Handwritten Text Segmentation in Scribbled Document via Unsupervised Domain Adaptation

Junho Jo*, Jae Woong Soh*, and Nam Ik Cho*
* Department of ECE, INMC, Seoul National University, Seoul, Korea
E-mail: {jottue, soh90815}@ispl.snu.ac.kr; nicho@snu.ac.kr

***Abstract*—Supervised learning methods have shown promising results for the handwritten text segmentation in scribbled documents. However, many previous methods have handled the problem as a connected component analysis due to the extreme difficulty of pixel-level annotations. Although there is an approach to solve this problem by using synthetically generated data, the resultant model does not generalize well to real scribbled documents due to the domain gap between the real and synthetic dataset. To alleviate the problems, we propose an unsupervised domain adaptation strategy for the pixel-level handwritten text segmentation. This is accomplished by employing an adversarial discriminative model to align the source and target distribution in the feature space, incorporating entropy minimization loss to make the model discriminative even for the unlabeled target data. Experimental results show that the proposed method outperforms the baseline network both quantitatively and qualitatively. Specifically, the proposed adaptation strategy mitigates the domain shift problem very well, showing the improvement of baseline performance (IoU) from 64.617% to 85.642%.**

***Index Terms*—handwritten text segmentation, synthetic dataset, unsupervised domain adaptation, domain shift, entropy minimization**

I. Introduction

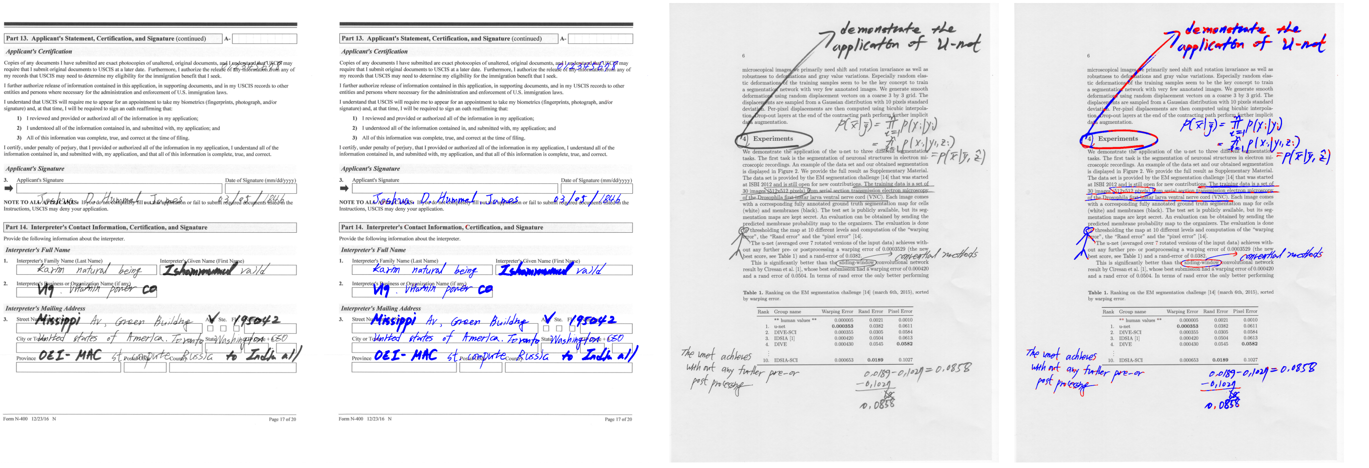
Convolutional Neural Networks (CNNs) have drastically innovated the field of computer vision, achieving the best performance in a multitude of tasks such as image classification [1], semantic segmentation [2], object detection [3], *etc*. The prerequisite of these successful results is the availability of abundant labeled training data for supervision. However, such abundances are a privilege only for certain well-known areas [1]–[3], and there are barren areas still under construction of dataset for the introduction of CNN. Additionally, obtaining annotated data remains a cumbersome and expensive process in the majority of applications. Semantic segmentation is one such task that requires great human effort as it involves annotating dense pixel-level labels [4]–[6]. Even with a lot of effort to complete the annotation, each individual's subjective annotation in the ambiguous boundaries or distorted regions may cause the network to converge to a sub-optimal solution. One promising approach that addresses the above issues is the utility of synthetically generated data for training without human's subjective decision [7]–[10].

There was an effort to introduce CNNs to handwritten text segmentation in scribbled documents [30]. In order to construct the dataset for training the segmentation network for supervision, they presented an algorithm for the synthesis of scribbled documents that can easily obtain pixel-level annotations of handwritten text. They used separately existing datasets: IAM dataset [11] as handwritten text and PRImA dataset [12] as scanned documents. To make realistic scribbled documents, they paid attention to the preservation of textures of handwritten text and consistent scan noise of documents, removing undesirable block artifacts from IAM dataset. Since pixel-level annotations of handwritten text were easily obtained through Otsu binarization [13], large amounts of training data were successfully established without human intervention. However, the network performs well for synthetic data that is participated in the training, but does not work well for real scribbled documents, showing lack of generalization ability as shown in Fig. 1.

