Real-time Domain Adaptation for Semantic Segmentation using Adversarial Training and Lovász-Softmax Loss

Amirhossein Lotf Ranaei
amirhossein.lotfranaei@studenti.polito.it

Abstract

Semantic segmentation is a fundamental task in computer vision, but the high cost of annotating pixel-level data has driven up interest in using synthetic datasets. However, models trained on synthetic data often fail to generalize to real-world images due to a problem known as domain shift. In this project, we investigate unsupervised domain adaptation (UDA) for the task of real-time semantic segmentation. We use the BiSeNet architecture, a real-time segmentation network, and address the domain gap between the synthetic GTA5 dataset and the real-world Cityscapes dataset. We first measure the performance drop caused by the domain shift, explore the use of data augmentations to reduce it, and improve the generalization to real-world images. We then implement an adversarial UDA framework for domain adaptation. Additionally, as an extension, we enhance the adversarial framework by incorporating the Lovász-Softmax loss, creating a composite segmentation objective that directly optimizes the Intersection-over-Union (IoU) metric alongside the standard cross-entropy loss. Our experiments show that strong photometric augmentation significantly improves the performance, but standard adversarial training does not improve performance further. However, our proposed extension with the hybrid Lovász-Softmax loss improves the final mean IoU to 33.04% on the Cityscapes validation set. This result shows a significant overall improvement from the initial baseline and demonstrates the effectiveness of combining adversarial adaptation with a task-specific loss function, particularly to improve performance on underrepresented classes.

1. Introduction

Semantic segmentation is the task of assigning a semantic label to every pixel of an image; It is an important part of modern computer vision systems and it's used in fields such as autonomous driving, medical imaging, and robotics. Thanks to deep convolutional neural networks (DCNNs), semantic segmentation has consistently pushed the state of the art [4]. However, training these deep networks requires a huge amount of pixel-perfect annotated data. The manual creation of such datasets is expensive, time-consuming, and even prone to human error, forming a major bottleneck in the development pipeline.

A popular and cost-effective alternative is to leverage large-scale, automatically annotated synthetic datasets generated from modern computer games, like the solution that was proposed in the paper *Playing for Data: Ground Truth from Computer Games* [6] where they managed to create pixel-accurate semantic label maps for images by extracting the ground truths from *Grand Theft Auto V*. Although these synthetic datasets are rich in detail and provide a perfect ground truth, they cannot generalize well to real-world data and face a significant performance drop when compared to them [10]; this performance drop is known as the *Domain shift* and is caused by differences in textures, lighting models, object geometries, and overall environmental distributions.

To bridge the semantic and visual gap, in this project, we explore Unsupervised Domain Adaptation (UDA). The goal of UDA is to adapt a model trained on a labeled synthetic (source) domain to an unlabeled real-world (target) domain, without requiring any manual annotations from the target domain. We specifically focus on a real-time network, BiSeNet [9], due to its direct relevance in applications like autonomous driving, where low-latency inference is needed. Our goal for this project is to first assess the Domain Shift in an urban-street environment, specifically for a real-time solution employed in autonomous driving; we will use the Cityscapes dataset [3] as the test dataset throughout this project.

Our systematic approach was the following:

1. Establishing an upper and lower performance bound by training the BiSeNet model first on real-world data and then on a synthetic dataset to quantify the domain shift.

2. Implement and analyze the results and the logical validity of various data augmentation techniques as a simple yet effective way of reducing the domain gap.

3. For our choice of UDA technique, we chose to implement an adversarial adaptation network as proposed by Tsai et al. [[7]], in which a discriminator guides the segmentation network to learn domain-invariant features

4. As an extension, we enhanced the segmentation objective by incorporating the **Lovász-Softmax loss**, a direct proxy for the Intersection over Union (IoU) metric, to further enhance segmentation quality.

