Sequence-to-Sequence Domain Adaptation Network for Robust Text Image Recognition

Yaping Zhang[1,2], Shuai Nie[1], Wenju Liu[1]*, Xing Xu[3,5], Dongxiang Zhang[4,5], Heng Tao Shen[3]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CASIA)
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[3]Center for Future Media & School of Computer Science and Engineering, University of Electronic Science and Technology of China
[4]College of Computer Science and Technology, Zhejiang University, China [5]Afanti AI Lab, China
{yaping.zhang, shuai.nie, lwj}@nlpr.ia.ac.cn, xing.xu@uestc.edu.cn, zhangdongxiang37@gmail.com, shenhengtao@hotmail.com

Abstract

Domain adaptation has shown promising advances for alleviating domain shift problem. However, recent visual domain adaptation works usually focus on non-sequential object recognition with a global coarse alignment, which is inadequate to transfer effective knowledge for sequence-like text images with variable-length fine-grained character information. In this paper, we develop a Sequence-to-Sequence Domain Adaptation Network (SSDAN) for robust text image recognition, which could exploit unsupervised sequence data by an attention-based sequence encoder-decoder network. In the SSDAN, a gated attention similarity (GAS) unit is introduced to adaptively focus on aligning the distribution of the source and target sequence data in an attended character-level feature space rather than a global coarse alignment. Extensive text recognition experiments show the SSDAN could efficiently transfer sequence knowledge and validate the promising power of the proposed model towards real world applications in various recognition scenarios, including the natural scene text, handwritten text and even mathematical expression recognition.

1. Introduction

Deep learning methods have achieved remarkable results on text image reading [3, 5, 7, 13, 21, 23, 31]. However, it remains challenging to build a robust text recognizer that can handle varying data in new scenarios effectively, due to the inevitable domain shift when the actual data is encountered at "test time". As shown in Figure 1, the text data distribution tends to be changed by multiple factors, such as, the different appearances in natural scene texts [21], various handwriting styles in handwritten texts [3], and even diverse structures in mathematical expressions [7]. To build a robust text recognizer for the shifted target text image, a general solution is to collect large scale annotated text images, while they are high-cost and cannot cover all diversities. However, unsupervised target text images are easily available. If we could take advantage of the unsupervised text images to reduce domain shift, it will be helpful.

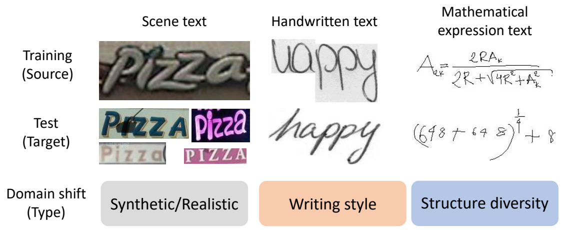

Figure 1. Examples of different types of domain shift in text image recognition scenarios.

Unsupervised domain adaptation is an effective way using the unlabeled target domain data to mitigate the domain shift, which is to align the feature distribution between the source and target domain. Recent research endeavors [30, 38] on domain adaptation have shown the potential results on character recognition. They generally optimize the global representation of a character to minimize some measure of domain shift, such as maximum mean discrepancy (MMD) [24, 38], correlation alignment distance (CORAL) [35, 41], or adversarial loss [9, 30, 36], where feature dimensions are fixed in the source and target domain. However, a text image is the combination of different characters, which is a variable-length label sequence instead of an isolation. Consequently, the most popular domain adaptation methods cannot be directly applied to the sequence prediction, since a global fixed-length representation lacks important fine-grained information at the character level, which in turn cannot appropriately describe the content of sequence-like images.

In this paper, to address the aforementioned issues, we

[*] Corresponding author.

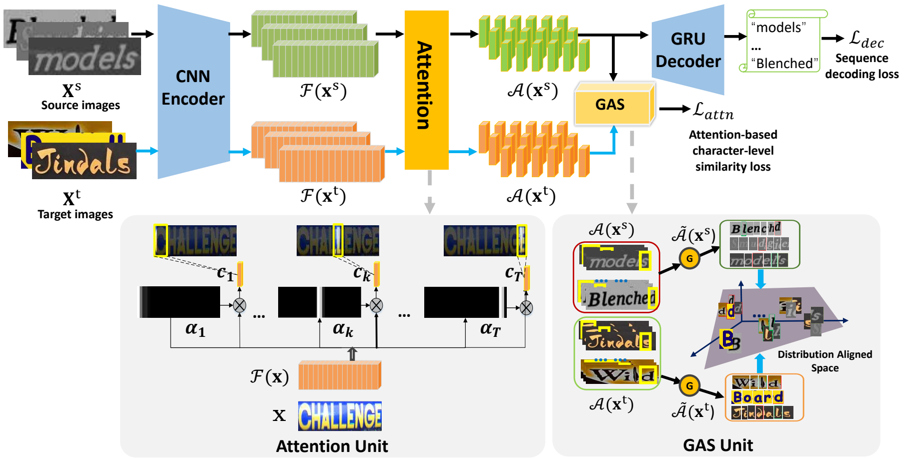
$\mathbf{X}^s$
Source images
$\mathbf{X}^t$
Target images
CNN Encoder
$\mathcal{F}(\mathbf{x}^s)$
$\mathcal{F}(\mathbf{x}^t)$
Attention
$\mathcal{A}(\mathbf{x}^s)$
$\mathcal{A}(\mathbf{x}^t)$
GRU Decoder
"models"
...
"Blenched"
$\mathcal{L}_{dec}$
Sequence decoding loss
GAS
$\mathcal{L}_{attn}$
Attention-based character-level similarity loss
c_1
c_k
c_T
α_1
α_k
α_T
$\mathcal{F}(\mathbf{x})$
$\mathbf{x}$
Attention Unit
$\tilde{\mathcal{A}}(\mathbf{x}^s)$
$\tilde{\mathcal{A}}(\mathbf{x}^t)$
G
Distribution Aligned Space
GAS Unit

