Geometry-Aware Domain Adaptation Network for Scene Text
Supplementary Material

Fangneng Zhan

1. Network Configuration

Different scene text recognition and detetion techniques have been developed from the earlier direct methods [9, 24, 16, 1, 6, 10] to the recent learning-based methods [17, 20, 21, 18, 23] and attention models [12, 3, 26]. This work [28] adopts Generative Adversarial Nets (GANs) to achieve the domain adaptation of scene texts, which performs pixel-level adaptation via continuous adversarial learning between generators and discriminators which has achieved great success in image generation [4, 15, 33], image composition [13, 32, 27, 30, 31] and image-to-image translation [34, 8, 19, 29]. Different approaches have been investigated to address pixel-level image transfer by enforcing consistency in the embedding space. [22] translates a rendering image to a real image by using conditional GANs. [2] studies an unsupervised approach to learn pixel-level transfer across domains. [14] proposes an unsupervised image-to-image translation framework using a shared-latent space. [5] introduces an inference model that jointly learns a generation network and an inference network. More recently, CycleGAN [34] and its variants [25, 11] achieve very impressive image translation by using cycle-consistency loss. [7] proposes a cycle-consistent adversarial model that adapts at both pixel and feature levels.

Generators. The generator G_X (or G_Y) consists of G_{X_A} (or G_{Y_A}) and G_{X_B} (or G_{Y_B}) whose structures are shown in Table 1 and Table 2, respectively.

Discriminators. There are three discriminators including D_X, D_Y and D_T. D_X and D_Y adopt the discriminator of PatchGAN [8] whose structure is shown in Table 3. D_T is the spatial transformation discriminator which will distinguish the transformation matrix from $X \rightarrow Y$ and the inverse transformation matrix from $Y \rightarrow X$. Table 4 gives detailed structures of D_T.

Table 1. The structure of G_{X_A} (or G_{Y_A}): 's' denotes the stride of convolutional layers; 'Out Size' is the size of the feature map in convolutional layers; 'Block3' contains 3 residual blocks.

Layers	**Out Size**	**Configurations**
Block1	240×240	$7 \times 7\ conv, 32, s\ 2$
Block2	120×120	$3 \times 3\ conv, 64, s\ 2$
Block3	120×120	$\begin{bmatrix} 3 \times 3\ conv, 64 \\ 1 \times 1\ conv, 64 \end{bmatrix} \times 3, s\ 1$
Block4	240×240	$3 \times 3\ deconv, 64, s2$
Block5	480×480	$3 \times 3\ deconv, 32, s2$
Block6	480×480	$7 \times 7\ conv, 3, s\ 1$

Table 2. The structure of G_{X_B} (or G_{Y_B}): 's' denotes the stride of convolutional layers; 'Out Size' is the size of the feature map in convolutional layers; 'Block4' contains 5 residual blocks.

Layers	**Out Size**	**Configurations**
Block1	480×480	$7 \times 7\ conv, 32, s\ 1$
Block2	240×240	$3 \times 3\ conv, 64, s\ 2$
Block3	120×120	$3 \times 3\ conv, 128, s\ 2$
Block4	120×120	$\begin{bmatrix} 3 \times 3\ conv, 256 \\ 1 \times 1\ conv, 256 \end{bmatrix} \times 5, s\ 1$
Block4	240×240	$3 \times 3\ deconv, 128, s2$
Block5	480×480	$3 \times 3\ deconv, 64, s2$
Block6	480×480	$7 \times 7\ conv, 3, s\ 1$

2. Implementation Detail

All input images X are resized to 480×480 as shown in Fig. 2 in the main manuscript. In the localization network, it will be further resized to 128. A *Spatial Code* with a length of 10 in the spatial module S_X is randomly sampled which is passed to two fully-connected layers to generate a feature map of size 128×128. The generated feature map and the input image are then concatenated and passed to the localization network LN_X for spatial transformation prediction. The predicted transformation is then applied to the original image X of size 480×480 by the transformation module T to generate the transformed image T_X as well

Table 3. The structure of the D_X (or D_Y): 's' denotes the stride of convolutional layers; 'Out Size' is the size of feature maps.

Layers	Out Size	Configurations
Block1	240×240	4×4 $conv, 64, s$ 2
Block2	120×120	4×4 $conv, 128, s$ 2
Block3	60×60	4×4 $conv, 256, s$ 2
Block4	30×30	4×4 $conv, 512, s$ 2
Block5	30×30	4×4 $conv, 512, s$ 1

Table 4. The structure of the spatial transformation discriminator D_T: 'FC' denotes fully-connected layers.

Layers	Out Size	Configurations
Block0	9×1	$Resize$
Block1	256	FC
Block2	128	FC
Block3	1	FC

as the transformation map m as illustrated in Fig. 1. The generated T_X and m are further concatenated as the input of generator G_{X_A} to complete the black region. The black region of T_X will be further replaced by the corresponding region in the output of G_{X_A} by:

$$Replaced\, T_X = T_X * m + G_{X_A}(T_X, m) * (1 - m) \quad (1)$$

The replaced T_X is then passed to G_{X_B} for appearance adaptation. If a single generator is used for the completion and appearance adaptation, the adapted image will tend to be blurry as shown in 'Single Generator' in Fig. 2.

For the learning in spatial space, D_X and D_Y will also distinguish the $Adapted\ X$ according to the realism in geometry and appearance spaces, which will further enhance the learning in spatial space. With better realism in spatial space, D_X and D_Y will concentrate on distinguishing the images according to the feature in appearance space, thus driving G_X and G_Y to learn better adaptation in appearance space. With better realism in appearance space, D_X and D_Y will also drive the spatial module to learn better adaptation in spatial space. The coordinated learning in spatial space and appearance space will drive network to achieve the best adaptation performance.