Like the above-mentioned example, training a CNN based on such visually appealing synthetic data, and then applying it to real-world images will give inferior performance due to the large differences in image characteristics which give rise to the *domain shift* problem [14]. From a probabilistic point of view, considering the network that is trained only by samples derived from a source distribution (synthetic data), the network will work well only if the test data is also sampled from the same distribution. In this respect, we can infer that the overfitted performance is derived from the discrepancies between synthetic and real images distribution.

A convincing solution to diminish the *domain shift* problem is the Domain Adaptation (DA) [15]–[28]. In principle, DA is achieved by minimizing some measure of distance between the source and the target distributions [15]. The general approaches of domain adaptation either attempt to learn an additional mapping layer to reduce gap in domain representation [18] or learn domain invariant representations in the same feature space [25].

Inspired by the latter approach, we propose a domain adaptation strategy for handwritten text segmentation. Specifically, adopting Jo *et al.*'s network [30] as a baseline of segmentation network, we apply DA process that is to transfer learned representations from a synthetic to a real dataset by fine-tuning the model on unlabeled target data to address the aforementioned *domain shift* problem. We focus on the practical case of the problem where no labels from the real domain are available, which is commonly referred to as Unsupervised Domain Adaptation (UDA). Also, while aligning the distribution of target data to source ones in feature space, we further incorporate the entropy minimization loss [28] to make the proposed network discriminative for unlabeled target data.



(a) synthetic image (b) (c) real image (d)

Fig. 1: Overfitted performances of Jo *et al.*'s network [30] that is trained only by synthetic training data: (a), (c) input images (synthetic and real), (b), (d) segmentation results (blue: correctly segmented pixels, red: missing or incorrect ones).

To the best of our knowledge, this is the first work to explore UDA for handwritten text segmentation. From our extensive experiments, it achieves plausible results in terms of objective measures.

The rest of this paper is organized as follows. We start by introducing related works in Section II. In Section III and IV, we present details of our proposed network architecture and experiments, respectively. Section V handles results with analysis. Finally, Section VI draws concluding remarks.

II. Related Work

A. Handwritten text segmentation

Document digitization has been an important topic for the decades, and a huge number of methods have been proposed to address many kinds of sub-tasks such as optical character recognition, layout analysis, and so on [31]–[33]. However, the performance of the methods can be severely degraded when there are scribbles on the document. Hence, many researchers addressed separating handwritten texts from the printed document by segmenting them as a unit of connected component (CC) [34]–[40]. In [38], they extracted CCs and assigned feature vectors to them by exploiting hand-crafted features between components. Finally, they classified each component by applying a k-nearest neighbor classifier. Similarly, Kandan *et al.* [36] classified each component by using support vector machines, improving descriptors to be robust to deformations. Li *et al.* [34] use CNNs to classify CCs, incorporating conditional random fields into their framework to consider relations with neighboring CCs. However, since these methods employ binarization and CC extraction as essential preprocessing steps, they have drawbacks that the final performance heavily depends on the performance of each module and lack of generalization ability. To alleviate these problems, Jo *et al.* [30] proposed a pixel-level handwritten text segmentation method based on an end-to-end CNN which does not need any preprocessing steps. They assigned '+1' for pixels of handwritten text and '-1' for others (background, machine-printed text, table boundaries, and so on). Also, to construct a dataset for training the network with supervision, they presented a promising synthesis algorithm that can generate realistic scribbled documents along with the pixel-level annotated labels. However, their network that is trained with synthetic data shows overfitted performances to synthetic data. To address this problem, we adopt Jo *et al.*'s network [30] as the baseline network for handwritten text segmentation and modify it to fit our framework.

B. Domain Adaptation

Domain Adaptation (DA) is a kind of transfer learning that leverages labeled data in one or more related source domains, to perform well for unlabeled data in a target domain [15]. This is generally achieved by minimizing some measure of domain variance, such as the Maximum Mean Discrepancy (MMD) [20], or by matching moments of the two distributions [21]. Recently, adversarial training approaches have shown convincing results, where adversarial generative models [22]–[24] aim to generate source-like data with target data, while adversarial discriminative models [25]–[27] focus on aligning distribution of representations of target domain to source domain in embedding spaces. These impressive strategies of DA have worked as a breakthrough of efficient learning methods using synthetic dataset, such as GTA5 [8], SYNTHIA [9], and so on [7], [10]. Inspired by these approaches [16]–[19], we apply adversarial discriminative models to alleviate the *domain shift* problem in handwritten text segmentation task.

