2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)

Historical Document Synthesis With Generative Adversarial Networks

Vinaychandran Pondenkandath[*†], **Michele Alberti**[*†], **Michaël Diatta**[*†], **Rolf Ingold**[†], **Marcus Liwicki**[†§]

[†]*Document Image and Voice Analysis Group (DIVA)*
University of Fribourg, Switzerland
{firstname}.{lastname}@unifr.ch

[§]*Machine Learning Group*
Luleå University of Technology, Sweden
marcus.liwicki@ltu.se

***Abstract*—This work tackles a particular image-to-image translation problem, where the goal is to transform an image from a source domain (modern printed electronic document) to a target domain (historical handwritten document). The main motivation of this task is to generate massive synthetic datasets of "historic" documents which can be used for the training of document analysis systems. By completing this task, it becomes possible to consider the generation of a tremendous amount of synthetic training data using only one single deep learning algorithm. Existing approaches for synthetic document generation rely on heuristics, or 2D and 3D geometric transformation-functions and are typically targeted at degrading the document. We tackle the problem of document synthesis and propose to train a particular form of Generative Adversarial Neural Networks, to learn a mapping function from an input image to an output image. With several experiments, we show that our algorithm generates an artificial historical document image that looks like a real historical document – for expert and non-expert eyes – by transferring the "historical style" to the classical electronic document.**

I. Introduction

In the past decade, Deep Learning approaches have shown impressive results for computer vision tasks such as image segmentation [1], object recognition [2] and image synthesis [3]. This state of affairs is mainly due to the advent of powerful computers, to the accessibility of larger datasets, but also to the improvement of algorithmic techniques that allow training deeper networks [4], [5].

Unfortunately, in the field of Document Image Analysis (DIA) for historical documents, labelled image datasets are a scarce resource. This lack of labelled training data makes it challenging to take advantage of several deep learning breakthroughs [6], [7].

The shortage of labelled data can be dealt with in two ways; acquiring larger labelled datasets or generating synthetic datasets. Acquiring large labelled task-specific data is often a very effective way to increase performance on a task; however, this is often very expensive both in terms of time and resources. Recent approaches based on crowd sourcing [8] show some possibilities to tackle this problem, but they require similarities in handwriting [9] or the domain (structured data records).

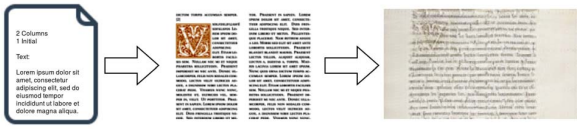


Fig. 1. The first step is to create modern electronic printed documents from a Latex specification document. The second step involves using a deep neural network to learn a mapping function to transform the modern printed document to a historical handwritten document.

An effective system for generating synthetic data would allow for generating large labelled datasets in a relatively cheap and efficient manner [10], [7]. This can be achieved by creating a system that takes the specifications of the pages to be generated in an electronic format (such as XML or Latex) and then produces the document image such that the ground truth matches the documents. Existing tools, however, are limited to modern documents or prints.

Contribution

In this paper, we want to show how it is possible to go one step further in the history of Image Synthesis Software for Historical Handwritten Documents. Contrary to several traditional methodologies [11], [12], we produce a new general-purpose generating framework that takes advantage of recent advancements in the design of Generative Adversarial Networks (GANs) and Neural Style Transfer Algorithms (NST). We tackle an Image-to-Image translation task by learning a mapping function that goes from the Source Domain S (modern printed electronic document) to the Target Domain T (historical handwritten document). Our primary goal is to generate synthetic historical handwritten document images, which looks like other historical handwritten documents from T. Our secondary goal is to demonstrate a new promising and more straightforward approach to create a large amount of complex synthetic handwritten historical documents based on different ground-truthed electronic documents. The proposed

[*] These authors contributed equally to this work.

978-1-7281-5054-3/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDARW.2019.40096

(a) Book 1 (b) Book 2

(c) Book 3 (d) Book 4

(e) Book 5 (f) Book 6

Fig. 2. Sample pages of each historical book used from the HBA 1.0 dataset. (a) Book 1: "Plutarchus, Vitae illustrium virorum". (b) Book 2: "Justinien, Institutes". (c) Book 3: "Girart d'Amiens, Meliacin ou le Cheval de fust". (d) Book 4: "Cy, commencent le Procès de Belial à l'encontre de Jhésus". (e) Book 5: "Voyage pittoresque de la Grèce". (f) Book 6: "La Chartreuse de Parme"

cycleGAN framework reaches the goals and performs the generation task in an integrated manner. We believe that our promising results show that this area of research requires further investigation.

