Bilinear Multimodal Discriminator for Adversarial Domain Adaptation with Privileged Information

Taylor Mordan[⋆1], Antoine Saporta[⋆2,3], Alexandre Alahi[1], Matthieu Cord[2,3], Patrick Pérez[3]

***Abstract*— Over the past few years, deep Convolutional Neural Networks have shown outstanding performance on semantic segmentation, which is an essential tool needed by self-driving cars to understand their environments. However, their training relies on large datasets with pixel-level ground truth annotations, which are costly and tedious to produce on real data, making application to new situations difficult. In this context, Unsupervised Domain Adaptation (UDA) from synthetic data is an approach of great interest since it leverages cost-free labeled synthetic datasets to help generalizing to unlabeled real ones. In this paper, we propose a new adversarial training strategy for UDA that uses additional privileged information on the synthetic domain during training to improve transfer to the real one. Our method introduces a multimodal discriminator for adversarial training, featuring a bilinear fusion between representations of segmentation and privileged information to exploit at best alignment between modalities. We evaluate our approach on real-world Cityscapes dataset, using synthetic labeled data with depth as privileged information from SYNTHIA dataset and show competitive results.**

I. Introduction

Robots for last mile mobility, such as self-driving cars or delivery robots, rely on the latest success of deep Convolutional Neural Networks (ConvNets) [1], [2], [3] to solve the fundamental perception step. However, ConvNets need large amounts of labeled images to be trained properly and to avoid overfitting. Even though the standard approach is to pre-train networks on a large-scale dataset, *e.g.*, ImageNet [4], before fine-tuning them on a target dataset adapted to the addressed task, the size of this second dataset is still of practical importance. Unfortunately, having a large number of precisely annotated images can take a prohibitively long time, therefore limiting the sizes of such datasets [5].

A possible solution to this issue is to use synthetic image generation to have a virtually unlimited number of perfectly annotated examples. It would indeed solve the annotation issue, since labeling can be done along with the image generation process, without requiring human input. On the other hand, the synthetic generation is not perfect, and, in the context of images, mismatches between synthetic and real images can appear (*e.g.*, in appearances or distribution of objects [6]). Therefore, models trained on synthetic images usually generalize poorly to real-world scenarios when applied directly. Domain Adaptation [7], [8] addresses this limitation, by jointly training on both kinds of data to improve transfer. In this work, we consider Unsupervised Domain Adaptation (UDA), where only the source domain (synthetic images here) comes with annotations, while the target domain (real images here) is completely unlabeled.

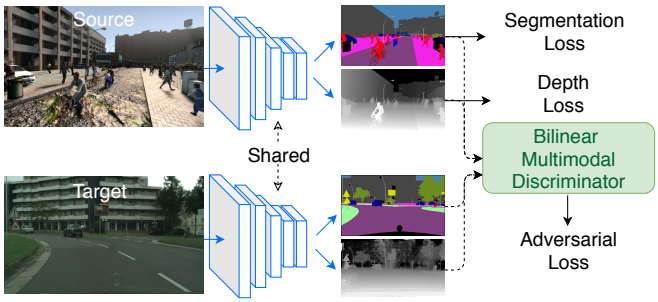


Fig. 1. **Bilinear multimodal adversarial learning.** On the labeled source domain, the network is trained with standard task (segmentation and depth) losses, while on the unlabeled target domain, a bilinear multimodal discriminator (detailed in Figure 2) aligns segmentation and depth predictions into a common representation of both modalities used for adversarial training.

UDA from synthetic data is particularly suited to semantic segmentation due to the heavy annotation cost associated with this task. Another advantage of synthetic imagery is the ease to obtain additional ground truth information, *e.g.*, other modalities, which can be considered as privileged information [9], [10], [11], [12], [13] and used to improve training through a primary Multi-Task Learning (MTL) framework [14] where an additional auxiliary task is associated with each kind of privileged information.