3. Experiment

In the scene text detection experiment, as ICDAR2015 and MSRA-TD500 have larger views compared with ICDAR2013, we crop 480×480 patches around the text region as the training reference according to the bounding box annotations.

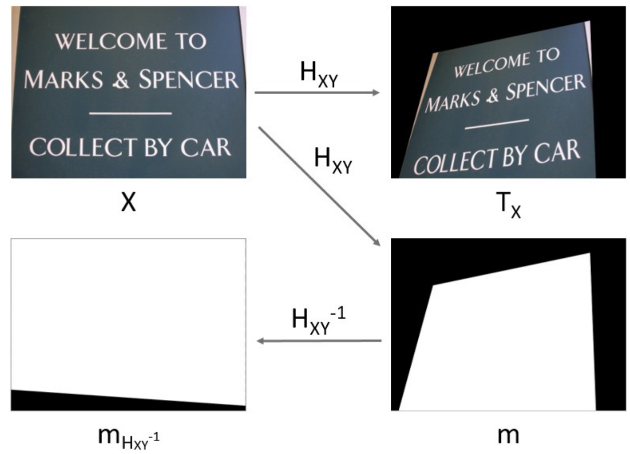

Figure 1. The transformation map and missing region: m and m_{XY}^{-1} are binary transformation maps in which 1 denotes the image region and 0 denotes the padded background. Through the inverse transformation H_{XY}^{-1}, the missing region in the spatial transformation cannot be recovered as shown in $m_{H_{XY}^{-1}}$.

Figure 2. Using a single generator to achieve completion and appearance adaptation will introduce blur as shown in 'Single Generator'. The use of two sub-generators improves the quality of the adapted image significantly as shown in 'Two Generators'.

Figure 3. The ST-GAN will lose the border region. So we constraint the range of the transformation parameters as predicted by the ST-GAN in the test phase, so that the transformed image can preserve all the information of original image as shown in ST-GAN(WC).

The original ST-GAN is for image composition, and we adapted it to achieve image translation in spatial space. As there is no mechanism in ST-GAN to preserve the information of input images, many images will lose their bordering region in spatial transformation as shown in the 'ST-GAN' of Fig. 3. For fair comparison, we constraint the range of the parameters in the transformation matrix so that all the information of the input image can be preserved as show in the 'ST-GAN(WC)' of Fig. 3 in the test phase.

Two NVIDIA GTX 1080TI GPUs are used to train the network with a batch size of 2. The learning rate is initialized with 0.001 and a polynomial decay mechanism of learning rate is applied in the training process. Adam is used as the optimizer.

References

[1] J. Almazn, A. Gordo, A. Forns, and E. Valveny. Word spotting and recognition with embedded attributes. *TPAMI*, 36(12):2552–2566, 2014.

[2] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017.

[3] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou. Focusing attention: Towards accurate text recognition in natural images. In *ICCV*, pages 5076–5084, 2017.

[4] E. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015.

[5] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville. Adversarially learned inference. *arXiv:1606.00704*, 2016.

[6] A. Gordo. Supervised mid-level features for word image representation. In *CVPR*, 2015.

[7] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, A. A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018.

[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[9] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014.

[10] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Deep structured output learning for unconstrained text recognition. In *ICLR*, 2015.

[11] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017.

[12] C.-Y. Lee and S. Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. In *CVPR*, pages 2231–2239, 2016.

[13] C.-H. Lin, E. Yumer, O. Wang, E. Shechtman, and S. Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *CVPR*, 2018.

[14] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017.

[15] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[16] J. A. Rodriguez-Serrano, A. Gordo, and F. Perronnin. Label embedding: A frugal baseline for text recognition. *IJCV*, 2015.

[17] B. Shi, X. Bai, and C. Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *TPAMI*, 39(11):2298–2304, 2017.

[18] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai. Robust scene text recognition with automatic rectification. In *CVPR*, 2016.

[19] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017.

[20] B. Su and S. Lu. Accurate scene text recognition based on recurrent neural network. In *ACCV*, 2014.

[21] B. Su and S. Lu. Accurate recognition of words in scenes without character segmentation using recurrent neural network. *PR*, 2017.

[22] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. In *ICLR*, 2017.

[23] C. Xue, S. Lu, and F. Zhan. Accurate scene text detection through border semantics awareness and bootstrapping. In *Proceedings of the European conference on computer vision (ECCV)*, pages 355–372, 2018.

[24] C. Yao, X. Bai, B. Shi, and W. Liu. Strokelets: A learned multi-scale representation for scene text recognition. In *CVPR*, 2014.

[25] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017.

[26] F. Zhan and S. Lu. Esir: End-to-end scene text recognition via iterative image rectification. In *CVPR*, pages 2059–2068, 2019.

[27] F. Zhan, S. Lu, and C. Xue. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–266, 2018.

[28] F. Zhan, C. Xue, and S. Lu. Ga-dan: Geometry-aware domain adaptation network for scene text detection and recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9105–9115, 2019.

[29] F. Zhan, Y. Yu, K. Cui, G. Zhang, S. Lu, J. Pan, C. Zhang, F. Ma, X. Xie, and C. Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15028–15038, 2021.

[30] F. Zhan, C. Zhang, W. Hu, S. Lu, F. Ma, X. Xie, and L. Shao. Sparse needlets for lighting estimation with spherical transport loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12830–12839, 2021.

[31] F. Zhan, C. Zhang, Y. Yu, Y. Chang, S. Lu, F. Ma, and X. Xie. Emlight: Lighting estimation via spherical distribution approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[32] F. Zhan, H. Zhu, and S. Lu. Spatial fusion gan for image synthesis. In *CVPR*, pages 3653–3662, 2019.

[33] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.