Figure 2. The structure of SSDAN consists of: a CNN encoder to map the input images into a sequence of high-level feature vectors , an attention unit between the encoder and decoder to adaptively focus on the location of character, a GRU decoder to convert encoded features into output strings recurrently, and a GAS unit to offer the guidance for model to adaptively find character-level domain-invariant features between the source and target domain. Overall, the unsupervised sequence-to-sequence domain adaptation is achieved by jointly minimizing character-level similarity loss $\mathcal{L}_{attn}$ and source decoding loss $\mathcal{L}_{dec}$.

develop a Sequence-to-Sequence Domain Adaptation Network (SSDAN) for robust text image recognition. As shown in Figure 2, the proposed SSDAN is an attention based encoder-decoder model for handling sequences, which is derived from [7, 21]. It could automatically concentrate on the most relevant region of the character while decoding, which frees a sequence-like text image from having to squash all the information of a source sequence into a global fixed-length vector. Furthermore, a gated attention similarity (GAS) unit is introduced to align distributions of the source and target domain at an attended character-level feature space, where we adopt a gate function to control the model focusing on effective character-level features, instead of global coarse alignment. In GAS unit, an unsupervised character-level similarity loss is used to guide the model to reduce the domain shift between the source and target sequence. The unsupervised sequence-to-sequence domain adaptation is then achieved by jointly minimizing unsupervised character-level similarity loss and supervised source decoding loss, which could learn both domain-invariant and discriminative features that are effective for the shifted target domain.

We summarize our contributions as follows:

- We propose a novel Sequence-to-sequence Domain Adaptation Network dubbed SSDAN for robust text image recognition, which could be generalized to different scenes, such as natural scene text, handwritten text and mathematical expression recognition.
- We introduce a novel GAS unit in SSDAN to bridge the sequence-like text image recognition and domain adaptation, which could adaptively transfer fine-grained character-level knowledge instead of performing domain adaptation by global features.
- The proposed SSDAN is capable of using unsupervised sequence data to reduce domain shift effectively.

Extensive experiments on six benchmark datasets validate the promising power of the proposed model towards large scale real world application in natural scene text, handwritten text and even more difficult mathematical expressions recognition.

2. Related Work

In this section, we review the literature of text recognition methods. Then we discuss the recent trials of applying domain adaptation techniques on text recognition.

Text Recognition Methods. Deep learning methods have achieved remarkable results on image text reading [3, 5, 7, 13, 20, 21, 23, 31]. However, the literature is relatively sparse on building a robust text recognizer that can handle varying data in abundance of scenarios effectively. Some methods were designed to handle perspective distortion exhibited in the scene text. For example, [32] and [22] introduce a spatial transformer network to rectify the entire text before recognition. Furthermore, CharNet [21] tried to introduce a character-level spatial transformer to rectify individual characters, which was capable of handling more complicated forms of distortion that cannot be modeled by a single global transformation easily. However, they were only designed for spatial affine distortions and hard to generalize to the distortion caused by handwriting styles or various structures in mathematical expressions. In summary, existing text image recognition methods are usually designed for a specific scenario, and cannot be generalized effectively to different tasks. While our domain adaptation model is designed for different scenarios, including the nature scene text, handwritten text, and mathematical expression recognition. Furthermore, the intrinsic domain shift in the text image data is commonly neglected in existing methods. On the contrary, our SSDAN utilizes the domain adaptation technique to tackle the domain shift problem, which adaptively performs the character-level adaption in text images.

Domain Adaptation For Text Recognition. There have been a plethora of recent works in the field of visual domain adaptation addressing the domain shift problem [30, 38, 41]. Some methods are evaluated on the character-level handwritten or natural scene digital dataset for recognition tasks and have shown effective performance. However, the majority of recent works use deep convolutional architectures to map the source and target domains into a shared space where the domains are aligned. They generally optimize the global representation via minimizing some measure of domain shift, such as MMD [24, 38], CORAL [35, 41], or adversarial loss [9, 30, 36]. Therefore, these methods cannot be directly applied on sequential text images with multiple characters, as the domain shift are locally in the characters rather than the global image. Recently, other methods have been proposed to adapt the different font styles for image-to-image translation via adversarial learning [1]. Similarly, these methods limitedly translate the font in different style of signal characters on a global image, which are still cannot be extended to text-line images. To address these problems, we develop a sequence-to-sequence domain adaptation to focus on fine-grained character-level features to transfer variable-length sequence knowledge successfully.

3. Proposed Method

In this paper, unsupervised sequence-to-sequence domain adaptation is developed for robust text recognition. Specifically, the source domain text images with well-annotated text labels (a sequence of characters or symbols) are available, while we only have an access to unlabeled text images in target domain, which is in a different distribution. More formally, we assume that there are N^s annotated source domain samples $X^s = \{\mathbf{x}_i^s\}_{i=0}^{N^s}$ with the corresponding labels $\mathcal{Y}^s = \{\mathbf{y}_i^s\}_{i=0}^{N^s}$, and N^t unlabeled target-domain samples $X^t = \{\mathbf{x}_i^t\}_{i=0}^{N^t}$ without any available annotated labels in the training time. For $\mathbf{y} \in \mathcal{Y}^s$, $\mathbf{y} = \{y_1, y_2, ..., y_T\}$, where y_k and T denotes a character label and the variable length of text, respectively.