In this paper, we adopt the scenario of adversarial UDA from synthetic images for semantic segmentation with depth as privileged information, in the context of autonomous driving. We argue that proper alignment between segmentation and depth maps should allow a better exploitation of correlations between spatial structures of both modalities. Bilinear models appear to be an interesting tool to achieve it, as they capture fine high-order correlations. We introduce BerMuDA (Bilinear Multimodal Domain Adaptation), a deep ConvNet-based UDA strategy exploring this idea. As illustrated in Figure 1, it contains a bilinear module aligning representations of multiple modalities in order to benefit the adversarial training. Our contributions are threefold: (i) we propose a new way to leverage structured privileged information such as depth maps for UDA, using a fine alignment step; (ii) we improve adversarial learning scheme with multimodal inputs by introducing a bilinear discriminator, generalizing bilinear fusion models to spatial inputs; (iii) we experimentally validate our approach on standard datasets.

⋆ Authors have contributed equally.
[1]EPFL, VITA, CH-1015 Lausanne, Switzerland
[2]Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
[3]valeo.ai, Paris, France

II. Related work

a) Unsupervised Domain Adaptation: Discrepancy-based approaches align statistic distributions on the two domains by introducing a statistical criterion to minimize [15], [16], [17], [18]. Adversarial approaches concurrently train their models with domain discriminators to encourage domain confusion, by encoding both domains in indistinguishable feature spaces [19], [20], [21] or converting source data to the target domain while maintaining corresponding annotations [22], [23], [24]. Reconstruction-based approaches use data reconstruction to regularize and ensure feature invariance, with an encoder-decoder structure [25], [26] or a cyclic mapping between the two domains [27], [28].

b) UDA from synthetic to real for segmentation: Semantic segmentation is often used as the main task for UDA from synthetic to real data, as the labeling requirements are much less in this case than for standard supervised training. Autonomous driving is a hot topic in this regard, as illustrated by the release of multiple datasets of urban scenes, both real [29], [30] and synthetic [31], [32].

Hoffman *et al.* [33] are the first to perform domain alignment using fully convolutional networks with adversarial training to distinguish domains given features. Following works exploit different levels in the segmenter: output predictions [34], with low [35] or all [36] levels in addition. This has been extended to additionally computed features, with reconstructed images [37], intermediate representations converted with GANs [38] and entropy maps [39]. Learning label distribution to identify domains is a common tool to regularize transfer [40], [41]. Scenes have a regular spatial structure that can be leveraged to identify their content [42], [43], [44]. Alignment between domains can be evaluated with the agreement of two classifiers, co-trained [45], [46] or sampled through Dropout [47], or based on the confidence of predictions [48]. Several successful methods use self-supervised learning to address the lack of annotation on target domain, iteratively refining the whole model by using the most confident predictions [49], [50].

c) Domain Adaptation with privileged information: Adversarial UDA using privileged information from synthetic images is studied by Ren *et al.* [51], who use edge, surface normal and depth annotations to learn features more transferable between domains. Regarding semantic segmentation, Lee *et al.* [52] use a GAN to convert images between domains, and leverage depth predictions to regularize the learning of the generator. Closer to our work, Vu *et al.* [53] fuse segmentation and depth predictions to produce depth-aware maps, which are then fed to a discriminator for adversarial training, focusing on closer objects.

d) Bilinear fusion: Bilinear models can fuse multiple modalities into a single representation, as studied for the Visual Question Answering task, where such models have been used with text and image modalities [54], [55], [56], [57]. Bilinear fusions have already been explored in Domain Adaptation too, to adapt more complex class-conditional feature distributions [21], or to align audio-visual multimodal distributions [58], but without privileged information.

III. BerMuDA model

We address UDA with depth as privileged information on the source domain. We build on standard adversarial training for UDA [34], and adapt both the segmentation and discriminator networks to accommodate to the additional modality, as illustrated in Figure 1. In particular, alignment between both modalities is performed with a bilinear fusion integrated within the discriminator.