II. Related Work

In this section, we briefly summarize the relevant related work to our scenario.

A. Annotation and Image Synthesis Softwares

Document Image Annotation software is designed to assist researchers in the process of annotating documents. The goal of such software is to assist the user in creating ground truth for documents. Pink Panther [13] and trueViz [14] require the user to manually annotate the entire document, but other software such as Aletheia [6], DIVAnnotation [15] and Graph-ManuScribble [16] assist the user with semi-automated annotations. Additionally, DIVADIAWI [17], DIVAServices [18] and Transcriptorium [19] provide an online platform to assist users with their annotation efforts. Despite the various attempts to produce more accessible, collaborative, web-based and automated platforms, the process of annotation using such software remains tedious and requires some expertise.

In parallel, the DIA community has expended significant efforts to develop software that allows for the generation of synthetic document datasets. Several approaches focus on creating defect models and associated generators [20], [21]. Other frameworks [10], [7] focus on creating synthetic documents by document structures defined with XML or Latex combined with background extraction and other data augmentation methods. These approaches lead to complex pipelines that involve human expertise as well as human biases over the structure and the definition of document pages. While there has been work that uses deep learning to generate synthetic document images [22], the approach has been limited in scope to generating only individual characters by font transfer using Neural Style Transfer (see Section II-C).

B. Image-to-Image Translation with GANs

An Image-to-Image translation problem can be considered as an image constrained synthesis process [23], [24], where the goal is to transform a particular image representation into a different image representation. Several computer vision problems like colourization, super-resolution, segmentation or depth estimation can be grouped into the same category of problems [25], [23], [26], [27].

Pix2Pix [24] is the first GAN framework that proposed a good solution for supervised Image-to-Image translation tasks, using paired images. This scenario requires that the images in the source domain have a corresponding paired image in the target domain. Algorithms that are trained with such paired image datasets aim to identify the mapping function between the source and target domains.

Since 2016, several unsupervised Image-to-Image translation frameworks have tried to achieve the same task for unpaired image datasets. Following the GAN classification description presented in [28], we can divide them into four categories. Approaches that use a cycle loss with at least a bi-directional reconstruction implementation [29], [30], [31]. Approaches that use a pair-wise distance constraint loss [32]. Approaches that use a task-specific domain adaptation auxiliary classifier like in [33], and approaches that use Variational Auto-Encoder (VAEs) with weight sharing as generator [34].

C. Image-to-Image Translation with Neural Style Transfer Algorithm

Another approach towards an Image-to-Image translation task, is to restrict the heterogeneous domain conversion problem [35] to a style transfer problem [23]. The goal is to learn

to separate the content from the style for two given domains to produce a synthetic output that mixes the two characteristics into one single image representation, which is called Neural Style Transfer Algorithm (NST) [23].

III. Datasets

In this section we describe the dataset we use in this work.

A. Historical Handwritten Document Dataset

The historical documents we use in this work come from the HBA 1.0 dataset[1]. The dataset is composed of eleven books, five manuscripts and six printed books written in different languages and typographies, and published between the twelfth and nineteenth centuries. It contains 4436 pages: 2,435 handwritten pages and 2,001 printed pages. For our purposes, we select 2553 images from six books (see Figure 2), of which four books have handwritten text and two books have printed text. The approximate average resolution of the documents is 2600×4000 pixels. Additionally, we remove all blank and binding pages to have only pages that contain text.

B. Electronic Modern Documents

We develop a framework that allows users to create document images of a modern style using Latex. The framework allows users to specify parameters such as the number of columns, the presence of decorative initials, font and size. As the images are generated programmatically, we have information about the coordinates of all of the various components of the images (such as decorations, layout and textual information). Once the required parameters are specified, the document images can be generated using Latex. For our purposes, we generate 1095 document images of approximate average resolution 1200×1600 pixels. The dataset consists of images in a single and double column, with three different types of decorations set to various sizes and styles. Some example documents generated using the framework can be seen in Figure 3.

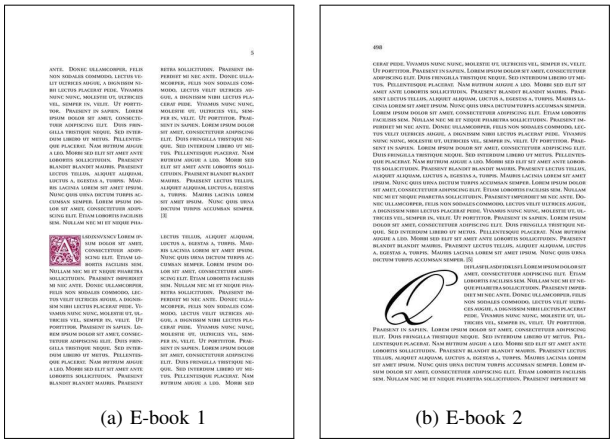
C. Domain Datasets Taxonomy

We partition the aforementioned datasets into different subsets that are used for training or pre-training the different components of the models.