Considering that typical global domain adaptation methods lack fine-grained character-level information, we develop a Sequence-to-Sequence Domain Adaptation Network (SSDAN) for robust text image recognition, aligning the distribution of the source and target sequence data in an attended character-level feature space rather than a global coarse alignment. As shown in Figure 2, the proposed SSDAN is an attention-based sequence encoder-decoder network, which encodes a text image into a sequence of attended character-level features that are then recomposed through a GRU decoder with an attention mechanism. In the proposed SSDAN, a GAS unit is further introduced to adaptively guide model finding the character-level domain-invariant features between the source and target domain.

3.1. Attentive Text Recognition

The attentive text recognition can be essentially considered as learning a mapping between a sequence of feature maps encoded from sequence-like text image $\mathbf{x}$, and a ground truth label sequence $\mathbf{y} = \{y_1, y_2, ..., y_T\}$. As shown in Figure 2, the attentive text recognition pipeline consists of: 1) a CNN encoder that learns high-level visual representations from an input image. 2) an attention model between the encoder and the decoder driving the focus of attention of the model towards a specific part of the sequence of encoded features. 3) a GRU decoder that generates a sequence of symbols as output, one at every time step.

CNN Encoder. CNN encoder $\mathcal{F}$ takes the raw input image $\mathbf{x}$ from the source or target domain, and produces a feature grid $\mathcal{F}(\mathbf{x})$ of size $H' \times W' \times D$, where D denotes the number of channels, H' and W' are the resulted feature map height and width, respectively. The encoder output is then reshaped as a grid sequence of L elements, $L = H' \times W'$. Each of these elements is a D-dimensional feature vector that corresponds to a local region of the image through its corresponding receptive field. Hence, the whole encoded image $\mathcal{F}(\mathbf{x})$ could be reformatted as,

$$\mathcal{F}(\mathbf{x}) = [\mathbf{f}_1, ..., \mathbf{f}_L], \mathbf{f}_i \in R^D, \tag{1}$$

where $\mathbf{f}_i$ corresponds to i-th grid of the encoded image $\mathcal{F}(\mathbf{x})$, which preserves specific spatial information of the input image $\mathbf{x}$.

Attention. Although the CNN encoder keeps the spatial information, we cannot decide the location of a specific character in a text image. Therefore, an attention model is introduced to learn which part of the text image is the most relevant to a decoding character. As shown in Figure 2, the attention is a T-step process, at time-step k, the representation of the most relevant part to character y_k of encoding feature map $\mathcal{F}(\mathbf{x})$ is defined as a context vector $\mathbf{c}_k$:

$$\mathbf{c}_k = \sum_{i=0}^{L} \alpha_{k,i} \mathbf{f}_i, \tag{2}$$

where, the attention weights $\alpha_{k,i}$ is calculated by

$$\alpha_{k,i} = \frac{\exp(\mathbf{s}_{k,i})}{\sum_{j=0}^{L} \exp(\mathbf{s}_{k,j})}, \tag{3}$$

where the attention score $\mathbf{s}_{k,i}$ indicates the probability of that the model attends to the i-th sub-region in the encoded map $\mathcal{F}(\mathbf{x})$ when decoding the k-th character of the text image. Following the past empirical work [7], we defined the attention score as

$$\mathbf{s}_{k,i} = \beta^{\top} \tanh(\mathbf{W}_h \mathbf{h}_{k-1} + \mathbf{W}_f \mathbf{f}_i), \tag{4}$$

where β, $\mathbf{W}_h$ and $\mathbf{W}_f$ are the parameters to be learnt, $\mathbf{h}_{k-1}$ is the previous decoding state in the decoder.

GRU Decoder. A GRU decoder is employed to predict the string of an input text image recurrently, where we use gated recurrent unit (GRU) neural network. At decoding time step k, the GRU leverages the context vector $\mathbf{c}_k$, previous state $\mathbf{h}_{k-1}$ and previous predicted character y_{k-1} to generate a new hidden state

$$\mathbf{h}_k = GRU(\mathbf{h}_{k-1}, y_{k-1}, \mathbf{c}_k), \tag{5}$$

where, $\mathbf{c}_k$ is generated by the attention mechanism, which focuses on the most relevant region of current decoding character. Then, the probability of current predicted symbol y_k is computed by :

$$p(y_k|y_{k-1}, \mathbf{c}_k) = g\left(\mathbf{W}_o \tanh(\mathbf{E}\tilde{\mathbf{y}}_{k-1} + \mathbf{W}_d\mathbf{h}_k + \mathbf{W}_c\mathbf{c}_k)\right), \tag{6}$$

where g denotes a softmax activation function, $\mathbf{W}_o$, $\mathbf{W}_d$ and $\mathbf{W}_c$ are the mapping matrices, $\mathbf{E}$ is the embedding matrix, and $\tilde{\mathbf{y}}_{k-1}$ is the one-hot vector of character label y_{k-1}.

The probability of the sequential labels $\mathbf{y}$ is finally given by the product of the probability of each label:

$$P(\mathbf{y}|\mathcal{A}(\mathbf{x})) = \prod_{k=1}^{T} p(y_k|y_{k-1}, \mathbf{c}_k), \tag{7}$$

where $\mathcal{A}(\mathbf{x}) = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_T\}$, which could be regarded as a sequence of attended character-level features from an input text image $\mathbf{x}$.

3.2. Gated Attention Similarity Unit

Given the misalignment of ground truth strings between the source and target sequence domain, we introduced Gated Attention Similarity (GAS) Unit, based on an attention encoder-decoder mechanism, to convert a variable-length input text image into a sequence of character features. By decomposing the text strings into a set of characters, the source and target domain will statistically share the same label space in character-level, and thus the influence of the misalignment problem can be alleviated. More formally, through attention mechanism, an input image $\mathbf{x}$ can be adaptively decomposed into a series of character-level feature set $\mathcal{A}(\mathbf{x}) = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_T\}$, where $\mathbf{c}_k$ presents the feature of k-th character in the text image $\mathbf{x}$. Specifically, a source text image $\mathbf{x}^s$ and a target text image $\mathbf{x}^t$ are decomposed into a source and target attended character-level feature set $\mathcal{A}(\mathbf{x^s})$ and $\mathcal{A}(\mathbf{x^t})$, respectively.