A. Overview of the approach

We consider two image domains: the synthetic *source* one, $\mathscr{D}_s$, and the real *target* one, $\mathscr{D}_t$. A source example $\left(x_s, y_s^{seg}, y_s^{dep}\right)$ is composed of a synthetic image $x_s \in \mathscr{D}_s$ with the corresponding ground truth segmentation map y_s^{seg} and depth map y_s^{dep}, represented as spatial arrays of one-hot vectors of C classes and of scalar distance values respectively. On the other hand, a target domain example only contains an unlabeled real image $x_t \in \mathscr{D}_t$. Our model is a main network F taking either a source or target image x as input and yielding both a segmentation z^{seg} and a depth map z^{dep} as output, denoted as $\left(z^{seg}, z^{dep}\right) = F(x)$ and illustrated in Figure 1. The model consists of a backbone network with two task-specific prediction branches for segmentation and depth prediction that are learned jointly, as is common practice in multi-task problems [59], [60].

a) Supervised training on source domain: Since annotations are available on the source domain, the model is learned in the standard supervised MTL way for source examples, with the supervised loss $\mathscr{L}_{sup}$ being a linear combination of all task losses:

$$\mathscr{L}_{sup} = \mathscr{L}_{seg} + \lambda_{depth}\mathscr{L}_{depth}, \tag{1}$$

where λ_{depth} is a hyper-parameter balancing both task losses. The segmentation loss $\mathscr{L}_{seg}$ is a pixel-wise cross-entropy loss across all C classes, summed over spatial dimensions:

$$\mathscr{L}_{seg}\left(z_s^{seg}, y_s^{seg}\right) = -\sum_{w,h,c} {y_s^{seg}}_{[w,h,c]} \log\left({z_s^{seg}}_{[w,h,c]}\right). \tag{2}$$

The depth prediction loss $\mathscr{L}_{depth}$ is a reverse Huber regression loss [61] applied on depth predictions z^{dep} in log space, to focus more on closer objects, as in [14]:

$$\mathscr{L}_{depth}\left(z_s^{dep}, y_s^{dep}\right) = \sum_{w,h} \text{berHu}\left({z_s^{dep}}_{[w,h]} - {y_s^{dep}}_{[w,h]}\right), \tag{3}$$

with the reverse Huber function defined by

$$\text{berHu}(e) = \begin{cases} |e| & \text{if } |e| \leq \tau, \\ \frac{e^2+\tau^2}{2\tau} & \text{if } |e| > \tau, \end{cases} \tag{4}$$

τ being a threshold set to 1/5 of the maximum error in the mini-batch.

b) Adversarial training on target domain: Target domain examples are used for learning through an adversarial training procedure as no annotation is available on this domain. For this, a discriminator D is learned concurrently to the main network F, and they compete against each other for optimizing exclusive objectives [62]. Following [34], for an

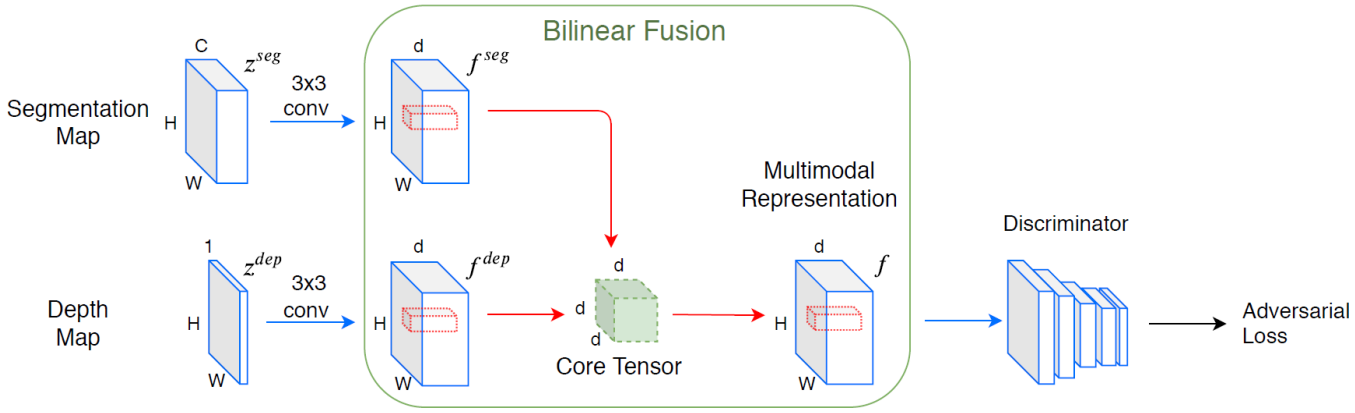


Fig. 2. **Bilinear Multimodal Discriminator.** The discriminator is composed of two parts: a bilinear fusion between segmentation and depth representations, yielding a multimodal representation of both inputs, and a fully convolutional classifier to predict the domain of the input image.

