Semi-Supervised Domain Adaptation Using CycleGAN

Sanket Lokegaonkar, Subhashree Radhakrishnan, and Jia Bin Huang

Virginia Tech

Abstract

Domain adaptation has been an interesting field in computer vision where a system learns a model from source distribution and performs well on the target distribution. Real world computer vision comprises of varied model distributions due to which recent studies have shown that tasks such as classification, segmentation show degradation in the state-of-the-art performance. One of the main reasons for this is attributed to domain shift from the biased training datasets. Thus, domain adaptation helps in redressing this problem by learning a common representation between source and target distributions that would reduce the domain shift. Adversarial learning methods are a promising approach to training robust deep networks, and can generate complex samples across diverse domains. [1]. However, adversarial loss suffers from the modal collapse problem and may learn a random distribution that is identical to the target distribution without individual pairing in two domains. The recent techniques proposed [2] consider completely unlabeled target domain and finetune later to achieve better performance. In this work, we propose approach which might alleviate these shortcomings by introducing CycleGAN for domain transfer instead of domain adversarial network. CycleGAN acts as better regularizer due to the addition of reconstruction loss and avoids mode collapse. We perform experiments to investigate the use of CycleGAN for domain adaptation. Qualitative evaluation of the representations learnt by the model and results were reported and analyzed.

1. Introduction

Cross-domain adaptation finds its application on various tasks including Image to Image translations, style transfer, classification and segmentation. The generalized application including all of these is called image-to-image translation [14]. Domain adptation is typically used in tackling problems with unlabeled datasets. Transfer learning and Zero-shot learning are approaches that are similar to cross domain adaptation.

Transfer learning is a problem defined over two domains. These two domains share the same feature space and class label space, but have significantly different distributions. One domain has sufficient labels, named as source domain, and the other domain has few labels, named as tar get do-main. The problem is to learn a effective classifier for the target doma in [12]. However transfer learning does not make use of the unlabeled images in the target domain to train the model.

Zeroshot Learning can be considered as a special case of transfer learning where the source and target do- mains have different tasks/label spaces and the target do- main is unlabelled, providing little guidance for the knowl- edge transfer. [13]. Though zero shot learning operates in an unsupervised setting, it does not take into account any image in the target domain that is unlabeled, during training. The predictions by zero shot learning is more of chance.

Computer vision tasks in a semi-supervised/ unsupervised setting is important problem as most of the real world data is unlabeled.

In this work, we propose *semi-unsupervised* learning framework for segmentation that can be easily extended to generalized image-image translation task, which requires labeled source domain data and partially labeled target domain.

Our work develops an ensemble framework of Cyclic GAN and FCN for segmentation task. By considering two sets of domains wherein the source domain is completely labeled and the target domain is partially labeled. The motivation of the proposed architecture is three fold.

- A stronger regularizer by using a cyclic GAN for the task of domain adaptation.
- Operates in a semi-supervised setting and learns a common representation with the available labels of the target domain.

- To propose a general framework for segmentation in the real world scenario such as self - driving cars where the labeling of real world images is sparse and biased (for e.g. Lesser number of day-night pairs).

2. Related work

Domain Adaptation Domain adaptation in computer vision has focused on the task of reducing domain shift and generalizing across varied distributions [16, 18, 19]. Recent work includes [20, 21, 22] which all learn a feature representation which encourages maximal confusion between the two domains. Other work aims to align the features [23, 24] by minimizing the distance between their distributions in the two domains.