2. Related Work

2.1. Semantic segmentation

The field of semantic segmentation was revolutionized with the introduction of Fully Convolutional Networks (FCN), which allowed end-to-end training for dense prediction. Subsequently, "classic" architectures like DeepLabV2 [2] introduced key innovations such as atrous (dilated) convolutions to expand the receptive field to capture multi-scale context without increasing the number of parameters or computational cost. DeepLabV2 also popularized the use of Conditional Random Fields (CRFs) as a post-processing step to refine segmentation boundaries, which leads to high accuracy but at the cost of significant computational overhead, making it unsuitable for real-time use.

2.2. Real time segmentation

For applications like autonomous driving, inference speed is as important and critical as accuracy. This has led to the development of lightweight, efficient architectures that are designed to run in real-time. BiSeNet [9] is an interesting example of this kind of architecture. It cleverly addresses the trade-off between rich spatial information and a large receptive field by featuring a two-path design:

- **Spatial Path (SP):** This path processes the original high-resolution input image with a shallow network to maintain fine spatial details and boundaries.

- **Context Path (CP):** This path uses a fast downsampling strategy with a lightweight backbone (like ResNet18) to obtain a large receptive field in order to capture high-level semantic context.

The features from both paths are fused by a Feature Fusion Module, allowing the network to achieve a strong balance between performance and speed.

2.3. Unsupervised Domain Adaptation (UDA)

UDA aims to transfer knowledge from a labeled source domain (e.g., GTA5 [6]) to an unlabeled target domain (e.g., Cityscapes [3]). According to a recent review by Zhao et al. [10], UDA techniques are broadly categorized into two main approaches.

- **Image-to-Image Translation:** This approach aims to modify the images of the source domain to match the style and appearance of the target domain using models like Generative Adversarial Networks (GANs). The translated source images can then be used to train a segmentation model in a supervised fashion. Methods like FDA [8] perform this adaptation efficiently in the frequency domain.

- **Adversarial Feature Adaptation:** This is the focus of our project. Inspired by GANs, this approach uses a discriminator network at the feature or output level. The discriminator is trained to distinguish between features extracted from source and target images, while the main segmentation network (the generator) is trained to fool the discriminator. This minimax game encourages the generator to learn features that are both effective for segmentation and invariant across domains. The work by Tsai et al. [7] is a foundational example of this method for the structured output space of semantic segmentation.

3. Methodology

3.1. Baseline Network

Throughout this project, we use the BiSeNet [9] architecture with a ResNet18 backbone pre-trained on ImageNet. Its dual-path structure makes it a suitable candidate for real-time applications where a trade-off between accuracy and efficiency is necessary.

3.2. Adversarial Domain Adaptation

To align the feature distributions of the source (GTA5) and target (Cityscapes) domains, we implement an adversarial UDA framework based on Tsai et al. [7]. This framework consists of two main components:

- **Generator (G):** The BiSeNet network, which takes an image I as input and produces a pixel-wise probability map $P \in \mathbb{R}^{H \times W \times C}$, where C is the number of classes.

- **Discriminator (D):** A fully convolutional network that takes a probability map P as input and predicts its domain of origin. Our discriminator is composed of 5 convolutional layers with a kernel size of 4x4 and LeakyReLU activations. As a fully convolutional network, it outputs a spatial map of domain predictions, and the loss is computed over all locations in this map, consistent with the approach in [7].

The networks are trained via a minimax game. The discriminator is trained to minimize the binary cross-entropy (BCE) loss for classifying source predictions as "real" (label 1) and target predictions as "fake" (label 0). The loss for a mini-batch of size M is:

$$\mathcal{L}_D = -\frac{1}{M}\sum_{i=1}^{M}\left[\log D(G(x_s^{(i)})) + \log(1 - D(G(x_t^{(i)})))\right] \tag{1}$$

where $x_s^{(i)}$ and $x_t^{(i)}$ are the i-th images from the source and target domain batches, respectively.

The generator is trained to minimize a composite loss, consisting of a segmentation loss on labeled source data and an adversarial loss designed to fool the discriminator:

$$\mathcal{L}_G = \mathcal{L}_{\text{seg}}(G(x_s), Y_s) + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}(G(x_t)) \tag{2}$$

Here, Y_s is the ground truth for the source image, and the adversarial loss is $\mathcal{L}_{\text{adv}} = -\frac{1}{M}\sum_{i=1}^{M}\log D(G(x_t^{(i)}))$, which encourages the generator to produce target predictions that the discriminator classifies as "real". The hyperparameter λ_{adv} balances the two objectives and was set to 0.002 in our experiments.