We notice that if the attention context vector fails to focus on the region of effective character, the adaptation on the attention context vector will not help. To overcome this problem, we introduce a gate mechanism to select effective attention context vectors to perform domain adaptation. An intuition is that if the current attention context vector $\mathbf{c}_k$ is distinguishable, the probability that $\mathbf{c}_k$ belongs to one specific character y_k will be relatively higher than others. Hence, we further introduce an adaption gate function $\delta(\mathbf{c}_k)$ to judge if a context vector $\mathbf{c}_k$ is attending to a valid character,

$$\delta(\mathbf{c}_k) = \begin{cases} 1 & \text{if } p\left(y_k|y_{k-1}, \mathbf{c}_k\right) > p_c \\ 0 & \text{if } p\left(y_k|y_{k-1}, \mathbf{c}_k\right) < p_c \end{cases}, \tag{8}$$

where p_c is a confidence threshold. Furthermore, a gate function set $\mathbf{G}$ is adaptively changed according to the specific input image $\mathbf{x}$, which is expressed as:

$$\mathbf{G}(\mathbf{x}) = \left\{\delta(\mathbf{c}_1), ..., \delta(\mathbf{c}_T)\right\}, \tag{9}$$

Through the gate function, we can update attention context vector set by adaptation gate function set $\mathbf{G}(\mathbf{x})$,

$$\tilde{\mathcal{A}}(\mathbf{x}) = \mathcal{A}(\mathbf{x}) \otimes \mathbf{G}(\mathbf{x}), \tag{10}$$

where $\otimes$ denotes element-wise product operator. Specifically, if $\mathbf{c}_k \times \delta(\mathbf{c}_k) = 0$, then current context vector $\mathbf{c}_k$ will not be added in a new attention context vector set.

A gated attention similarity loss $\mathcal{L}_{attn}$ is accordingly introduced to measure the distance on the valid attended character-level feature set of source and target domain as

$$\mathcal{L}_{attn} = E_{[\mathbf{x}^s \in \mathbf{X}^s, \mathbf{x}^t \in \mathbf{X}^t]} \left\{ dist\left(\tilde{\mathcal{A}}(\mathbf{x}^s), \tilde{\mathcal{A}}(\mathbf{x}^t)\right)\right\}. \tag{11}$$

There are multiple choices for the distance function $dist$, such as (1) MMD [24] computing the norm of difference between two domain means, (2) CORAL [35] computing the distance of covariance of two domain, or even (3) adversarial loss [9] minimizing the loss of a domain classifier to learn a representation that is simultaneously discriminative of source labels while not being able to distinguish between domains. In the experiment, we have explored these different measurements, and experimentally found that CORAL is more appropriate for our model. Specifically, CORAL is to align the second-order statistics-correlation of the source and target data , which is defined as

$$dist(\mathcal{U}_s, \mathcal{U}_t) = \frac{1}{4d^2}||cov(\mathcal{U}_s) - cov(\mathcal{U}_t)||_F^2, \tag{12}$$

where $\mathcal{U}_s = \{\mathbf{u}_i^s\}$, $\mathbf{u}^s \in \mathcal{R}^d$, $\mathcal{U}_t = \{\mathbf{u}_i^t\}$, $\mathbf{u}^t \in \mathcal{R}^d$, and $||\cdot||_F^2$ denotes the squared matrix Frobenius norm, $cov(\mathcal{U}_s)$ is the covariance matrix of samples $\mathcal{U}_s$, denoted by

$$cov(\mathcal{U}_s) = \frac{1}{N-1}\left(\mathcal{U}_s{}^\top\mathcal{U}_s - \frac{1}{N}(\mathbf{1}^\top\mathcal{U}_s)^\top(\mathbf{1}^\top\mathcal{U}_s)\right), \tag{13}$$

where $\mathbf{1}$ is a column vector with all elements equal to 1, N is the number of samples $\mathcal{U}_s$, and $\mathcal{U}_s(i,j)$ ($\mathcal{U}_t(i,j)$) indicates the j-th dimension of the i-th source (target) data example.

In our GAS unit, $\mathcal{U}_s$ and $\mathcal{U}_t$ are replaced by the valid attended character-level feature set $\tilde{\mathcal{A}}(\mathbf{x}^s)$ and $\tilde{\mathcal{A}}(\mathbf{x}^t)$, respectively. Note that $\tilde{\mathcal{A}}(\mathbf{x}^s)$ and $\tilde{\mathcal{A}}(\mathbf{x}^t)$ need to be reformatted as a matrix, respectively. Specifically, suppose $\tilde{\mathcal{A}}(\mathbf{x}^s)$ and $\tilde{\mathcal{A}}(\mathbf{x}^t)$ contain T_1 and T_2 elements, respectively. Then $\tilde{\mathcal{A}}(\mathbf{x}^s)$ and $\tilde{\mathcal{A}}(\mathbf{x}^t)$ could be reformatted as matrices with $T_1 \times D$ and $T_2 \times D$ elements, and their covariance matrices are with the same dimension $D \times D$.