input image x in either domain, the discriminator D takes the output $F(x)$ of network F as input, and makes a prediction $\delta = D(F(x))$ for the domain of x, as is common practice in Domain Adaptation [37]. We here introduce a new kind of discriminator network D, which we call bilinear multimodal discriminator, to handle the multiple modalities output by the network F, as showcased in Figure 1. Its structure is further described in Section III-B. The discriminator D is learned to correctly identify the domain, through minimization of a binary cross-entropy loss $\mathscr{L}_{advD}$:

$$\mathscr{L}_{advD}(\delta, x) = -\mathbb{1}_{\{x\in\mathscr{D}_s\}} \log\left(\delta_{[x\in\mathscr{D}_s]}\right) - \mathbb{1}_{\{x\in\mathscr{D}_t\}} \log\left(\delta_{[x\in\mathscr{D}_t]}\right). \quad (5)$$

At the same time, in order to obtain predictions more transferable across domains, the network F is trained to confuse the discriminator D by minimizing an adversarial binary cross-entropy loss $\mathscr{L}_{advF}$ on target domain examples:

$$\mathscr{L}_{advF}(\delta, x) = -\mathbb{1}_{\{x\in\mathscr{D}_t\}} \log\left(\delta_{[x\in\mathscr{D}_s]}\right). \quad (6)$$

c) Full training on both domains and deployment: The full loss function $\mathscr{L}_F$ minimized by the network F is then the weighted sum of the supervised loss (Equation (1)) and the adversarial loss (Equation (6)):

$$\mathscr{L}_F = \mathscr{L}_{seg} + \lambda_{depth}\mathscr{L}_{depth} + \lambda_{adv}\mathscr{L}_{advF}, \quad (7)$$

with λ_{adv} a hyper-parameter controlling the weight of the adversarial loss. The training therefore alternates between optimizing the network F (Equation (7)) and the discriminator D (Equation (5)).

When deployed, our model does not induce any overhead compared to a standard segmentation network, as the depth branch is not used. The associated auxiliary task is leveraged during training only, under the primary MTL framework [14].

B. Bilinear Multimodal Discriminator

The discriminator D, detailed in Figure 2, is composed of a bilinear fusion and a fully convolutional classifier. The last part is commonly used alone in adversarial training, and we here add a prior merging layer to handle multimodal inputs.

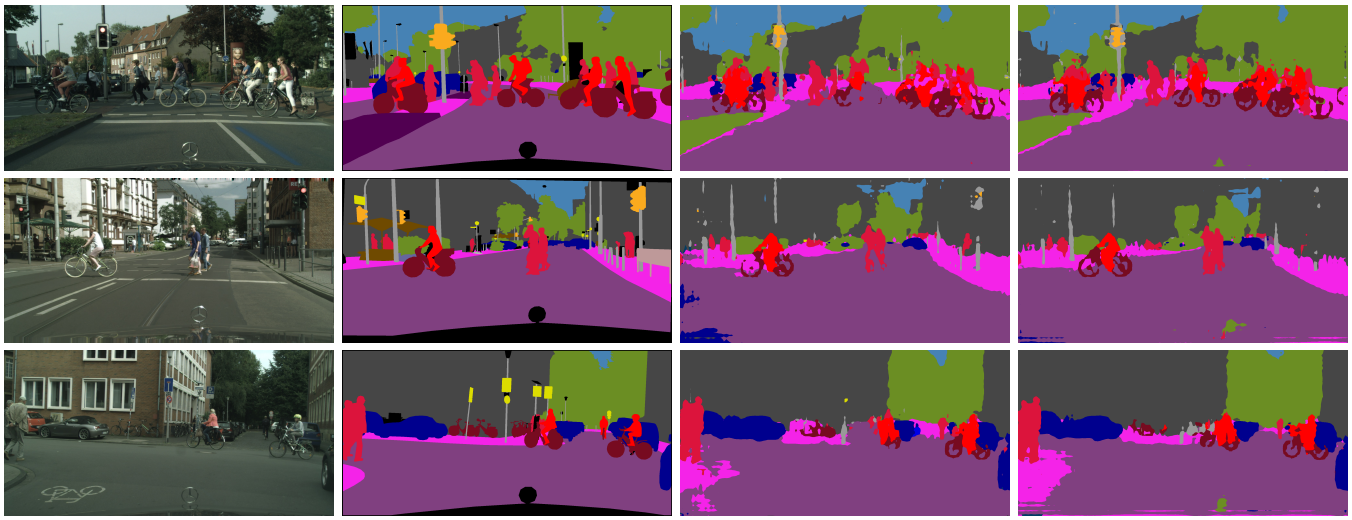
This additional step takes both segmentation z^{seg} and depth z^{dep} predictions from F as inputs, and projects them onto f^{seg} and f^{dep}, in a feature space of higher dimension d through two separate 3×3 spatial convolutions. Using convolutions instead of pointwise mappings [56], [57] is a way to generalize to spatial inputs, and therefore to locally aggregate spatial context, which is especially useful to handle depth prediction z^{dep} composed of a single channel. We then rely on the block-diagonal fusion of [57], with K full-rank blocks, to perform the actual alignment, represented by the green core tensor in Figure 2 (note that the K-block decomposition is not shown for clarity). This decomposition allows modelling fine interactions while still controlling the number of parameters in this core tensor. At each spatial position (w,h), ${f^{seg}}_{[w,h]}$ and ${f^{dep}}_{[w,h]}$ are fused into $f_{[w,h]}$ through the action of the core tensor, shown with red arrows in Figure 2. For this, they are uniformly divided into K chunks ${\tilde{f}_k^{seg}}_{[w,h]}$ and ${\tilde{f}_k^{dep}}_{[w,h]}$. Each pair of features is then bilinearly fused to yield a multimodal feature $\tilde{f}_{k[w,h]}$:

$$\tilde{f}_{k[w,h]} = \left({\tilde{f}_k^{seg}}_{[w,h]}\right)^\top A_k \left({\tilde{f}_k^{dep}}_{[w,h]}\right) + b_k, \quad (8)$$

with A_k and b_k the weight matrix and bias learned for chunk k in the core tensor. The final multimodal representation $f_{[w,h]}$ at spatial position (w,h) is the concatenation of all K $\tilde{f}_{k[w,h]}$. The complete multimodal map f is then fed into the fully convolutional classifier to output the prediction δ of the domain of input image.

Discussion

SPIGAN [52] and DADA [53] are two related approaches, also doing UDA from synthetic images and using depth as privileged information. The first one uses depth as an additional way to regularize the training of the generator that translates images from the source domain to the target one. Privileged information is therefore not directly used to enhance the segmentation network, which should yield less transfer between tasks. On the other hand, DADA follows an idea similar to BerMuDA, but with additional features in both the main network architecture and the adversarial training procedure. Indeed, it can also be interpreted as a multimodal



(a) Image (b) Ground truth (c) AdaptSegNet (d) BerMuDA

Fig. 3. **Qualitative segmentation results.** Input images are presented in (a) with corresponding ground truths in (b). We show segmentation predictions by the baseline AdaptSegNet [34][2]in (c) and by our proposed model BerMuDA in (d). It is noticeable that BerMuDA obtains finer segmentation masks than AdaptSegNet for some practically important classes, such as the *pedestrian* and *rider* ones in these examples.

TABLE I

COMPARISON OF MODALITY FUSION APPROACHES IN MEAN INTERSECTION OVER UNION (%).

Model				Results	
Name	Approach used in	Privileged information	Modality fusion	mIoU-16	mIoU-13
Segmentation discriminator	AdaptSegNet [34]	-	-	39.0	45.6
Independent discriminators	-	✓	-	39.0	45.9
Joint concatenation	-	✓	concatenation	39.3	46.0
Joint product	DADA [53]	✓	product	39.5	46.3
Joint bilinear	BerMuDA [ours]	✓	bilinear	40.2	47.0

discriminator, with the main difference lying in the fusion operation used. While DADA uses an element-wise product, which only focuses on closer objects, BerMuDA opts for a bilinear transformation. Overall, the main idea behind a bilinear multimodal discriminator is to optimize over a large family of bilinear functions applied on segmentation and depth predictions, encompassing a wide variety of relevant representations and generalizing the element-wise product.