3.3. Extension: Hybrid Segmentation Loss with Lovász-Softmax

The standard segmentation loss, $\mathcal{L}_{seg}$, is typically the pixel-wise cross-entropy (CE) loss. While effective, CE is an imperfect proxy for the Jaccard Index (Intersection-over-Union), which is the primary metric for evaluating segmentation quality, especially in scenarios with significant class imbalance. As our extension, we enhance the segmentation loss by incorporating the Lovász-Softmax loss, proposed by Berman et al. [1]. This loss is a differentiable surrogate for the IoU metric, designed for its direct optimization. Instead of replacing CE entirely, we form a hybrid loss that leverages the stability of CE with the metric-focused optimization of Lovász-Softmax. Our final segmentation loss objective is a weighted sum:

$$\mathcal{L}_{seg} = \mathcal{L}_{CE} + w_{lovasz} \cdot \mathcal{L}_{Lovasz} \tag{3}$$

Where w_{lovasz} is a weighting factor. We hypothesize that by optimizing this composite objective, the model can learn more robust and accurate segmentation boundaries, particularly for smaller or underrepresented classes that are often marginalized by the standard CE loss.

4. Experiments and results

4.1. Experimental Setup

- **Datasets:** We use the GTA5 dataset [6] (2500 images) as the source domain, with images resized to 1280 x 720 pixels for training. The Cityscapes dataset [3] (1572 training images, 500 validation images) serves as the target domain, with all images resized to 1024 x 512 pixels for both the adversarial training phase and final validation. GTA5 labels are mapped to the 19 common classes of Cityscapes using a pre-processing step.
- **Training Details:** All models are trained for 50 epochs using the Adam optimizer [5] and a batch size of 8. The initial learning rate for the generator (BiSeNet) was set to $1e^{-4}$, while the initial learning rate for the discriminator was set to $2.5e^{-5}$. A polynomial learning rate decay scheduler with a power of 0.9 was used. For our extension, the weight for the Lovász-Softmax loss component, w_{lovasz}, was set to 0.5.
- **Implementation:** Our implementation is based on PyTorch. All training was conducted on a single NVIDIA A100 GPU using mixed-precision computation for efficiency. The complete source code is available at: `https://github.com/arman-rn/MLDL-SemSeg`.

4.2. Baselines and Domain Shift Analysis

To establish clear performance upper and lower bounds, we first evaluated a classic, high-performance model (DeepLabV2 [2]) and a real-time network (BiSeNet) on the Cityscapes dataset. The results are shown in Table 1.

Model	mIoU (%)	Latency (ms)	FLOPs (G)	Params (M)
DeepLabV2	63.94	23.45	374.66	43.90
BiSeNet	**54.86**	**5.23**	**25.78**	**12.58**

Table 1. Performance of classic and real-time networks trained and tested on Cityscapes. BiSeNet proves to be a strong choice for a real-time baseline as it trades accuracy for vast improvements in speed (seen in the number of parameters).

Next, to quantify the domain shift, we trained BiSeNet on the synthetic GTA5 dataset and evaluated it on the Cityscapes validation set. As shown in Table 2, the performance dropped drastically from 54.86% to 22.47% mIoU, a relative decrease of nearly 60%. The per-class scores show that performance is particularly poor for smaller, less common classes like 'bicycle', 'rider', and 'motorcycle', which are almost completely missed. This highlights the severity of the domain gap.

4.3. Mitigating Domain Shift with Data Augmentation

We investigated four data augmentations to reduce the domain shift: 'HorizontalFlip', 'ColorJitter', 'ISONoise', and 'CoarseDropout'. Our experiments showed that photometric augmentations were far more effective than other

types of augmentations, as shown in Table 2. ‘ColorJitter‘ alone provided a massive boost, improving mIoU from 22.47% to 31.62%. On the other hand, ‘HorizontalFlip‘ slightly degraded performance to 21.90%, suggesting that some geometric invariances learned from synthetic data may not transfer well to the real domain’s specific camera poses. Our best combination of augmentations, consisting of ‘ColorJitter‘, ‘ISONoise‘, and ‘CoarseDropout‘, achieved an mIoU of **31.97%**, recovering over 9 mIoU points and serving as a very strong baseline for domain adaptation.