Generative adversarial networks (GANs) Since the seminal work by Goodfellow et al. [25] in 2014, a series of GAN-family methods have been proposed for a wide variety of problems.Generative Adversarial Networks have been recent addition to deep generative models, which instead of parameterizing the distribution , learn it in min-max game formulation of Generator and Discriminator. Generator is tasked with generating data samples and Discriminator with distinguishing real samples from the generated ones. [25]. The following gives a brief description of the various architectures of GAN deployed so far for specific tasks.
In [11] a conditional GAN is used with auxiliary classifier loss that takes as input: latent variable (noise) z where $z \in \mathtt{U}(-1, 1)$ and class label c , learns to output a fake image consistent to the real image coming from class label c. The resolution of the resulting images were comparatively poor as the loss considered was not pixel wise. lesser. Liu *et al*. proposed coupled generative adversarial network to learn a joint distribution of images from both source and target datasets [26]. Taigman et al. [37] proposed unsupervised mechanism for cross-domain image conversion presented by can train an image-conditional generator without paired images, but relies on a sophisticated pre-trained function that maps images from either domain to an intermediate representation, which requires labeled data in other formats. Unlike this work the proposed approach can handle data with sparse labeling and completely unlabeled data. [3]The paper introduces approach for pixel-level prediction in semi-supervised and unsupervised segmentation of images using GAN. The generator takes in random noise z + class information to G(z). Discriminator judges the generated image and generates segmentation map for each class + fake class. This work deals with weakly super vised data and hence requires labels such as bounding boxes. The general-purpose solution for image-to-image translation proposed by Isola et al. [14] requires significant number of labeled image pairs.

[2] proposes a fully convolutional network have been shown to work well for dense prediction like semantic segmentation, but they perform poorly to domain shifts. This work considers fully unsupervised setting and uses a domain adversarial loss for domain shift. In contrast our work is flexible and leverages the presence of partial labels and uses a cyclic consistency loss to enforce domain shift.

Generative models that enforces source data to comply with target data have been used for other tasks include super-resolution [28], texture synthesis [29], style transfer from normal maps to images [40], and video prediction [31], whereas few others were aiming for general-purpose processing image to image translation [14, 37]. But these works do not concentrate on semantic segmentation problem.

3. Method

Our proposed method is designed for the dedicated task of segmentation wherein a cyclic GAN is used for domain shift and FCN is used for segmentation.

3.1. Cycle Consistency Loss

Domain adversarial training learn mappings G and F that produce outputs identically distributed as target domains Y and X respectively where a domain classifier learns a distance function by classifying whether output of the generator belongs to that domain or not. [25]. However, with large enough capacity, a network can map the same set of input images to any random permutation of images in the target domain, where any of the learned mappings can induce an output distribution that matches the target distribution. Thus, an adversarial loss alone cannot guarantee that the learned function can map an individual input x_i to a desired output y_i. To further reduce the space of possible mapping functions, [6] propose that for each image x from domain X, the image translation cycle should be able to bring x back to the original image, i.e. $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. We call this *forward cycle consistency*. Similarly, or each image y from domain Y, G and F should also satisfy *backward cycle consistency*: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. This is the cyclic consistency loss .

3.2. Segmentation Loss

We use segmentation loss i.e pixel wise cross entropy loss between the source and the segmented labeled pairs. The segmentation loss ensures that the generated sementation map is close to the ground truth.

3.3. Objective

A segmentation model is learnt by minimizing the domain shift and penalizing the output when far from the