3.3. Overall Objective Function

With the well-annotated source-domain data, we could learn an optimized source text image recognizer by minimizing a supervised decoding loss, where we can use the negative log likelihood of sequential probability as the decoding loss $\mathcal{L}_{dec}$ to measure the differences between the predicted and the source labeled character sequences:

$$\mathcal{L}_{dec} = E_{(\mathbf{x}^s,\mathbf{y}^s)\sim(X^s,\mathcal{Y}^s)} \left\{-\log p(\mathbf{y}^s|\mathcal{A}(\mathbf{x}^s))\right\}. \tag{14}$$

Directly optimizing $\mathcal{L}_{dec}$ may cause overfitting in source domain, and thus fails to perform well for the shifted target domain. The GAS unit in our model is introduced to offer guidance to learn domain-invariant features between the source and target domain. The learnt robust representations should work effectively on the target domain, where they are also required to be discriminative. Therefore, the attention similarity loss $\mathcal{L}_{attn}$ in Eq. 11 is combined with the discriminative decoder loss L_{dec} in source domain. The overall objective function of the attentional domain adaptation model is defined as:

$$\mathcal{L}_{SSDAN} = \mathcal{L}_{dec} + \lambda\mathcal{L}_{attn}, \tag{15}$$

where λ is a hyper-parameter to balance two terms. The model parameters can be directly optimized by minimizing the overall objective through stochastic gradient descent optimization algorithms.

4. Experiments

Datasets. We conduct extensive experiments to validate the proposed SSDAN on six general recognition benchmark datasets, including three different types of text image, *i.e.*, scene text, handwritten text, and mathematical expressions with more complex structure, as shown in Figure 1.

- **ICDAR-2003 (IC-03)** [25] contains 860 cropped scene text images, following the protocol used in [31].
- **ICDAR-2013 (IC-13)** [18] contains 857 cropped scene images after filtering as did in IC-03.
- **Street View Text (SVT)** [37] consists of 647 test scene word images from Google Street View.
- **IIIT5K-words (IIIT5K)** [27] contains $3,000$ cropped test scene text images from the Internet.
- **IAM** [26] is a handwritten English text dataset, written by 657 different writers. It is partitioned into writer-independent training, validation and test partitions of 6161, 976 and 2915 lines, respectively. That contains a total of 46945, 7554 and 20306 correctly segmented words in each partition.
- **CROHME 2014** [28] is a handwritten mathematical expression dataset. It contains 8836 training and 986 test math expressions. There are 101 math symbols. The handwritten expressions or LaTeX notations in the test set never appear in the train set.

Evaluation Metric. For different recognition task, we adopt different evaluation metric as follows:

- **Scene text.** The word prediction accuracy is used to evaluate scene text recognition model, following several benchmark [21, 31].
- **Handwritten text.** Two metrics are used to evaluate the handwritten text recognition model: the Character Error Rate (CER) and the Word Error Rate (WER) [3, 34]. CER is defined as the Levenstein distance between the predicted and real character sequence of the word. WER denotes the percentage of words improperly recognized. For CER and WER, small values indicate better performance.
- **Mathematical expression.** We use a global performance metric expression recognition rate (ExpRate) to denote the percentage of predicted formula sequences matching the real formula sequences [7].

Implementation Details. The architecture of the CNN encoder is derived from the DenseNet [14], where dense blocks are densely concatenation of 1×1 convolution layers and 3×3 convolution layers. while the transition layers are composed of 1×1 convolution and 2×2 average pooling, and the channel 0.5 refers to the compression rate, which is to reduce the number of feature map of each block to half. All convolutions are followed by batch normalization layer [15] and rectified linear unit (Relu) activation function [29] . In order to make the encoder suitable for recognizing text, we use the pooling layer with kernel size 2×1 to reduce feature dimension along the height axis only. As a result, the resolution of feature maps produced by encoder is $H/32 \times W/4$, where the values of H and W are set according to the specific dataset. After the CNN encoder, we use a bi-directional LSTM to capture more context information for attention, and each LSTM has 256 hidden units. For the decoder, we use a GRU cell with 512 memory blocks.

All of our experiments are implemented with Tensorflow. The complete model is initially pre-trained to minimize the decoding loss of the source training data, and then is fine-tuned to minimize the overall domain adaptation objective with unsupervised target data. The model is trained with the Adadelta optimizer [39].

4.1. Comparison with Existing Methods

In this section, we investigate the generalization of our model in three different domains, including scene text, handwritten text and mathematical expressions. To validate the performance of our SSDAN model, we focus on unconstrained text recognition without any language model or lexicon. On each task , we consider a baseline for SSDAN as *SSDAN-base* that omits the GAS unit to switch off the domain adaption process. SSDAN-base is used to investigate the capability of SSDAN for domain adaptation on the text image recognition task.

Results on Scene Text. In this scenario, we explore the capability of SSDAN for domain adaptation on the scene text recognition, where synthetic dataset MJSYNTH [17] is used as the source training data, and the real scene text data is used as target test data. MJSYNTH [17] contains 8 millions annotated synthetic images, which are generated to simulate natural scene text images. Table 1 presents the test results on four real scene text datasets. Compared to the baseline model SSDAN-base, our SSDAN method could obtain consistent improvement in different datasets. It's mainly attributed to sequence-to-sequence domain adaptation, which is able to learn more domain-invariant features. Furthermore, we investigate the performance of our model among the recent state-of-the-art approaches [10, 11, 16, 20–22, 33], which are tailored for scene text recognition. We can observe that the performance of our baseline SSDAN-base are at average level. However, the SSDAN model with sequence domain adaptation can achieve comparable results with the best competitor [21, 33]. It's notable that the motivations in our method are substantially different from these works. For example, RARE [32], STAR-Net [22], ASTER [33] and Char-Net [21] target the irregular scene text recognition, which are designed for spatial distortions. They would not be easily generalized to different distortions, such as various handwriting style and complex structures in mathematical expressions. In contrast, our method aims to perform sequence-to-sequence domain adaptation to reduce the domain shift, and correspondingly allows us to relieve different distortions using a general framework in different scenarios.

Table 1. Scene text recognition accuracies on general scene text recognition benchmarks.