4.4. Adversarial Domain Adaptation

Building on the best augmentation setting, we trained our adversarial UDA model. The training dynamics are visualized in Figure 1. The plots show the generator’s segmentation loss steadily decreasing while the adversarial losses for the generator and discriminator fluctuate, indicating a stable minimax game where neither player overpowers the other.

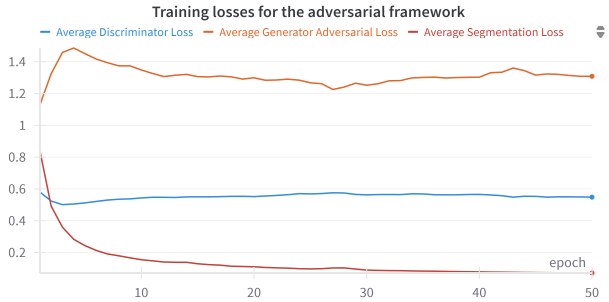


Figure 1. Training losses for the adversarial framework. It shows the smoothed per-epoch losses for segmentation, the generator’s adversarial loss, and the discriminator loss.

Surprisingly, the final performance of the standard adversarial model was **31.74% mIoU**. As detailed in Table 2, this is a slight decrease compared to the augmentation-only baseline (31.97%).

4.5. Improving Adaptation with Lovász-Softmax Loss

Finally, we trained the adversarial model using our proposed hybrid segmentation loss ($L_{\text{seg}} = L_{\text{CE}} + w_{\text{lovasz}}.L_{\text{Lovasz}}$). This model achieved a final mIoU of **33.04%**, outperforming both the augmentation-only and the standard adversarial models. The comparison of validation mIoU curves in Figure 2 clearly shows that the Lovász-Softmax model achieved consistently higher and more stable mIoU scores throughout the later stages of training. The full results for all experimental stages are compiled in Table 2.

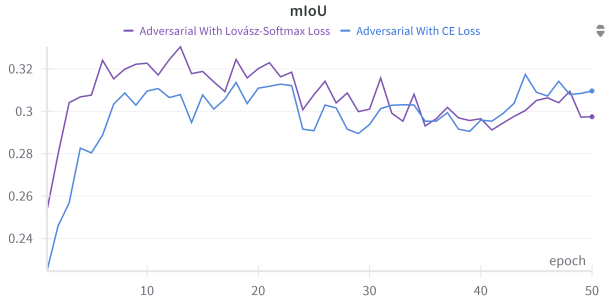


Figure 2. Comparison of validation mIoU over 50 epochs.

5. Discussion and Analysis

Our experiments provide a clear narrative of tackling the domain shift problem. The initial baseline established a severe performance gap of over 32 mIoU points, confirming that models trained on synthetic data do not generalize well to real images.