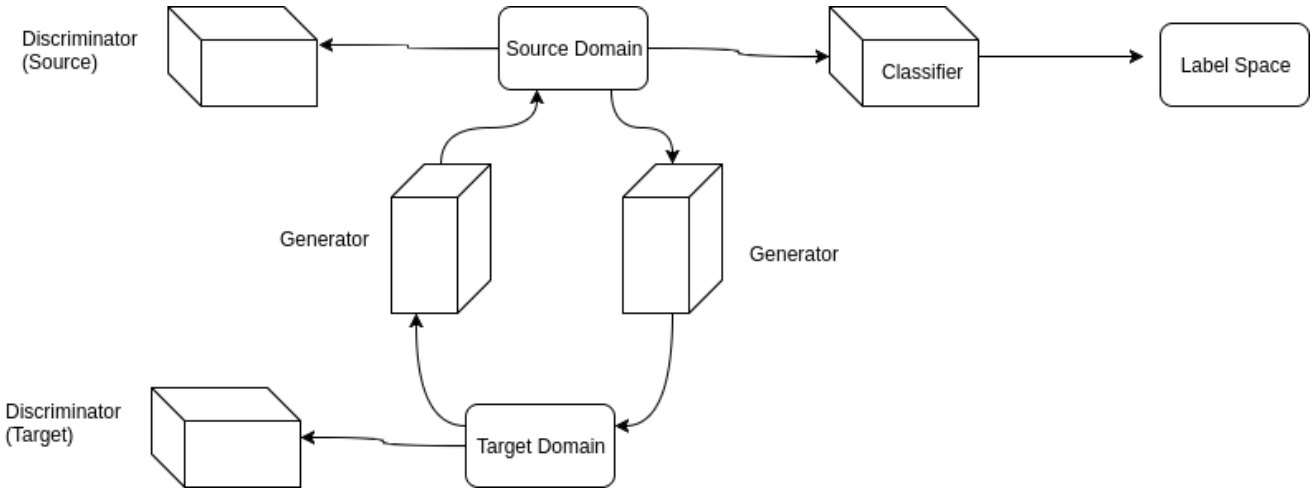

Figure 1: Architecture for the proposed approach of Cycle-GAN+FCN

ground truth segmentation. Thus the overall objective is a combination of the cyclic consistency loss and segmentation loss.

$$\mathcal{L}(I_{\mathcal{B}}, L_{\mathcal{B}}, I_{\mathcal{A}}) = \mathcal{L}_{seg}(I_{\mathcal{B}}, L_{\mathcal{B}}) + \text{cyc}(G, F) + \mathcal{L}_{seg}(I_{\mathcal{A}}, L_{\mathcal{A}})$$

The above equation gives the overall objective where $I_{\mathcal{A}}$, $L_{\mathcal{A}}$ are the target image and labels respectively, $I_{\mathcal{B}}$, $L_{\mathcal{B}}$ are the source domain images and labels . cyc(G, F) represents the cyclic loss. The cyclic loss is given by the following equation. The cycleGAN used is same as defined in [6].

3.4. Network Configuration

A network architecture is same as the one used for cycleGAN [6]. The architecture for cycleGAN is based on generative networks from Johnson et al [7] that were implemented for the tasks of super resolution and neural style transfer. We use identical network architecture for G_A and G_B.This network contains two stride-2 convolutions, several residual blocks [17], and two fractionally-strided convolutions with stride $\frac{1}{2}$. We use 6 blocks for 240×240 images, and higher-resolution training images. Similar to Johnson et al. [7], we use instance normalization [9]. For the discriminator networks we use 70×70 PatchGANs [14, 29, 28], which try to classify whether 70×70 overlapping image patches are real or fake. The overview of the proposed architecture is given in figure 1. Since smaller portions of the input image was used, the network was lesser memory intensive. We had also tested L1 loss for FCN which did not give significant improvement in the performance and hence had used cross entropy loss. The output from the cycleGAN is fed to FCN and the loss is back-propagated throughout the network.

3.5. Training procedure

We pre-train CycleGAN subnetwork in an unsupervised manner using source domain and target domain images. We use the techniques shown with original CycleGAN implementation to stabilize our model training procedure.
First, for LGAN (Equation 2), we replace the negative log likelihood objective by a least square loss [30]. This loss performs more stably during training and generates higher quality results as reported.

$$\mathrm{L}_{LSGAN} = \mathrm{E}_{y \sim p_{data}(y)}[D_Y(y-1)^2] + \mathrm{E}_{x \sim p_{data}(x)}[D_Y(G(x))^2] \quad (1)$$

We experiment with 2 different training approaches for domain adaptation.
1. Sequential Domain Adaptation: We train the Source Domain to Labels for k epochs. Then we train the CycleGAN for k epochs and finally finetune for limited target domain image to label pairs for k epochs.
2. Mixmatch Domain Adaptation: We pool all training pairs Random Source Domain to Target Domain (AB) , Source Domain to Labels (BC) and Target Domain to Labels (AC) together. We train on randomly sampled pairs from the pool.
We update the classifier subnetwork for the BC pair with cross-entropy Loss. We update the CycleGAN subnetwork for AB pair with L_{LSGAN} loss. For the AC Pair, we update the Generator (Target to Source Domain) + Classifier subnetwork with cross entropy loss. The networks are trained on the ADAM optimization solver with batchsize of 1. All networks were trained from scratch, and trained with learning rate of 0.001.