III. Methodology

In this section, we provide details of the DA model for handwritten text segmentation. Let synthetic images $x_s \in X_s \subset \mathbb{R}^{H \times W}$ and the corresponding one-hot encoded binary segmentation map $y_s \in Y_s \subset \mathbb{R}^{H \times W \times 2}$ as samples from

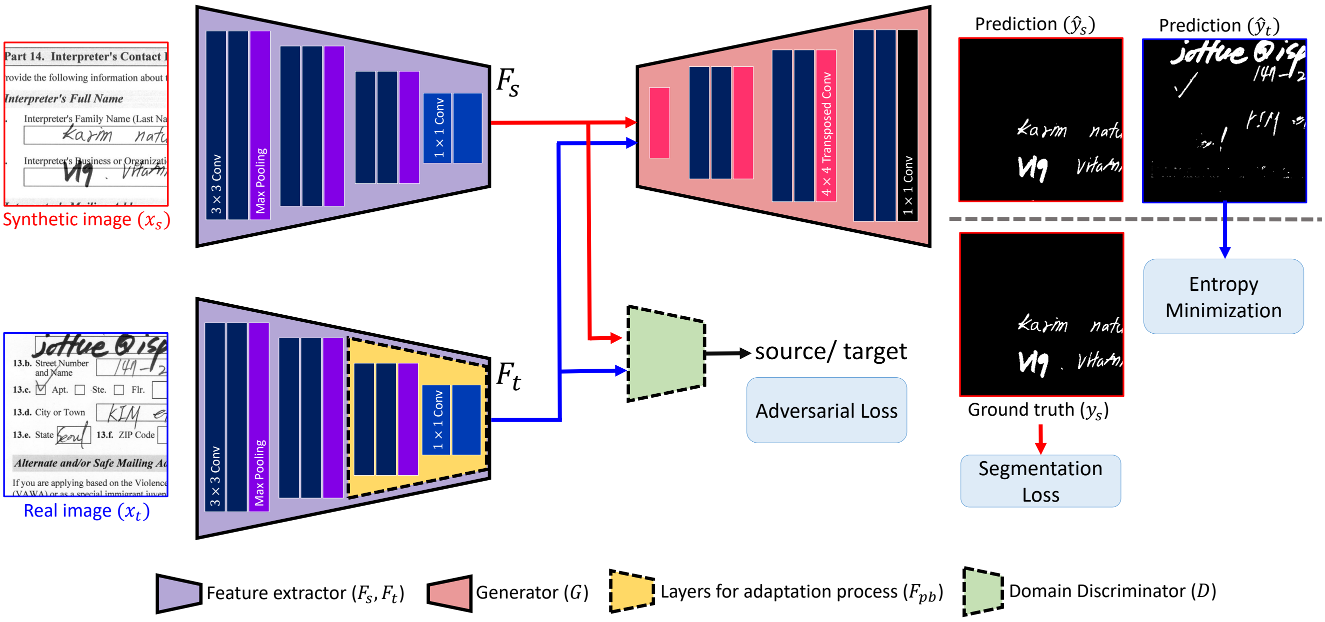


Fig. 2: Overall architectures of proposed network. The networks drawn by the dotted line represent the parts that are updated by participating in the adversarial adaptation process.

a source distribution, $\mathcal{S}(x, y)$. In the case of real images, $x_t \in X_t \subset \mathbb{R}^{H \times W}$ are drawn from a target distribution, $\mathcal{T}(x, y)$. In consideration of the practical issues, labels $y_t \in Y_t \subset \mathbb{R}^{H \times W \times 2}$ cannot be accessed. The baseline network, Jo *et al.*'s [30], can be divided into two components: Feature extractor $F(\cdot)$ that embeds an input image to latent representation spaces, and Generator $G(\cdot)$ that yields the corresponding segmentation map from the latent representations. We aim that learned informative representations from synthetic are well applied to the real data by fine-tuning the model on unlabeled target data through adversarial training in the manner of UDA. Overall architecture of proposed network is illustrated in Fig. 2.