- The Handwritten Document Dataset (HDD) is composed of Books 2, 3 and 4 from the modified HBA Dataset.
- The Printed Document Dataset (PDD) is composed of Books 5 and 6 from the modified HBA Dataset and the E-book 2 from our Latex -framework.
- The GAN Training Dataset (GTD) is composed of Book 1 from the modified HBA Dataset and by E-book 1 from our Latex -framework.

The HDD and PDD datasets are used only for pre-training the auto-encoder parts of our cycleGAN mode as defined in Section IV. The GTD is used for training the final cycleGAN model as a whole.

[1] http://hba.litislab.eu/index.php/dataset/



(a) E-book 1 (b) E-book 2

Fig. 3. The figure shows two-page samples of the generated Latex modern electronic documents. These samples depict two different column modes, initial style and size. E-book 1 is part of the GAN Training Dataset. E-book 2 is used for the pre-training phase and is a member of the Handwritten Document Dataset.

IV. Experimental Setting

In this section we explain the task, model architecture and the experimental set-up for the different approaches we use for historical document image synthesis.

A. Model Architecture

As the dataset that we use does not contain paired images between the source and target domains, we use a variant of GANs, called the cycleGAN [29], that uses the cycle consistency loss. The cycleGAN architecture performs a transformation of the images from the source to the target domains and vice-versa. The cycle consistency loss with the bi-directional mapping function coupled with the L1 distance loss increases the learning stability of the adversarial framework in an unpaired image setting [28].

For the Neural Style Transfer, we choose the VGG-19 Convolutional Neural Network implementation [23] where the goal is to minimize the content loss and the style functions conjointly.

B. Task

As shown in Figure 1, we tackle the problem of historical document image synthesis in two steps. The first step of generating source domain documents is achieved with a Latex framework as described in Section III-B. In the second step, we train a neural network to learn a mapping function between the source domain (modern document) and the target domain (historical handwritten document). The second step can be further divided into three tasks, a reconstruction task and classification task which are used to pre-train the networks and the final generation task.

- The reconstruction task is used to pre-train the two encoder components of the generators of the cycleGAN model.

(a) Target Domain Samples (b) Source Domain Samples (c) cycleGAN Synthetic Samples (d) NST Synthetic Samples

Fig. 4. Examples of images generated by the cycleGAN and the NST after training on the Complete Document images. The first and second columns contains samples from the Target Domain and Source Domain respectively. The third column contains samples generated by the cycleGAN trained from scratch. The samples generated by the NST model (pre-trained on the PDD dataset) can be seen in the fourth column. Every sample contains a zoomed-in view to see the quality of the generated pages.

- The classification task is used to pre-train the Neural Style Transfer model and the discriminator component of the cycleGAN model.
- The generation task is the main Image-to-Image translation task that uses unpaired data. This task is performed with two different approaches, the cycleGAN model and the Neural Style Transfer model.

C. Data Pre-processing

As seen in Section III, the dataset we use contain high resolution images. The high resolution of these images makes it difficult to train the cycleGAN and the VGG-19 models on the native resolution image due to computational memory constraints. As a workaround, we consider two different data pre-processing methods to prepare the data before training:

- Complete Document: We use the entire document image as input to the network after resizing it to 256×256 pixels. This reduces the fine detail, but preserves the global structure and look of the document.
- Random Crop: We select a random crop of size 256×256 pixels from the central portion (to avoid blank border regions) of the document. In this scenario, the fine detail of the pages are preserved, and crops have almost the same number of lines and words in both the historical and modern domains.

D. cycleGAN Training Procedure

We train the cycleGAN model in two different settings – from scratch and from a pre-trained model.

When training the cycleGAN from scratch, we train the model for 50 epochs with a batch size of 1. The learning rate is 0.0002 with a linear decay starting from epoch 25. We also use a history buffer that stores the 50 most recently generated images. This history buffer is used to update the discriminator and reduce model oscillations during training.

When training the cycleGAN in the pre-trained scenario, we initialize the generator and discriminator with weights obtained from the reconstruction and classification tasks respectively. The weights of the encoder part of the generators are initialized with the weights of the encoder component of an auto-encoder that is trained for reconstruction on the HDD and PDD datasets (see Section III-C).