Model	IIIT5K	SVT	IC-03	IC-13
ANN [16]	–	71.7	89.6	81.8
STAR-Net [22]	83.3	83.6	89.9	89.1
R^2AM [20]	78.4	80.7	88.7	90.0
CRNN [31]	81.2	82.7	91.9	89.6
RARE [32]	81.9	81.9	90.1	88.6
Ghosh et al [12]	–	75.1	89.3	–
Gao et al [10]	81.8	82.7	89.2	88.0
ASTER [33]	83.2	**87.6**	**92.4**	89.7
Char-Net [21]	83.6	84.4	91.5	90.8
SSDAN-base	81.1	82.1	91.2	91.0
SSDAN	**83.8**	84.5	92.1	**91.8**

Results on Handwritten Text. To verify the generalization capability of our model, we evaluate our model on IAM to validate the effectiveness of the sequence-to-sequence domain adaptation on the handwriting recognition. In this case, the source and target data are the writer-independent training and test data, respectively. Various handwriting styles are primary causes of domain shift. What's more, it may suffer character-touching problem, which is different from scene text. We note that [6] achieved a state-of-the-art performance on IAM, however, the test data used in [6] wasn't same with [2, 3, 34]. For a fair comparison, we only show the results using the same test data and without any language model. Table 2 illustrates the handwriting recognition results. Although the performance of our baseline is not better than sueiras2018offline [34], our SSDAN model can still achieve significant improvement, which demonstrates the effectivity of model.

Table 2. Results on handwritten text.

Method	WER	CER	Average
bluche2015deep [2]	24.7	**7.3**	16.00
bluche2016joint [3]	24.6	7.9	16.25
sueiras2018offline [34]	23.8	8.8	16.30
SSDAN-base	23.9	9.2	16.55
SSDAN	**22.2**	8.5	**15.35**

Results on Handwritten Mathematical Expression. To show the flexibility of our model, we also conduct experiments on handwritten mathematical expression recognition. Handwritten mathematical expression recognition is to convert an image into structured language, such as LaTex strings, which not only denotes the text itself but also denotes its structural information. It's a more complex problem than traditional scene text or handwriting recognition. In particular, it suffers variant scales of handwritten math symbols with more complicated structure, which results in more difficult domain shift. In the experiment, the training expressions and the unseen test expressions are used as source and target data, respectively. The results of the expression recognition rate (ExpRate) are listed in Table 3. We note that the WAP approach [40] achieved the state-of-the-art result, which involved 5 ensemble models to improve the performance. However, our model doesn't use any ensemble trick, which might be investigated as the future work. Compared with the best 3 systems in CROHME 2014 competition, which only use the CROHME training data, our model obviously outperforms these participating systems with large gaps. Furthermore, compared to [7, 8, 19], our SSDAN model can still achieve better performance. It's remarkable that our model is competitive, especially after domain adaptation. Hence, we can believe that our SSDAN model is able to capture the complex domain shift in structural images.

4.2. Ablation Study

We firstly evaluate the necessity of character-level domain adaptation. Then we analyze the contributions of different components, and investigate the effect of different

Table 3. Results on handwritten mathematical expression.

Method	ExpRate
I [28]	37.2
VI [28]	25.7
VII [28]	26.1
WYCIWYS [8]	28.7
Le et al [19]	35.2
IM2TEX [7]	38.7
SSDAN-base	39.9
SSDAN	**41.6**

domain shift measurement and parameter sensitivity. Furthermore, we visualize some recognition results, and explore the effect of unsupervised data.

Comparison to Standard Domain Adaptation. Measuring similarity on CNN outputs directly can be treated as a STandard Domain Adaptation method (STDA), which lacks of fine-grained character-level information in a text image. Our SSDAN method introduces a GAS unit to adaptively perform domain adaptation on a set of character-level feature vectors via attention scheme, which focuses on most relevant region towards a specific character instead of global CNN outputs. To validate the the necessity of character-level domain adaptation, we have done experiment to compare the STDA with our SSDAN on IAM dataset. The results in Table 4 show that STDA obtains worse results than the baseline SSDAN-base, while SSDAN gets significant improvement. It validates the advantages of our SSDAN that the fine-grained character-level knowledge transfer between the source and target sequence data is more effective in the decoder than the CNN outputs.

Table 4. Comparison to standard domain adaptation.

Method	SSDAN-base	SSDAN	STDA
WER	23.9	**22.2**	25.0
CER	9.2	**8.5**	11.1

Component Analysis. In this scenario, we evaluate the contribution of different components of the proposed model. These variants include: 1) using different CNN encoder to investigate the contribution of encoder, *i.e.*, V1, V3 and V5 using VGG-based (VGG) [31], ResNet-based (ResNet) [4], and DenseNet encoder, respectively; 2) introducing the GAS unit for different encoder to evaluate the effect of GAS unit among different encoders, *i.e.*, V2, V4, and V6 are developed based on the V1, V3, and V5, respectively. For the analysis, we choose handwritten text dataset IAM to evaluate model from both CER and WER, and all the experiments are on the same training protocol. Table 5 presents the comparison between the variants of our model. Firstly, we can observe that DenseNet is a more powerful encoder from the comparisons among the model V1, V3, and V5. Furthermore, the comparison pairs (V1, V2), (V3, V4) and (V5, V6) show that the GAS unit could always improve performance despite of the types of encoders, which demonstrates that considering the sequence-to-sequence domain adaptation makes sense.

Table 5. Component Analysis.