A. *Proposed Approach*

We initially train the baseline $\{F_s(\cdot), G_s(\cdot)\}$ on only synthetic data with full supervision. As a result, the segmentation performance of the network is prone to overfit to source data expressed as

$$\textbf{error}(\mathcal{P}) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[||G_s(F_s(x)) - y||_1], \tag{1}$$

$$\textbf{error}(\mathcal{S}) \ll \textbf{error}(\mathcal{T}), \tag{2}$$

suggesting biased training derived from the differences in image characteristics between source and target. If we utilize the same generator, $G_s(\cdot) = G_t(\cdot) = G(\cdot)$, that learns the posteriors, $\mathcal{P}(Y|Z)$, *i.e.*, the prediction of the pixel-level category of the object (Y), given latent representations (Z), we can infer that the *domain shift* problem is caused by the difference on the marginal distribution $\mathcal{P}(Z)$. In other words, to address the problem, we should enforce the distributions of representation from two domains to be indistinguishable in the feature space, *i.e.*, $\mathcal{P}(F_s(X_s)) = \mathcal{P}(F_t(X_t))$, for the subsequent prediction. For setting the range of adaptation in $F_t(\cdot)$, we refer to [29] regarding the transferability of learned features in hierarchical deep networks. They informed that as we move from lower to higher layers, there are increases in domain discrimination capability. In other words, discrepancies between feature distributions are derived from some of higher layers in $F_t(\cdot)$, and then, we empirically apply adaptation process to only higher 4 convolutional layers of $F_t(\cdot)$, named as $F_{pb}(\cdot)$, and fixed shared parameters of other layers.

We start the adaptation process by initializing $F_t(\cdot)$ with the supervisely trained weights of $F_s(\cdot)$. To measure the discrepancies of latent distributions between source and target, we employ domain discriminator $D(\cdot)$ that discriminates origin of latent representations between the source and target in feature spaces. With each iteration, $D(\cdot)$ and $F_{pb}(\cdot)$ are adversarially trained. $F_{pb}(\cdot)$ is updated to learn which embeddings are more likely to fool $D(\cdot)$, *i.e.*, to embed indistinguishably in feature spaces between source and target. As training progresses, the features become more domain invariant, therefore, the prediction performance gradually improves. We further incorporate the entropy minimization loss [28] to make proposed network discriminative for unlabeled target data during aligning feature distribution of target to source,.

In rest of this section, we explain the objective functions for adversarial adaptation. Details on adversarial adaptation training process is summarized as a pseudocode in Algorithm 1. The optimization steps are implemented using stochastic gradient updates of each minibatch.

B. Objective Functions

1) Segmentation loss: Cross entropy loss is widely used in segmentation task [2], [41]. However, as [30] stated, there are *class imbalance* problems that the number of background pixels is approximately 20 times larger than that of text pixels which make network converge to sub-optimal solution. To alleviate the problems, they proposed dynamically balanced cross entropy loss $\mathcal{L}_{\text{DBCE}}$ incorporating with *focal* loss. In our case, to deviate the instability of $\mathcal{L}_{\text{DBCE}}$ for adversarial training, we only adopt *focal* loss [42] given as,

$$\mathcal{L}_{seg}(F(\cdot), \mathcal{S}) = \mathbb{E}_{(x_s,y_s)\sim\mathcal{S}}[\mathbf{FC}(G(F(x_s)), y_s)], \quad (3)$$

where

$$\mathbf{FC}(p, q) = || - q \odot (1-p)^\gamma \odot \log(p)||_1, \quad (4)$$

where $p, q \in \mathbb{R}^{H\times W\times 2}$ denote prediction probability map through proposed network and one-hot encoded label map, respectively. γ is the hyperparameter that determines the boost degree of the penalty. The scaling factor $(1-p)^\gamma$ automatically lessens the contribution of easy examples and makes the model focus on hard examples, balancing the training.

2) Adversarial loss: To enforce the distributions of representation from two domains to be closer in feature space, discrepancies between them are measured by $D(\cdot)$ that trained to maximize the probability of assigning the correct label to both source examples and target examples as 1 and 0, respectively. For $D(\cdot)$ and $F_{pb}(\cdot)$, the adversarial loss is described as

$$\begin{aligned}\mathcal{L}_{ad:D}(\mathcal{S}, \mathcal{T}) = \mathbb{E}_{x_s\sim\mathcal{S}}[(D(F_s(x_s)) - 1)^2] \\ + \quad \mathbb{E}_{x_t\sim\mathcal{T}}[(D(F_t(x_t)))^2], \quad (5)\end{aligned}$$

$$\mathcal{L}_{ad:F}(\mathcal{T}) = \mathbb{E}_{x_t\sim\mathcal{T}}[(D(F_t(x_t)) - 1)^2]. \quad (6)$$

We employed a least-squares GAN loss [47] to stabilize the training of $D(\cdot)$.

