, h_{cir}^{r,s})$, $\mathcal{G}_t(h_{cr}^{ir,s}, h_{cir}^{ir,t})$ and $\mathcal{G}_s(h_{cr}^{ir,t}, h_{cir}^{ir,s})$.

Parameters Analysis: To show the function of feature alignment on task-irrelevant dataset, we change the value of λ from *0.1* to *15* and record results (*N-*



(a) Confusion Matrix on MNIST (b) Confusion Matrix on Fashion-MNIST

Fig. 4: Visualization of Confusion Matrix. Experiments are performed on adaptation from *N-domain* to *G-domain*. For these two experiments, we select EMNIST as the task-irrelevant datasets.

domain→*G-domain*) in Figure 2 (b). With the increasing of λ, HGNet achieves higher accuracy, illustrating that such feature alignment manner has positive effect on solving the domain shift issue on task-relevant dataset.

Visualization of Latent Space: To further analyse distribution of high-level feature, we draw feature visualization and confusion matrix on MNIST and Fashion-MNIST in Figure 2 (c) and Figure 4. For these experiments, we select EMNIST as task-irrelevant datasets and transfer negative images (*N-domain*) into gray-scale modality (*G-domain*). From the performance, we know that HGNet learns clear boundary between various categories, which significantly promotes feature discriminative.

5 Conclusion

Zero-shot Domain Adaptation (ZSDA) assumes that we hardly access target samples during training stage. To fight off ZSDA more effectively, we propose a novel approach named Hybrid Generative Network (HGNet) including feature extractor, adaptive feature separation module, hybrid generator and classifier. Concretely, feature extractor learns representations from visual signals, and then adaptive feature separation module distinguishes classification-relevant units from classification-irrelevant ones storing meaningless semantic information. Moreover, we adopt two manners to perform high-quality feature separation. One is to use annotation as supervision to generate discriminative feature. Another is to exploit hybrid generative strategy to extract association across various available datasets. Finally, extensive experimental results validate the effectiveness of HGNet on solving ZSDA problem.

References

1. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE transactions on pattern analysis and machine intelligence **33**(5), 898–916 (2010)
2. Cai, R., Li, Z., Wei, P., Qiao, J., Zhang, K., Hao, Z.: Learning disentangled semantic representation for domain adaptation. In: IJCAI. vol. 2019, p. 2060. NIH Public Access (2019)
3. Chen, T.H., Liao, Y.H., Chuang, C.Y., Hsu, W.T., Fu, J., Sun, M.: Show, adapt and tell: Adversarial training of cross-domain image captioner. In: ICCV. pp. 521–530 (2017)
4. Chen, Z., Zhuang, J., Liang, X., Lin, L.: Blending-target domain adaptation by adversarial meta-adaptation networks. In: CVPR. pp. 2248–2257 (2019)
5. Cicek, S., Soatto, S.: Unsupervised domain adaptation via regularized conditional alignment. In: ICCV. pp. 1416–1425 (2019)
6. Cohen, G., Afshar, S., Tapson, J., Van Schaik, A.: Emnist: Extending mnist to handwritten letters. In: IJCNN. pp. 2921–2926. IEEE (2017)
7. Ding, Z., Li, S., Shao, M., Fu, Y.: Graph adaptive knowledge transfer for unsupervised domain adaptation. In: ECCV. pp. 37–52 (2018)
8. Ding, Z., Ming, S., Fu, Y.: Latent low-rank transfer subspace learning for missing modality recognition. In: AAAI (2014)
9. Dong, J., Cong, Y., Sun, G., Zhong, B., Xu, X.: What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In: CVPR (June 2020)
10. Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., Schölkopf, B.: Covariate shift by kernel mean matching. Dataset shift in machine learning **3**(4), 5 (2009)
11. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: CVPR. pp. 4893–4902 (2019)
12. Koniusz, P., Tas, Y., Porikli, F.: Domain adaptation by mixture of alignments of second-or higher-order scatter tensors. In: CVPR. pp. 4478–4487 (2017)
13. Kumar, A., Sattigeri, P., Wadhawan, K., Karlinsky, L., Feris, R., Freeman, B., Wornell, G.: Co-regularized alignment for unsupervised domain adaptation. In: NeurIPS. pp. 9345–9356 (2018)
14. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
15. Liu, H., Long, M., Wang, J., Jordan, M.: Transferable adversarial training: A general approach to adapting deep classifiers. In: ICML. pp. 4013–4022 (2019)
16. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: ICML. pp. 2208–2217. JMLR. org (2017)
17. Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In: CVPR. pp. 2507–2516 (2019)
18. Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Unified deep supervised domain adaptation and generalization. In: ICCV. pp. 5715–5725 (2017)
19. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: ICML. pp. 10–18 (2013)
20. Pei, Z., Cao, Z., Long, M., Wang, J.: Multi-adversarial domain adaptation. In: AAAI (2018)

21. Peng, K.C., Wu, Z., Ernst, J.: Zero-shot deep domain adaptation. In: ECCV. pp. 764–781 (2018)
22. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: ICCV. pp. 1406–1415 (2019)
23. Pinheiro, P.O.: Unsupervised domain adaptation with similarity learning. In: CVPR. pp. 8004–8013 (2018)
24. Roy, S., Siarohin, A., Sangineto, E., Bulo, S.R., Sebe, N., Ricci, E.: Unsupervised domain adaptation using feature-whitening and consensus loss. In: CVPR. pp. 9471–9480 (2019)
25. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR. pp. 3723–3732 (2018)
26. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR. pp. 2517–2526 (2019)
27. Wang, J., Jiang, J.: Conditional coupled generative adversarial networks for zero-shot domain adaptation. In: ICCV. pp. 3375–3384 (2019)
28. Wang, X., Li, L., Ye, W., Long, M., Wang, J.: Transferable attention for domain adaptation. In: AAAI. vol. 33, pp. 5345–5352 (2019)
29. Wulfmeier, M., Bewley, A., Posner, I.: Addressing appearance change in outdoor robotics with adversarial domain adaptation. In: IROS. pp. 1551–1558. IEEE (2017)
30. Xia, H., Ding, Z.: Structure preserving generative cross-domain learning. In: CVPR. pp. 4364–4373 (2020)
31. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
32. Xie, S., Zheng, Z., Chen, L., Chen, C.: Learning semantic representations for unsupervised domain adaptation. In: ICML. pp. 5423–5432 (2018)
33. Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R.: Deconvolutional networks. In: CVPR. pp. 2528–2535. IEEE (2010)
34. Zhang, W., Ouyang, W., Li, W., Xu, D.: Collaborative and adversarial network for unsupervised domain adaptation. In: CVPR. pp. 3801–3809 (2018)
35. Zhang, Y., Tang, H., Jia, K., Tan, M.: Domain-symmetric networks for adversarial domain adaptation. In: CVPR. pp. 5031–5040 (2019)
36. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: ICCV. pp. 2020–2030 (2017)