Algorithm 1 Sequential Domain Adaptation training procedure

Require: labeled Image set ac,labeled image set bc, unlabeled image set ab, GAN AB (Source to Target Domain) with generator parameters θ_{AB}, GAN BA (Target to Source Domain) with generator parameters θ_{BA} and discriminator parameters ω_A, and discriminator parameters ω_B, Classifier C with generator parameters θ_C , K epochs

Randomly initialize the generator and discriminator parameters θ_{AB} , θ_{BA} , ω_A, ω_B
repeat
  for i in bc **do**
    Update the Classifier C with cross-entropy loss for labeled pair i
  end for
until K epochs
repeat
  for i in ba **do**
    Update the Cycle GAN Generators GAN AB , GAN BA and Discriminators ω_A, ω_B with L_{LSGAN} and L_{Cycle}
  end for
until K epochs
repeat
  for i in ac **do**
    Update the GAN AB + Classifier with cross-entropy loss
  end for
until K epochs

Algorithm 2 Mixmatch Domain Adaptation training procedure

Require: labeled Image set ac,labeled image set bc, unlabeled image set ab, GAN AB (Source to Target Domain) with generator parameters θ_{AB}, GAN BA (Target to Source Domain) with generator parameters θ_{BA} and discriminator parameters ω_A, and discriminator parameters ω_B, Classifier C with generator parameters θ_C , K epochs

Randomly initialize the generator and discriminator parameters
Pool training pairs from ab, ac, bc into q θ_{AB} , θ_{BA} , ω_A, ω_B
repeat
  for i in q **do**
    if i from ab **then**
      Update the Cycle GAN Generators GAN AB , GAN BA and Discriminators ω_A, ω_B with L_{LSGAN} and L_{Cycle} loss
    end if
    if i from bc **then**
      Update the Classifier C with cross-entropy loss for labeled pair i
    end if
    if i from ac **then**
      Update the GAN AB + Classifier with cross-entropy loss
    end if
  end for
until K epochs

4. Experimental Results and Evaluation

In this section, we will be reporting experimental results on the domain adaptation task: simulated $\rightarrow$ real segmentation .

We initially planned on comparing our proposed method with FCN in the Wild model. But we faced difficulties during the implementation phase for the model. During the domain adversarial training stage, generator overpowered the discriminator, and the discriminator never recovers from it. Note: For training the FCN in the Wild model, We had used the training procedure of Generator, Discriminator pair from CycleGAN due to time constraints.

Because of the above issue, we were only able to perform preliminary experiments comparing our proposed approach: CycleGAN+Classifier with FCN. We have used FCN(fully convolutional network) based on 16 layers VGGNet as our base model for our experiments.

4.1. Datasets:

4.1.1 Cityscapes

Cityscapes contains 34 categories in high resolution, 2048×1024. The whole dataset is divided into three parts: 2, 975 training samples, 500 validation samples and 1, 525 test samples. The split of this dataset is city-level, which covers individual European cities in different geographic and population distribution

4.1.2 GTA5

GTA5 contains 24,966 high quality labeled frames from realistic open-world computer games, Grand Theft Auto V (GTA5). Each frame, with high resolution 1914×1052, is generated from fictional city of Los Santos, based on Los Angeles in Southern California. We take subset of 2000 images with labels compatible to Cityscapes categories for synthetic $\rightarrow$ real adaptation.