3) Entropy minimization loss: To obtain discriminative features on unlabeled target examples, we need to cluster target features far from the decision boundary of $G(\cdot)$ without supervision due to absences of labels. We adopt the entropy minimization loss $\mathcal{L}_{\text{entropy}}$ on target data given as

$$\mathcal{L}_{\text{entropy}}(\mathcal{T}) = \mathbb{E}_{x_t\sim\mathcal{T}}[\mathbf{H}(G(F_t(x_t)))], \quad (7)$$

where

$$\mathbf{H}(p) = || - p \odot \log(p)||_1, \quad (8)$$

where $p \in \mathbb{R}^{H\times W\times 2}$ denote prediction probability map through proposed network. Derived gradient by this term only flows to F_{pb}, enforcing feature embedding of target data far from the decision boundary of $G(\cdot)$, which is the way that reducing the self-entropy, *i.e.*, decreasing the uncertainty of class probability, resulting in the desired discriminative features.

Algorithm 1 Training Procedure for Adaptation

STEP 1: Training of baseline network
for N steps **do**
Sample k patches from $\mathcal{S}$: $\mathcal{X}_\mathcal{S} := \{x_s^i, y_s^i\}_{i=1}^k$

$$\theta^*_{F_s}, \theta^*_G \leftarrow \arg\min_{\theta_{F_s},\theta_G} \mathcal{L}_{seg}(F_s(\cdot), \mathcal{X}_\mathcal{S})$$

end for

STEP 2: Initialize for adaptation

$$\theta^*_{F_t} \leftarrow \theta^*_{F_s}$$

while $\mathcal{L}_{ad:D} > \epsilon$ **do**
Sample k patches from $\mathcal{S}$: $\mathcal{X}_\mathcal{S} := \{x_s^i, y_s^i\}_{i=1}^k$
Sample k patches from $\mathcal{T}$: $\mathcal{X}_\mathcal{T} := \{x_t^i\}_{i=1}^k$

$$\theta^*_D \leftarrow \arg\min_{\theta_D} \mathcal{L}_{ad:D}(\mathcal{X}_\mathcal{S}, \mathcal{X}_\mathcal{T})$$

end while

STEP 3: Adversarial training between $D(\cdot)$ and $F_{pb}(\cdot)$
while do
Sample k patches from $\mathcal{S}$: $\mathcal{X}_\mathcal{S} := \{x_s^i, y_s^i\}_{i=1}^k$
Sample k patches from $\mathcal{T}$: $\mathcal{X}_\mathcal{T} := \{x_t^i\}_{i=1}^k$

$$\theta^*_D \leftarrow \arg\min_{\theta_D} \mathcal{L}_{ad:D}(\mathcal{X}_\mathcal{S}, \mathcal{X}_\mathcal{T})$$

$$\theta^*_{F_{pb}} \leftarrow \arg\min_{\theta_{F_{pb}}} \{\lambda_1 \mathcal{L}_{seg}(F_t(\cdot), \mathcal{X}_\mathcal{T}) + \mathcal{L}_{ad:F}(\mathcal{X}_\mathcal{T}) + \lambda_2 \mathcal{L}_{\text{entropy}}(\mathcal{X}_\mathcal{T})\}$$