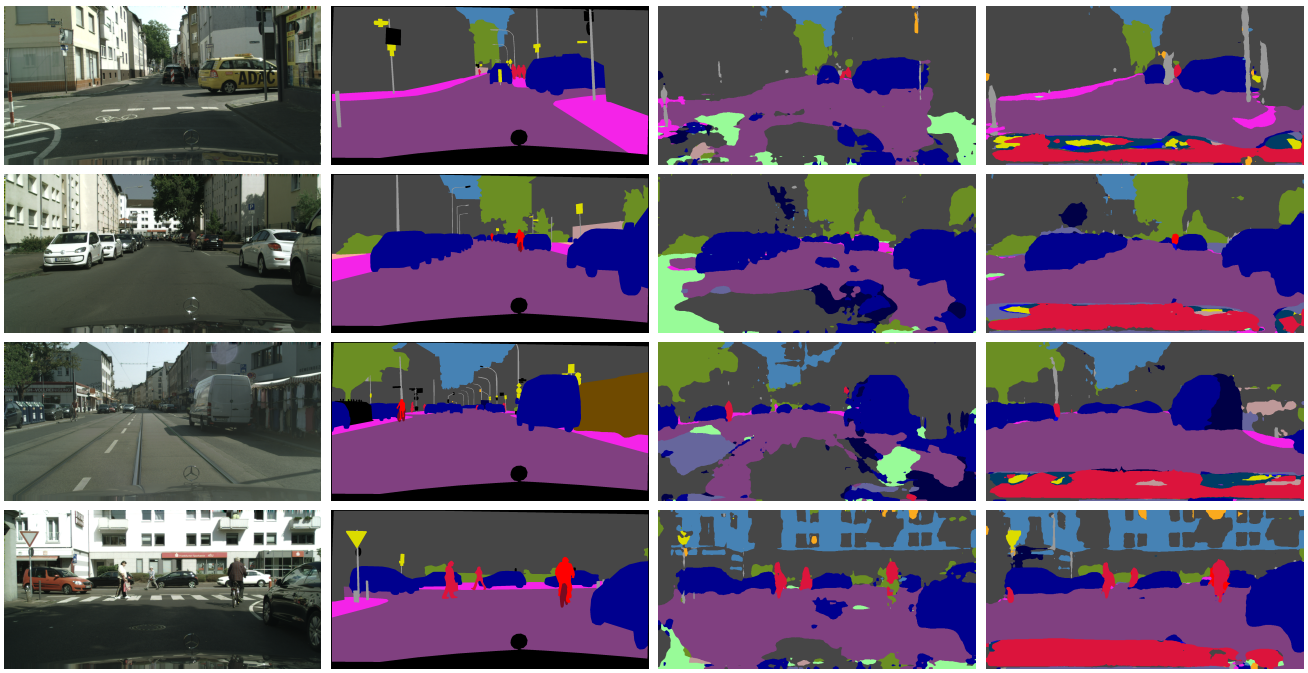
The most striking initial finding was the power of data augmentation. Photometric augmentations, especially ‘ColorJitter‘, were far more effective at bridging the domain gap than geometric or noise-based ones. This suggests that a primary component of the domain shift between GTA5 and Cityscapes is in the color, lighting, and contrast distributions, which these augmentations directly target. The resulting 31.97% mIoU model served as a very competitive baseline.

The most unexpected result came from the standard adversarial training, which failed to improve upon this strong augmentation baseline. While the training dynamics (Fig. 1) appeared stable, the final mIoU was slightly lower. This suggests several possibilities: (1) The feature alignment provided by this specific adversarial approach may not provide additional benefits beyond what strong photometric augmentations already offer. (2) The BiSeNet architecture might be more sensitive to the noisy gradients from the discriminator. (3) The adversarial loss weight λ_{adv} may require more careful tuning to balance the segmentation and adaptation objectives effectively.

However, the success of the Lovász-Softmax extension validates our final hypothesis. By directly optimizing a surrogate of the IoU metric, our model achieved a new best result for this project of **33.04%**. A closer look at Table 2 shows that the improvements were not uniform across all classes. The Lovász model significantly boosted performance on difficult, spatially small, or underrepresented classes like ‘sidewalk‘ (26.12% $\rightarrow$ 40.30%), ‘rider‘ (0.75% $\rightarrow$ 17.38%), and ‘bicycle‘ (4.97% $\rightarrow$ 8.76%) compared to the model based on the standard adversarial approach. This is consistent with the known properties of the Lovász-Softmax loss, which is more robust to class imbalance than

Table 2. Full per-class Intersection-over-Union (IoU) results on the Cityscapes validation set for all major experimental stages.