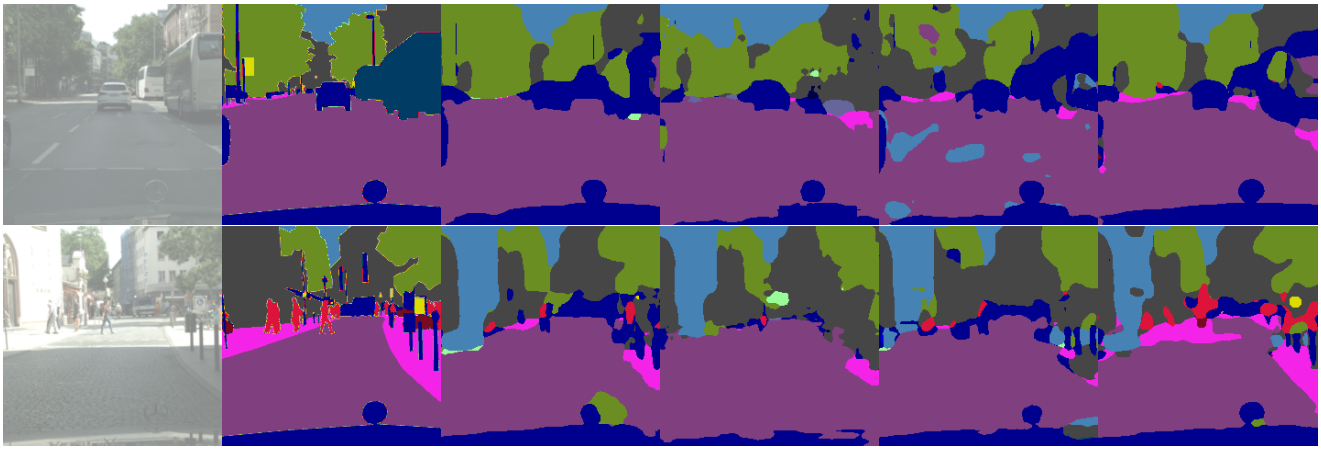

Figure 2: Qualitative results on adaptation from game->real world. [1]:Sequential , [2]: Mismatch Training

4.2. Baselines:

Fully Convolutional Networks(FCN): In this method we train the FCN network only on the simulated image dataset.
Finetuned Fully Convolutional Networks: In this method we train the FCN network on the simulated image dataset plus subset of 200 images from the real-world dataset.
FCN in the Wild: This method attempts to reduce minimize the global distance of feature space between the two domains through adversarial learning of domain discriminator/classifier. We wont be using this method for evaluations in the current draft.

4.3. Segmentation Metrics

To evaluate the performance of simulated $\rightarrow$ real segmentation task, we use the standard metrics from the Cityscapes benchmark, including per-pixel accuracy, per-class accuracy, and mean class Intersection-Over-Union (Class IOU).

4.4. Quantitative Comparison

We ran all the baselines for the same number of epochs. For our two approaches, we pretrained the CycleGAN in an unsupervised manner with combined dataset of simulated and real world dataset of 4000 images.
As seen in the Table, we were not able to produce good results with the baselines. We fall short of 0.2 mean class IoU over the baseline. We see weaker class accuracy for pedestrians , truck , traffic light classes. We believe the lower accuracy is also due to training on less number of epochs.

4.5. Qualitative Comparison

We can see from the segmentation results (Figure 2) that the outlines of larger objects like car are detected but the outlines of smaller objects are not clearly learnt. This issue is potentially due to training on less number of epochs. With closer look on qualitative results of the pretrained CycleGAN (Figure 3), we see that the while transferring from simulated to real-world, some of the trees become badly shaped buildings and higher level semantics such as tree locations/size, are sometimes lost. Pedestrians outlines are maintained but the details and the color is lost in the transition. This effect severly affects the output labels while performing segmentation with FCN.

5. Analysis and Limitations

One of the limitations of the proposed approach is that it does not strongly enforce the higher level semantics are conserved during the domain shift transformation.Only stylistic features are transfered. This, we believe causes poor segmentation results.
We believe one potential approach of enforcing the transformation that retains high level semantics is by ensuring that original source image - label pair retains the label mapping when transferred to the target domain. i.e Given an source domain image-label pair, it should give the same label mapping for F(source image) , where F is transformation from source domain to target domain. We will be exploring this possibility further in the future.

Figure 3: Qualitative results on transformations learnt by CycleGAN.
From left to right, the results show the
RealA: real image from cityscapes(target domain)
FakeB: the image generated by transforming RealA to source domain (game images)
RealB: image from GTAV (source domain),
FakeA represents the image generated by transforming RealB to target domain

	Per-pixel acc.	Per-class acc.	Class IOU
FCN	0.71	0.21	0.14
FinetunedFCN	0.81	0.26	0.20
CycleFCNT1	0.75	0.22	0.16
CycleFCNT2	0.79	0.23	0.18

Table 1: The segmentation accuracy for simulated $\rightarrow$ real world label task. CycleFCNT1 : Proposed architecture with mixmatch training. CycleFCNT2 : Proposed architecture with sequential training. In the training set for FinetunedFCN ,CycleFCNT1 and CycleFCNT2, 300 images from the target domain are added.