end while

IV. Experiments

A. Dataset

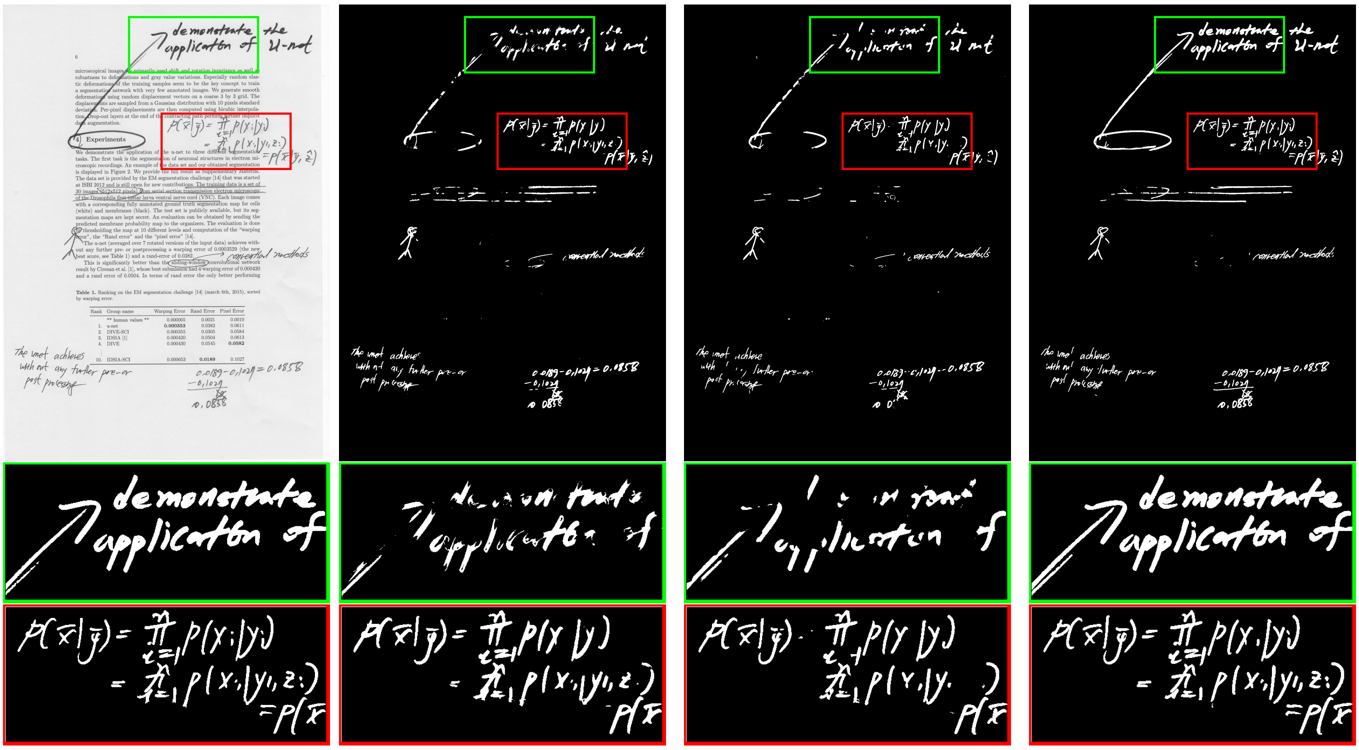
As source dataset, we used synthetically generated scribbled documents that Jo *et al.* [30] released. This dataset is composed of $146,391$ patches (128×128) of synthetic scribbled documents with perfectly annotated pixel-level labels. In the case of target dataset, we manually assembled a wide range of the scribbled documents without any annotations. For utilizing in adaptation training procedure, we cropped and augmented, and then finally, made $23,596$ patches with the same size of source ones. So, unlabeled training patches are used in unsupervised manners, due to absence of annotations. Owing to absence of annotations, real images are participated in training only with unsupervised manner.

B. Specification of training

We have trained the network using RMSProp [48] optimizer with a mini-batch size of 32. To stabilize the adversarial training, we set the initial learning rate as a small value (0.00005) with 0.96 decay rate in every 30 epochs. In case of others hyper-parameters, we empirically set $\gamma = 1$, $\lambda_1 = 0.01$, and $\lambda_2 = 0.1$, respectively.

C. Convergence issues

There are two critical convergence issues in the early stages of adversarial training. First, when setting the time



(a) original image (b) Jo *et al.* [30] (c) Baseline (d) ours

Fig. 3: Segmentation results on a real scribbled document through each captioned model.

of intervention for adaptation training, there are some trade-offs: The faster the intervention time, the less time was spent learning the informative representations solely on the source data, and then performance was poor. Conversely, the slower the intervention time, the more severe the discrepancies and the alignment becomes impossible. Considering these facts, we empirically intervene after $20,000$ iteration of training for baseline.

The second problem occurs in the early stages of adversarial learning when $D(\cdot)$ does not work well, *i.e.*, provide bad information as the value of the gradient used to train $F_{pb}(\cdot)$. As a result, we could see that weights that trained with the synthetic dataset were meaningless. To prevent the significantly lower performance of $D(\cdot)$ from contaminating the informative representations from $F_s(\cdot)$, We started adversairal training process after learning $D(\cdot)$ to give some performance.

V. Results

In this section, we present a thorough ablation study to see whether the objective functions contribute to the overall performance. We did not perform the comparisons with other existing works [34]–[40] except Jo *et al.* [30]. Since there is none that publicly provides the code and data to compare the performances. Also critically, they dealt with CCs or region-level results that can not be directly compared to ours, *i.e.*, pixel-level results.

For the quantitative comparisons, mean of pixel-level intersection-over-union (mIoU) among the classes has been

TABLE I: IoU results on real scribbled documents. The best results are highlighted in **bold face** and the second best results are <u>underlined</u>. H: handwritten text

Method			non-H (%)	H (%)
Jo *et al.* [30]			98.818	64.617
baseline			98.698	58.933
$\mathcal{L}_{\text{seg}}$	$\mathcal{L}_{\text{ad}}$	$\mathcal{L}_{\text{entropy}}$		
	✓		99.394	81.279
✓	✓		<u>99.470</u>	<u>83.660</u>
✓	✓	✓	**99.533**	**85.642**

widely evaluated in semantic segmentation task. In our case, due to severe imbalance between non-handwritten text pixels and handwritten text ones, mean value of IoU has numerically meaningless, instead, evaluate IoU of each class.