IV. EXPERIMENTS

A. Experimental setup

a) Datasets: As source domain, we use SYNTHIA dataset [31] which is composed of synthetic images annotated with pixel-wise semantic labels as well as depth maps. In particular, we adopt the split SYNTHIA-RAND-CITYSCAPES that matches the Cityscapes annotation style. For the target domain, we use Cityscapes dataset [29], without any annotation. The models are trained on the union of SYNTHIA and Cityscapes training sets with the common classes between them, and are evaluated on the 16-class and 13-class (excluding *wall*, *fence* and *pole* classes) subsets, reported as the 'mIoU-16' and 'mIoU-13' metrics respectively, of Cityscapes validation set.

b) Implementation details: As it is common practice, we adopt DeepLab-V2 [63] as the base semantic segmentation architecture, with a ResNet-101 [64] backbone model pre-trained on ImageNet [65]. The predictors for segmentation and depth estimation are two separate Atrous Spatial Pyramid Pooling (ASPP) modules applied in parallel on the last layer from the backbone. The discriminator is composed of a bilinear fusion block with $K = 50$ full-rank chunks in dimension $d = 200$, followed by five convolutional layers with kernel 4×4, stride 2, $\{64, 128, 256, 512, 1\}$ channels respectively, and leaky ReLU as activation function.

The main network is learned with SGD, with an initial learning rate of 2.5×10^{-4} polynomially decayed with a factor of 0.9, a momentum of 0.9 and a weight decay of 5×10^{-4}. The loss weights are set to $\lambda_{depth} = 0.001$ and $\lambda_{adv} = 0.001$. The discriminator is trained with Adam, with 10^{-4} as initial learning rate for the same polynomial decay, and $(0.9, 0.99)$ for momentum. Each mini-batch contains one source image of size 1280×760px and one target image of size 1024×512px.

[2]Results obtained by running the code released by the authors of [34] at `https://github.com/wasidennis/AdaptSegNet`

B. Results and comparison between approaches

In this work, we study several ways to leverage privileged depth information for UDA and to integrate it within a simple baseline, in order to focus on this aspect with fair comparisons. The improvement brought by BerMuDA is detailed in Table I.

The first row shows the baseline method, where no depth information is used, and the discriminator is applied on segmentation output only. This correspond to the approach used by AdaptSegNet [34]. All subsequent rows leverage depth in different ways. On the second row is a model integrating depth in a simple way, with two independent discriminators applied on segmentation and depth outputs respectively. Its results are comparable with the baseline's ones, with a slight improvement of 0.3 points only in mIoU-13, and show that adding depth supervision does not improve transfer between domains by itself if used in a naive way. The last three rows present several variants of a joint discriminator, taking both modalities as input, with differences lying in the way to fuse them. The first version uses a simple concatenation and the second an element-wise product, similar to DADA [53]. The last one is our proposed model, with a bilinear discriminator. It yields the best performance, with improvements of 1.2 and 1.4 points in both metrics with respect to not using depth information. The other two variants of a joint discriminator have more limited improvements, indicating that the alignment step is useful to leverage depth effectively. A qualitative visualization of segmentation results is displayed in Figure 3.

It is noticeable that latest approaches, *e.g.*, DISE [44] or BDL [50], explore both pseudo-labeling and CycleGAN-based image translation strategies, which yield great results but are also computationally heavy to train. Since they do not affect the architecture of the discriminator, they should be complementary to the use of depth as privileged information through a multimodal discriminator.

V. Conclusions

In this paper, we have introduced BerMuDA to address Unsupervised Domain Adaptation in the presence of privileged information. Under an adversarial training framework, its discriminator first aligns all modalities thanks to a bilinear fusion operation. This alignment step enables the learning of the privileged modality to better transfer to the adaptation between domains. We have applied BerMuDA on semantic segmentation with depth ground truth as privileged information, on a synthetic-to-real adaptation scenario, with learning on SYNTHIA dataset and evaluation on Cityscapes dataset. In addition, results could certainly be further improved by integrating BerMuDA into more recent frameworks that should be complementary to our contributions.

Acknowledgment

The authors thank Tuan-Hung Vu for helpful discussions and advice. This work was partially funded by grant DeepVision (ANR-15-CE23-0029, STPGP-479356-15), a joint French/Canadian call by ANR & NSERC.