Components	Model	V1	V2	V3	V4	V5	V6
Encoder	VGG	✓	✓				
	ResNet			✓	✓		
	DenseNet					✓	✓
Adaptation	GAS		✓		✓		✓
Evaluation	WER	32.8	26.9	29.9	27.9	23.9	**22.2**
	CER	15.9	12.6	14.3	13.1	9.2	**8.5**

Effect of Different Domain Shift Measurement. Our SSDAN learns domain invariant representations by minimizing some measure of domain shift between the distributions of attended fine-grained character-level features from the source and target text images. In this scenario, we investigate the effect of different domain shift measurement among CORAL, MMD and adversarial loss (Adversarial). The adversarial loss is measured by an extra domain classifier, which is a single layer fully-connected network with 128 hidden units. Specifically, adversarial loss based method needs to minimize the adversarial loss with respect to parameters specific to the domain classifier, while maximizing it with respect to the parameters of text image recognizer. To unify the training procedure in a single step, we use a a gradient reversal layer [9] for the minimax optimization. Table 6 shows the results using different measurement of domain shift on the IAM dataset. We can observe that CORAL-based method outperforms the MMD-based method and adversarial loss-based method. This may show that CORAL is more appropriate for adapting attended fine-grained character-level features.

Table 6. Effect of Different Domain Shift Measurement.

Method	CORAL	MMD	Adversarial
WER	**22.2**	22.7	24.6
CER	**8.5**	8.8	10.8

Parameter Sensitive Analysis. In this subsection, we evaluate the sensitiveness of the hyper-parameter p_c and λ in the Eq. 9 and Eq. 15, respectively. Here, we conduct the experiments on the MJSYNTH → SVT task. Specifically, we explore the different λ and p_c from $\{0.01, 0.1, 1, 10\}$ and $\{0, 0.1, 0.2, 0.4, 0.8\}$, respectively. The evaluation is conducted by changing one parameter while keeping the other hyper-parameters fixed. The λ in the objective function Eq. 15 controls the contribution of sequence domain adaptation. $\lambda = 0$ indicates the proposed model switching off the sequence domain adaptation, which equals to the SSDAN-base. While $\lambda > 0$ means to perform domain adaptation. Furthermore, the p_c in the gate function of Eq. 9

decides whether an attended feature performs domain adaptation or not. Specifically, if the probability that the current feature vector belongs to a valid character is larger than p_c, the vector will be performed domain adaptation, otherwise, it will be neglected as a noise. From other perspective, if $p_c = 0$, the gate function will not work, which means performing sequence domain adaptation on character-level feature without any guidance. While p_c is too large, the gate function will be too strict to select enough valid features. Figure 3 shows different gains of p_c values, where $\lambda = 1$. The results experimentally prove that the gate function is important to the overall performance.

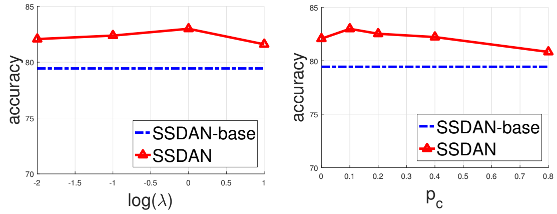

Figure 3. The effect of model parameters λ (left) and p_c (right).

Visualization. In this section, we visualized some recognition results from IAM. The results are shown in the pair of attention visualization and prediction text. The selected attention visualization shows the attending location at one specific time, where the SSDAN-base model suffers recognition error. As shown in Figure 4, while SSDAN-base failed to deal with the distortion of individual character caused by handwriting style, SSDAN successfully worked. As the first two cases shown in Figure 4, even though the SSDAN-base and SSDAN model attend to the same location at one specific time, SSDAN could achieve a better performance through alleviating the domain shift. More interestingly, we find the SSDAN model can learn more precise alignment, according to the last two cases in Figure 4. These results again validate the effectiveness of SSDAN.

Effect of Unsupervised Data. In order to quantify the effectiveness of unsupervised data, we train our model with different size of labeled data and unlabeled data, while keep other hyper-parameters fixed. Figure 5 presents the results with different data size. Firstly, we observe the SSDAN-base model, which is a full supervised learning, with different amounts of labeled samples randomly sampling from the MJSYNTH dataset. The more labeled samples are used, the higher accuracies on real test datasets get. It's notable that using additional unlabeled samples can get consistent performance improvement by SSDAN, where the size of unsupervised data is in accordance with the amount of labeled data. It indicates that our SSDAN is able to learn the knowledge from unsupervised data. We also observe that our model could get significant improvement when available annotated data is small.

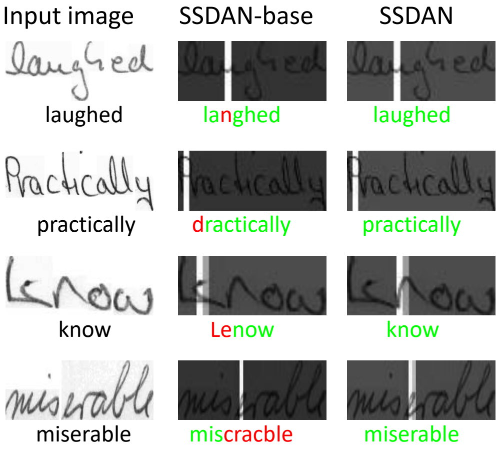

Figure 4. Examples showing the recognition result, the left column is the input image with ground truth, the second column and the last column denote the recognition result without and with domain adaptation, respectively. Each result is shown in the pair of attention visualization and prediction text.

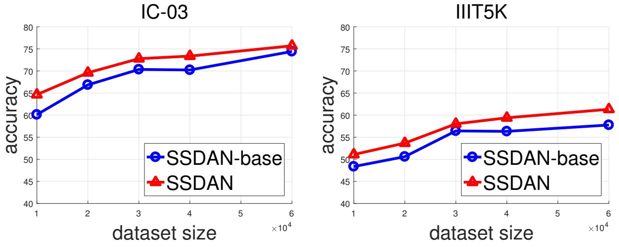

Figure 5. The effect of training dataset size on IC-03 (left) and IIIT5K (right).