A. Ablation Study

TABLE I demonstrates the effect of each individual objective function for the overall performance. Alone with the adversarial loss $\mathcal{L}_{ad}$ for adaptation process, overfitted performance to synthetic data is quite mitigated, showing significant gain from 58.932% to 81.279%. In this respect, we can conclude that applying adaptation technique to baseline is meaningfully functioning to address the *domain shift* problem.

$G(\cdot)$ was trained to predict segmentation map exactly fitted to the features of synthetic data. In this respect, if $F_t(\cdot)$ has no constraints to maintain informative representations about source, $G(\cdot)$ could not work its role for performance. By imposing the constraints about maintenance of information through $\mathcal{L}_{seg}$, we achieve the advantages of 2.381%p as shown in TABLE I.

While we observe that using only the adversarial loss and baseline loss terms does improve performance, the entropy minimization loss is needed to get the full performance benefit. As shown in TABLE I, we can conclude that usage of entropy minimization objective on unlabeled data makes the network discriminative for unlabeled data.

B. Comparisons with other approach

Fig. 3 demonstrates the comparisons of segmentation performance on each methods, where the first column shows the input and the label of segmentation results on green and red-box region. We compare our method to Jo *et al.*'s network [30] that we used it as our baseline in modified form. They tried efforts to diminish the overfitting problems, such as applying l_2-weight decay as regularization and removal of undesirable block artifacts in synthetic data for realistic. As stated in Section III-B1, we could not use $\mathcal{L}_{\text{DBCE}}$ to address task-specific imbalance problems due to the convergence issues of adversarial training. Although our baseline network performance is degraded from 64.617% to 58.933% against Jo *et al.*'s network [30], applying DA technique, proposed network outperforms with the significant margin as shown in TABLE I. Additionally, as shown in Fig. 3, proposed networks have a great enhancement of segmentation performance, addressing the *domain shift* problem.

VI. CONCLUSION

In this paper, we have addressed overfitted performance of previous handwritten segmentation network to synthetic training dataset, which is mainly due to the different characteristics of synthetic and real training images. We have proposed domain adaptation strategy to alleviate this domain shift problem, and also demonstrated the proposed method's effectiveness and superiority in segmentation performance through extensive experiments on real scribbled documents dataset. These are achieved by applying adversarial discriminative models to align feature distribution and entropy minimization to make network discriminative to real data. Note that there are no regularization to prevent overfitting and no supervision for real image, we can conclude that the proposed domain adaptation strategy alleviates overfitting problem very well. As a future work, we would like to extend this unsupervised approach to a semi-supervised one for better performance, explicitly providing the labels of real scribbled documents.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.NI190004, Development of AI based Robot Technologies for Understanding Assembly Instruction and Automatic Assembly Task Planning).

REFERENCES

[1] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[2] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[3] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.
[4] Cordts, Marius, et al. "The cityscapes dataset for semantic urban scene understanding." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[5] Everingham, Mark, et al. "The pascal visual object classes (voc) challenge." International journal of computer vision 88.2 (2010): 303-338.
[6] Silberman, Nathan, et al. "Indoor segmentation and support inference from rgbd images." European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012.
[7] Dosovitskiy, Alexey, et al. "CARLA: An Open Urban Driving Simulator." Conference on Robot Learning. 2017.
[8] Richter, Stephan R., et al. "Playing for data: Ground truth from computer games." European Conference on Computer Vision. Springer, Cham, 2016.
[9] Ros, German, et al. "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[10] Shafaei, Alireza, James J. Little, and Mark Schmidt. "Play and Learn: Using Video Games to Train Computer Vision Models."
[11] U-V Marti and Horst Bunke, "The iam-database: an english sentence database for offline handwriting recognition," International Journal on Document Analysis and Recognition, vol. 5, no. 1, pp. 39–46, 2002.
[12] Apostolos Antonacopoulos, David Bridson, Christos Papadopoulos, and Stefan Pletschacher, "A realistic dataset for performance evaluation of document layout analysis," in Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on. IEEE, 2009, pp. 296–300.
[13] Otsu, Nobuyuki. "A threshold selection method from gray-level histograms." IEEE transactions on systems, man, and cybernetics 9.1 (1979): 62-66.
[14] Gretton, Arthur, et al. "Covariate shift by kernel mean matching." Dataset shift in machine learning 3.4 (2009): 5.
[15] Ben-David, Shai, et al. "A theory of learning from different domains." Machine learning 79.1-2 (2010): 151-175.
[16] Sankaranarayanan, Swami, et al. "Generate to adapt: Aligning domains using generative adversarial networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[17] Nath Kundu, Jogendra, et al. "Adadepth: Unsupervised content congruent adaptation for depth estimation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[18] Murez, Zak, et al. "Image to image translation for domain adaptation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[19] Su, Hao, et al. "Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views." Proceedings of the IEEE International Conference on Computer Vision. 2015.
[20] Long, Mingsheng, et al. "Learning transferable features with deep adaptation networks." arXiv preprint arXiv:1502.02791 (2015).
[21] Peng, Xingchao et al. "Moment Matching for Multi-Source Domain Adaptation." CoRR abs/1812.01754 (2018): n. pag.
[22] Liu, Ming-Yu, and Oncel Tuzel. "Coupled generative adversarial networks." Advances in neural information processing systems. 2016.
[23] Bousmalis, Konstantinos, et al. "Unsupervised pixel-level domain adaptation with generative adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[24] Taigman, Yaniv, Adam Polyak, and Lior Wolf. "Unsupervised cross-domain image generation." arXiv preprint arXiv:1611.02200 (2016).
[25] Tzeng, Eric, et al. "Simultaneous deep transfer across domains and tasks." Proceedings of the IEEE International Conference on Computer Vision. 2015.
[26] Hoffman, Judy, et al. "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation." arXiv preprint arXiv:1612.02649 (2016).
[27] Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." The Journal of Machine Learning Research 17.1 (2016): 2096-2030.