E. NST Training Procedure

We use the VGG-19 based NST model in two different settings – using ImageNet weights and using weights from a pre-trained model. When using the model with the ImageNet weights, only the last layers of the network are reinitialized, and then the NST procedure is applied to the images.

In the case of the pre-training scenario, we first train the VGG-19 on our PDD dataset (see Section III-C) for a

classification task. The network is trained for 25 epochs with a batch size of 4, learning rate of 0.001 and momentum of 0.9. The weights of this model are then used for the NST procedure.

V. Results

The quality of the results produced by generative models is typically evaluated quantitatively or qualitatively. We can divide quantitative methods into two subcategories, the quantitative perceptual studies that are human-based and the quantitative metrics that are machine-based and task-dependent. To grasp the cognitive realism of the generated synthetic images, we use a qualitative approach, based on the subjective perceptual appreciation of the results.

A. Complete Document

Figure 4 shows sample results generated by the cycleGAN and the NST on the complete document images.

The synthetic images generated by the cycleGAN appear significantly better than those generated with NST. Regarding the semantic content (font shape, readability of words and letters, marginal annotations), we can notice many similarities between the target domain samples and the synthetic samples. The overall style content of the target domain (background colour, texture, paper degradation, initials style) is well expressed. However, in a structural content point of view (column-mode, number and presence of initials, textual artifacts), the initials are not well detected and expressed. The two column-mode is not at all expressed. This might be because the that the "Plutarchus" historical book from the GTD dataset (see Section III-C) contains only one column pages. When considering the synthetic documents produced with the NST, the structural content is better preserved. However, the style is mixed and standardized over the entire synthetic document, leading to the presence of a lot of coloured artifacts.

B. Random Crop

Figure 5 shows sample results generated by the cycleGAN and the NST on the random crop images.

In the synthetic images generated by the cycleGAN, we can see that the structural content shown in the samples matches the source domain input in terms of the number of words and lines. The shape of the synthetic words is consistent with that of the source domain input. The structure of the document is preserved even when the documents are a double column. However, there is still some stylistic mismatch when considering the initials and colour. From a semantic point of view, the fonts in the synthetic images look similar to the target domain. However, the letters do not correspond to the Latin letters from the input image.

When considering the NST synthetic images, the structural and semantic content are the same as the input images. However, the font does not change as compared to the synthetic images generated by the GAN.

(a) TD (b) SD (c) cycleGAN (d) NST

Fig. 5. Examples of images generated by the cycleGAN and NST after training on Random Crop images. The first and second columns contains samples from the Target Domain (TD) and Source Domain (SD) respectively. The third and fourth columns contain samples generated by the cycleGAN and the NST models. The synthetic samples generated by the cycleGAN look similar to the target domain, however they do not correspond perfectly to the text or layout seen in the corresponding source domain images.

VI. Conclusion

We have presented a framework for generating synthetic historical handwritten documents using two stages. The first stage produces a modern style document that can be customised with several parameters. The second stage uses deep neural networks to transform the modern style printed document to a historical handwritten document. With two different approaches (GAN and NST) for the second stage, we demonstrate that it is feasible to transform the domain of an input document image from modern printed to historical handwritten. Our research demonstrates the ability of GANs to generate synthetic documents that preserve the global characteristics of medieval manuscripts, however there remains scope for significant generative improvements at lower levels for characters and words.

A. Future Work

We plan to further investigate the effectiveness of the generated synthetic images, by pre-training neural networks on the synthetic data and comparing the performance of these pre-trained networks against randomly initialised networks for other DIA tasks. Additionally, we aim at improving the quality of the generated synthetic documents by fusing the global and local generation procedures to produce documents that have the correct global structure as well as fine-grained details. Finally, we plan on using accelerators with increased memory to train neural networks on entire pages from large document datasets to study the effects of working with larger input footprints.

ACKNOWLEDGMENT

The work presented in this paper has been partially supported by the HisDoc III project funded by the Swiss National Science Foundation with the grant number 205120_169618.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99.

[3] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images." in *ICML*, 2016, pp. 1349–1357.

[4] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[6] C. Clausner, S. Pletschacher, and A. Antonacopoulos, "Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments," in *2011 International Conference on Document Analysis and Recognition*, 2011, pp. 48–52.

[7] N. Journet, M. Visani, B. Mansencal, K. Van-Cuong, and A. Billy, "DocCreator: A New Software for Creating Synthetic Ground-Truthed Document Images," *Journal of imaging*, vol. 3, no. 4, p. 62, 2017.