5. Conclusion

In this paper, we present a novel SSDAN model for robust text image recognition, which bridges the sequence-like text image recognition and domain adaptation. It's capable of taking advantage of unsupervised sequence data to learn more robust representations. The proposed model could also be generalized to different scenes, which include scene text, handwritten text and mathematical expression recognition. Comprehensive experimental results on several datasets and extensive analyses have demonstrated the effectiveness of our algorithm. An interesting open issue for future research is to further adjust SSDAN framework to better deal with various sequence domain shift.

Acknowledgements. The first and the fourth authors have equal contribution in this work. This work was supported by the China National Nature Science Foundation with following No. 61573357, No. 61503382, No. 61602089, No. 61403370, No. 61273267 and No. 91120303; Project of Xueersi Online School (No. Y9D2M10101).

References

[1] Samaneh Azadi, Matthew Fisher, Vladimir Kim, Zhaowen Wang, Eli Shechtman, and Trevor Darrell. Multi-content gan for few-shot font style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 11, page 13, 2018.

[2] Théodore Bluche. *Deep neural networks for large vocabulary handwritten text recognition*. PhD thesis, Université Paris Sud-Paris XI, 2015.

[3] Théodore Bluche. Joint line segmentation and transcription for end-to-end handwritten paragraph recognition. In *Advances in Neural Information Processing Systems*, pages 838–846, 2016.

[4] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 5086–5094. IEEE, 2017.

[5] Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Aon: Towards arbitrarily-oriented text recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] Arindam Chowdhury and Lovekesh Vig. An efficient end-to-end neural model for handwritten text recognition. In *BMVC*, 2018.

[7] Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M Rush. Image-to-markup generation with coarse-to-fine attention. In *International Conference on Machine Learning*, pages 980–989, 2017.

[8] Yuntian Deng, Anssi Kanervisto, and Alexander M Rush. What you get is what you see: A visual markup decompiler. *arXiv preprint arXiv*, 1609, 2016.

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 1180–1189. JMLR. org, 2015.

[10] Yunze Gao, Yingying Chen, Jinqiao Wang, and Hanqing Lu. Reading scene text with attention convolutional sequence modeling. *arXiv preprint arXiv:1709.04303*, 2017.

[11] Yuting Gao, Zheng Huang, and Yuchen Dai. Double supervised network with attention mechanism for scene text recognition. *arXiv preprint arXiv:1808.00677*, 2018.

[12] Suman K Ghosh, Ernest Valveny, and Andrew D Bagdanov. Visual attention models for scene text recognition. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2017.

[13] Pan He, Weilin Huang, Yu Qiao, Chen Change Loy, and Xiaoou Tang. Reading scene text in deep convolutional sequences. In *AAAI*, volume 16, pages 3501–3508, 2016.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017.

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 448–456. JMLR. org, 2015.

[16] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep structured output learning for unconstrained text recognition. *arXiv preprint arXiv:1412.5903*, 2014.

[17] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014.

[18] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 1484–1493. IEEE, 2013.

[19] Anh Duc Le and Masaki Nakagawa. Training an end-to-end system for handwritten mathematical expression recognition by generated patterns. In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, volume 1, pages 1056–1061. IEEE, 2017.

[20] Chen-Yu Lee and Simon Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2231–2239, 2016.

[21] Wei Liu, Chaofeng Chen, and Kwan-Yee K Wong. Charnet: A character-aware neural network for distorted scene text recognition. In *AAAI Conference on Artificial Intelligence*. New Orleans, Louisiana, USA, 2018.

[22] Wei Liu, Chaofeng Chen, Kwan-Yee K Wong, Zhizhong Su, and Junyu Han. Star-net: A spatial attention residue network for scene text recognition. In *BMVC*, volume 2, page 7, 2016.

[23] Zichuan Liu, Yixing Li, Fengbo Ren, Wang Ling Goh, and Hao Yu. Squeezedtext: A real-time scene text recognition by binary convolutional encoder-decoder network, 2018.

[24] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.

[25] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, et al. Icdar 2003 robust reading competitions: entries, results, and future directions. *International Journal of Document Analysis and Recognition (IJDAR)*, 7(2-3):105–122, 2005.

[26] U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46, 2002.

[27] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC-British Machine Vision Conference*. BMVA, 2012.

[28] Harold Mouchere, Christian Viard-Gaudin, Richard Zanibbi, and Utpal Garain. Icfhr 2014 competition on recognition of on-line handwritten mathematical expressions (crohme 2014). In *Frontiers in handwriting recognition (icfhr), 2014 14th international conference on*, pages 791–796. IEEE, 2014.

[29] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[30] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *AAAI Con-*

ference on Artificial Intelligence, 2018.

[31] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2017.

[32] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4168–4176, 2016.

[33] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[34] Jorge Sueiras, Victoria Ruiz, Angel Sanchez, and Jose F Velez. Offline continuous handwriting recognition using sequence to sequence neural networks. *Neurocomputing*, 289:119–128, 2018.

[35] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016.

[36] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4, 2017.

[37] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1457–1464. IEEE, 2011.

[38] Baoyao Yang, Andy Jinhua Ma, and Pong C Yuen. Domain-shared group-sparse dictionary learning for unsupervised domain adaptation. In *AAAI*, 2018.

[39] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[40] Jianshu Zhang, Jun Du, Shiliang Zhang, Dan Liu, Yulong Hu, Jinshui Hu, Si Wei, and Lirong Dai. Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. *Pattern Recognition*, 71:196–206, 2017.

[41] Junbao Zhuo, Shuhui Wang, Weigang Zhang, and Qingming Huang. Deep unsupervised convolutional domain adaptation. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 261–269. ACM, 2017.