[28] Grandvalet, Yves, and Yoshua Bengio. "Semi-supervised learning by entropy minimization." Advances in neural information processing systems. 2005.

[29] Yosinski, Jason, et al. "How transferable are features in deep neural networks?." Advances in neural information processing systems. 2014.

[30] Jo, Junho et al. "Handwritten Text Segmentation via End-to-End Learning of Convolutional Neural Network." (2019).

[31] Ray Smith, "An overview of the tesseract ocr engine," in Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on. IEEE, 2007, vol. 2, pp. 629–633.

[32] Jewoong Ryu, Hyung Il Koo, and Nam Ik Cho, "Language-independent text-line extraction algorithm for handwritten documents," IEEE Signal processing letters, vol. 21, no. 9, pp. 1115–1119, 2014.

[33] Hyung Il Koo and Nam Ik Cho, "Text-line extraction in handwritten Chinese documents based on an energy minimization framework," IEEE Transactions on Image Processing, vol. 21, no. 3, pp. 1169–1175, 2012.

[34] Xiao-Hui Li, Fei Yin, and Cheng-Lin Liu, "Printed/handwritten texts and graphics separation in complex documents using conditional random fields," in 2018 13th IAPR International Workshop on Document Analysis Systems (DAS). IEEE, 2018, pp. 145–150.

[35] Mathias Seuret, Marcus Liwicki, and Rolf Ingold, "Pixel level handwritten and printed content discrimination in scanned documents," in Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on. IEEE, 2014, pp. 423–428.

[36] R Kandan, Nirup Kumar Reddy, KR Arvind, and AG Ramakrishnan, "A robust two level classification algorithm for text localization in documents," in International Symposium on Visual Computing. Springer, 2007, pp. 96–105.

[37] Xujun Peng, Srirangaraj Setlur, Venu Govindaraju, and Ramachandrula Sitaram, "Handwritten text separation from annotated machine printed documents using markov random fields," International Journal on Document Analysis and Recognition (IJDAR), vol. 16, no. 1, pp. 1–16, 2013.

[38] J¨urgen Franke and Matthias Oberlander, "Writing style detection by statistical combination of classifiers in form reader applications," in Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on. IEEE, 1993, pp. 581–584.

[39] Konstantinos Zagoris, Ioannis Pratikakis, Apostolos Antonacopoulos, Basilis Gatos, and Nikos Papamarkos, "Handwritten and machine printed text separation in document images using the bag of visual words paradigm," in Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on. IEEE, 2012, pp. 103–108.

[40] Abdel Bela¨ıd, KC Santosh, and Vincent Poulain d'Andecy, "Handwritten and printed text separation in real document," arXiv preprint arXiv:1303.4614, 2013.

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[42] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Doll´ar, "Focal loss for dense object detection," IEEE transactions on pattern analysis and machine intelligence, 2018.

[43] Zhi-Hua Zhou and Xu-Ying Liu, "Training costsensitive neural networks with methods addressing the class imbalance problem," IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 1, pp. 63–77, 2006.

[44] U-V Marti and Horst Bunke, "The iam-database: an english sentence database for offline handwriting recognition," International Journal on Document Analysis and Recognition, vol. 5, no. 1, pp. 39–46, 2002.

[45] Apostolos Antonacopoulos, David Bridson, Christos Papadopoulos, and Stefan Pletschacher, "A realistic dataset for performance evaluation of document layout analysis," in Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on. IEEE, 2009, pp. 296–300.

[46] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[47] Mao, Xudong, et al. "Least squares generative adversarial networks." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[48] Hinton, Geoffrey, Nitish Srivastava, and Kevin Swersky. "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent." Cited on 14 (2012).