Model	**mIoU (%)**	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
BiSeNet (GTA5 → CS)	22.47	48.95	16.94	62.49	5.63	10.36	13.88	14.68	7.17	74.92	5.72	63.27	43.01	0.00	47.61	5.96	5.16	1.02	0.03	0.04
BiSeNet + Aug (Horizontal Flip)	21.90	35.71	12.28	61.40	4.06	10.12	17.44	15.66	11.73	72.24	3.88	59.28	44.23	0.04	58.06	2.58	6.44	0.79	0.00	0.12
BiSeNet + Aug (Color Jitter)	31.62	85.63	28.24	73.78	23.07	14.08	24.74	17.31	10.45	78.25	21.32	77.17	**45.21**	2.85	74.29	19.57	2.39	0.02	2.10	0.35
BiSeNet + Aug (ISO Noise)	21.72	48.76	10.71	68.22	7.44	6.78	19.43	9.50	5.25	72.96	6.06	70.95	43.67	0.29	38.91	3.20	0.37	0.00	0.23	0.02
BiSeNet + Aug (Coarse Dropout)	21.60	40.66	13.90	61.86	6.98	8.43	15.68	11.07	6.09	74.52	4.20	62.76	37.48	0.03	58.59	2.62	4.14	0.56	0.79	0.00
Aug Combinations (Color Jitter + ISO Noise + Coarse Dropout)	31.97	**83.68**	24.95	75.47	19.41	14.39	24.85	18.72	7.07	**79.68**	**25.80**	**84.56**	42.08	6.64	71.80	12.93	6.75	**1.41**	**6.53**	0.71
+ Adversarial (CE)	31.74	86.62	26.12	78.55	**26.49**	7.03	**27.66**	21.49	**13.92**	75.76	16.03	76.19	41.89	0.75	79.68	17.32	0.55	0.51	1.51	4.97
+ Adversarial (Lovász)	**33.04**	78.18	**40.30**	**79.46**	19.34	**14.99**	25.07	**24.39**	11.75	78.21	22.04	78.79	19.75	**17.38**	**75.90**	**23.20**	**7.55**	0.00	2.76	**8.76**



Input Image | Ground Truth | Base Model Prediction | Our Final Model's Prediction

Figure 3. Qualitative results on samples from the Cityscapes validation set. Our model produces more accurate segmentations, especially for complex objects and boundaries.

cross-entropy.

The results in Figure 3 would visually confirm these findings. The prediction from the base model is noisy and inaccurate. The final model produces significantly cleaner and more semantically correct segmentation maps that are much closer to the ground truth.

6. Conclusions

In this project, we successfully implemented and evaluated a series of unsupervised domain adaptation techniques for real-time semantic segmentation. We demonstrated that strong data augmentation can serve as a powerful baseline for UDA, significantly closing the synthetic-to-real performance gap. While our standard adversarial implementation did not yield further improvements over this strong baseline, the enhancement of the training objective with a hybrid Cross-Entropy and Lovász-Softmax loss pushed the performance to 33.04% mIoU. This represents a total recovery of over 10 mIoU points from the initial domain-shifted model.

Our key takeaway is that successful domain adaptation is a multi-faceted problem. It requires not only aligning feature distributions between domains but also choosing a loss function that is tightly coupled with the final evaluation metric, especially to handle the inherent class imbalance in real-world scenes. This work highlights the potential of such integrated approaches.

For future work, several avenues are promising: (1) Exploring more advanced adversarial techniques, such as those with more complex discriminator architectures or different stabilization methods. (2) Combining our approach with self-supervised learning on target domain data to further leverage unlabeled images. (3) Investigating the impact of these methods on a wider range of real-time network architectures beyond BiSeNet.

References

[1] Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. *CoRR*, abs/1705.08790, 2018. 3

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2017. 2, 3

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016. 1, 2, 3

[4] Shijie Hao, Yuan Zhou, and Yanrong Guo. A brief survey on semantic segmentation with deep learning. *arXiv preprint arXiv:2111.12763*, 2021. 1

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 3

[6] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. *CoRR*, abs/1608.02192, 2016. 1, 2, 3

[7] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. *CoRR*, abs/1802.10349, 2018. 2

[8] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4085–4095, 2020. 2

[9] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. *CoRR*, abs/1808.00897, 2018. 1, 2

[10] Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E. Gonzalez, Alberto L. Sangiovanni-Vincentelli, Sanjit A. Seshia, and Kurt Keutzer. A review of single-source deep unsupervised visual domain adaptation, 2020. 1, 2