[8] A. Fornés, J. Lladós, J. Mas, J. M. Pujades, and A. Cabré, "A bimodal crowdsourcing platform for demographic historical manuscripts," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*. ACM, 2014, pp. 103–108.

[9] M. Moyle, J. Tonra, and V. Wallace, "Manuscript transcription by crowdsourcing: Transcribe bentham," *Liber Quarterly*, vol. 20, no. 3/4, pp. 347–356, 2011.

[10] S. Capobianco and S. Marinai, "DocEmul: A Toolkit to Generate Structured Historical Documents," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1. IEEE, 2017, pp. 1186–1191.

[11] C. A. Mello and R. D. Lins, "Generation of images of historical documents by composition," in *Proceedings of the 2002 ACM symposium on Document engineering*. ACM, 2002, pp. 127–133.

[12] C. A. Mello, "Synthesis of images of historical documents for web visualization," in *Multimedia Modelling Conference, 2004. Proceedings. 10th International*. IEEE, 2004, pp. 220–226.

[13] B. A. Yanikoglu and L. Vincent, "Pink panther: a complete environment for ground-truthing and benchmarking document page segmentation," *Pattern Recognition*, vol. 31, no. 9, pp. 1191–1204, 1998.

[14] C. H. Lee and T. Kanungo, "The architecture of trueviz: A groundtruth/metadata editing and visualizing toolkit," *Pattern recognition*, vol. 36, no. 3, pp. 811–825, 2003.

[15] M. Seuret, M. Bouillon, F. Simistira, M. Würsch, M. Liwicki, and R. Ingold, "A semi-automatized modular annotation tool for ancient manuscript annotation," in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 2018, pp. 340–344.

[16] A. Garz, M. Seuret, F. Simistira, A. Fischer, and R. Ingold, "Creating ground truth for historical manuscripts with document graphs and scribbling interaction," in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*. IEEE, 2016, pp. 126–131.

[17] H. Wei, K. Chen, M. Seuret, M. Würsch, M. Liwicki, and R. Ingold, "DIVADIAWI – a Web-based Interface for Semi-automatic Labeling of Historical Document Images," *Digital Humanities 2015*, 2015.

[18] M. Würsch, R. Ingold, and M. Liwicki, "DivaServicesA RESTful web service for Document Image Analysis methods," *Digital Scholarship in the Humanities*, vol. 32, no. 1, pp. 150–156, 2017. [Online]. Available: http://dx.doi.org/10.1093/llc/fqw051

[19] B. Gatos, G. Louloudis, T. Causer, K. Grint, V. Romero, J. A. Sánchez, A. H. Toselli, and E. Vidal, "Ground-truth production in the transcriptorium project," in *2014 11th IAPR International Workshop on Document Analysis Systems*. IEEE, 2014, pp. 237–241.

[20] H. S. Baird, "Document image defect models," in *Structured Document Image Analysis*. Springer, 1992, pp. 546–556.

[21] M. Seuret, K. Chen, N. Eichenbergery, M. Liwicki, and R. Ingold, "Gradient-domain degradations for improving historical documents images layout analysis," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 1006–1010.

[22] G. Atarsaikhan, B. K. Iwana, A. Narusawa, K. Yanai, and S. Uchida, "Neural font style transfer," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 5. IEEE, 2017, pp. 51–56.

[23] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[24] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," *arXiv preprint arXiv:1611.07004*, 2016.

[25] S. J. Pan, Q. Yang, W. Fan, and S. J. P. (ph. D, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[26] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[27] X. Wu, K. Xu, and P. Hall, "A survey of image synthesis and editing with generative adversarial networks," *Tsinghua Science and Technology*, vol. 22, no. 6, pp. 660–674, 2017.

[28] H. Huang, P. S. Yu, and C. Wang, "An Introduction to Image Synthesis with Generative Adversarial Nets," *arXiv preprint arXiv:1803.04469*, 2018.

[29] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[30] Z. Yi, H. R. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised Dual Learning for Image-to-Image Translation." in *ICCV*, 2017, pp. 2868–2876.

[31] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 1857–1865.

[32] S. Benaim and L. Wolf, "One-Sided Unsupervised Domain Mapping," in *Advances in neural information processing systems (NIPS)*, 2017, pp. 752–762.

[33] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, no. 2, 2017, p. 7.

[34] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised Image-to-Image Translation Networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 700–708.

[35] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," *arXiv preprint arXiv:1702.05374*, 2017.