6. Conclusion

We propose a novel semi-supervised approach using Cyclic GAN as a better regularizer for the problem of domain adaptation combined with FCN for the task of segmentation. We tested the proposed architecture on the task of segmentation for semi-labeled real world images by shifting the domain to simulated image domain. We report a IoU of 17.9 for the proposed architecture of Cycle GAN+FCN. The semi-supervised approach extends the usability to more real world applications and at the same time leverages partial available labeled data for Real world images.

Though our method was outperformed by Domain Adversarial GAN with FCN [2], we believe that the performance of the proposed architecture can be increased further if trained for greater number of epochs. The proposed architecture is deeper and is trained from scratch. Whereas, the baseline FCN in the wild model is finetuned with initial layers frozen due to which the proposed architecture is prone to vanishing gradient problem. This could be overcome by using techniques such as batch normalization and layer wise pre-training. Future directions include testing cycleGAN with pix2pix for domain adaptation in style transfer/image to-image translation tasks like sketch $\rightarrow$ image, day $\rightarrow$ night apart from segmentation. . Since model collapse is yet another major shortcoming of Domain adversarial training, the recent WGAN and Unrolled GAN can also be experimented. [8] [4].

References

[1] Tzeng, Eric and Hoffman, Judy and Saenko, Kate and Darrell, Trevor Adversarial discriminative domain adaptation. In *arXiv preprint arXiv:1702.05464*, 2017 1

[2] Hoffman, Judy and Wang, Dequan and Yu, Fisher and Darrell, Trevor FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation. In *arXiv preprint arXiv:1612.02649*, 2016 1, 2, 7

[3] Souly, Nasim and Spampinato, Concetto and Shah, Mubarak Semi and Weakly Supervised Semantic Segmentation Using Generative Adversarial Network. In *arXiv preprint arXiv:1703.09695*, 2017 2

[4] Metz, Luke and Poole, Ben and Pfau, David and Sohl-Dickstein, Jascha Unrolled Generative Adversarial Networks. In *arXiv preprint arXiv:1611.02163*, 2016 7

[5] Yi, Zili and Zhang, Hao and Gong, Ping Tan and others DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *arXiv preprint arXiv:1704.02510*, 2017

[6] Zhu, Jun-Yan and Park, Taesung and Isola, Phillip and Efros, Alexei A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv preprint arXiv:1703.10593*, 2017. 2, 3

[7] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 3

[8] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 7

[9] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 3

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[11] Liu, Ming-Yu and Breuel, Thomas and Kautz, Jan Unsupervised Image-to-Image Translation Networks. *arXiv preprint arXiv:1703.00848*, 2017. 2

[12] Wang, Hongqi and Xu, Anfeng and Wang, Shanshan and Chughtai, Sunny Cross domain adaptation by learning partially shared classifiers and weighting source data points in the shared subspaces. *Neural Computing and Applications,1–12*, 2017. 1

[13] Kodirov, Elyor and Xiang, Tao and Fu, Zhenyong and Gong, Shaogang Unsupervised domain adaptation for zero-shot learning. *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 1

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 1, 2, 3

[15] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (TOG)*, 33(4):149, 2014.

[16] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[18] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011. 2

[19] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012. 2

[20] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, 2015. 2

[21] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 2

[22] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016. 2

[23] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 2

[24] M. Long, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *NIPS*, 2016. 2

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[26] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016. 2

[27] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

[28] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016. 2, 3

[29] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016. 2, 3

[30] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[31] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 2

[32] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[33] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[34] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 3, 2016.

[35] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[36] L. Sharan, R. Rosenholtz, and E. Adelson. Material perception: What can you see in a brief glance? *Journal of Vision*, 9(8):784–784, 2009.

[37] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 2

[38] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012.

[39] R. Tyleček and R. Šára. Spatial pattern templates for recognition of objects with regular structure. In *German Conference on Pattern Recognition*, pages 364–374. Springer, 2013.

[40] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*, pages 318–335. Springer, 2016. 2

[41] Kingma, Diederik, and Jimmy Ba. "Adam: A method for stochastic optimization." InarXiv preprint arXiv:1412.6980 (2014).

[42] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2009.

[43] Y. Xia, D. He, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma. Dual learning for machine translation. *arXiv preprint arXiv:1611.00179*, 2016.

[44] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.

[45] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 513–520. IEEE, 2011.