References

[1] K. Fukushima, "Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[2] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[3] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[5] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2016.

[6] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2107–2116.

[7] G. Csurka, *Domain Adaptation in Computer Vision Applications*, ser. Advances in Computer Vision and Pattern Recognition. Springer, 2017.

[8] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[9] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks*, vol. 22, no. 5-6, pp. 544–557, 2009.

[10] V. Sharmanska, N. Quadrianto, and C. Lampert, "Learning to transfer privileged information," in *arXiv:1410.0389*, 2014.

[11] V. Vapnik and R. Izmailov, "Learning using privileged information: Similarity control and knowledge transfer," *Journal of Machine Learning Research (JMLR)*, vol. 16, pp. 2023–2049, 2015.

[12] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 826–834.

[13] Z. Shi and T.-K. Kim, "Learning and refining of privileged information-based RNNs for action recognition from depth sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] T. Mordan, N. Thome, G. Henaff, and M. Cord, "Revisiting multi-task learning with ROCK: a deep residual auxiliary block for visual detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[15] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv:1412.3474*, 2014.

[16] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[17] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2016.

[18] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[19] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

[20] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[22] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] Z. Wu, X. Han, Y.-L. Lin, M. Gokhan Uzunbas, T. Goldstein, S. Nam Lim, and L. Davis, "DCAN: Dual channel-wise alignment networks for unsupervised scene adaptation," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2018, pp. 518–534.

[25] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2016.

[26] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[27] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[28] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[29] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. [Online]. Available: https://www.mapillary.com/dataset/vistas

[31] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2016, pp. 102–118.

[33] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "FCNs in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv:1612.02649*, 2016.

[34] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[35] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully convolutional adaptation networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6810–6818.

[36] H. Huang, Q. Huang, and P. Krahenbuhl, "Domain transfer through deep activation matching," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2018, pp. 590–605.

[37] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[38] W. Hong, Z. Wang, M. Yang, and J. Yuan, "Conditional generative adversarial network for structured domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1335–1344.

[39] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] Y. Zhang, P. David, H. Foroosh, and B. Gong, "A curriculum domain adaptation approach to the semantic segmentation of urban scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[41] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. Frank Wang, and M. Sun, "No more discrimination: Cross city adaptation of road scene segmenters," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1992–2001.

[42] Y. Chen, W. Li, and L. Van Gool, "ROAD: Reality oriented adaptation for semantic segmentation of urban scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7892–7901.

[43] F. Saleh, M. Sadegh Aliakbarian, M. Salzmann, L. Petersson, and J. Alvarez, "Effective use of synthetic data for urban scene semantic segmentation," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2018, pp. 84–100.

[44] W.-L. Chang, H.-P. Wang, W.-H. Pengg, and W.-C. Chiu, "All about structure: Adapting structural information across domains for boosting semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[45] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3723–3732.

[46] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[47] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Adversarial dropout regularization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[48] X. Zhu, H. Zhou, C. Yang, J. Shi, and D. Lin, "Penalizing top performers: Conservative loss for semantic segmentation adaptation," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2018, pp. 568–583.

[49] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2018, pp. 289–305.

[50] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6936–6945.

[51] Z. Ren and Y. J. Lee, "Cross-domain self-supervised multi-task feature learning using synthetic imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[52] K.-H. Lee, G. Ros, J. Li, and A. Gaidon, "SPIGAN: Privileged adversarial learning from simulation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[53] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "DADA: depth-aware domain adaptation in semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[54] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[55] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[56] H. Ben-Younes, R. Cadène, N. Thome, and M. Cord, "MUTAN: Multimodal tucker fusion for visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[57] H. Ben-Younes, R. Cadene, N. Thome, and M. Cord, "BLOCK: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[58] F. Qi, X. Yang, and C. Xu, "A unified framework for multimodal domain adaptation," in *Proceedings of the ACM International Conference on Multimedia*, 2018, pp. 429–437.

[59] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[60] I. Kokkinos, "UberNet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[61] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks,"

in *Proceedings of the IEEE International Conference on 3D Vision (3DV)*, 2016, pp. 239–248.

[62] 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 (NIPS)*, 2014, pp. 2672–2680.

[63] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[65] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.