Closer Look at the Uncertainty Estimation in Semantic Segmentation under Distributional Shift

Sebastian Cygert[§]
Multimedia Systems Department
Gdańsk University of Technology
Gdańsk, Poland
sebcyg@multimed.org

Bartłomiej Wróblewski[§]
Gdańsk University of Technology
bart.wroblew@gmail.com

Karol Woźniak
Gdańsk University of Technology

Radosław Słowiński
Gdańsk University of Technology

Andrzej Czyżewski
Multimedia Systems Department
Gdańsk University of Technology

***Abstract*—Semantic segmentation plays a very important role in many robotic tasks, i.e. autonomous driving. While recent computer vision algorithms achieve impressive performance on many benchmarks, they lack robustness - presented with an image from different distribution (e.g. weather conditions unseen during training), may produce erroneous prediction. As such it is desired that a system would be able to reliably predict its confidence measure. However, it is unknown, how well the uncertainty estimation methods for semantic segmentation methods work in the real-world scenario, under distributional shift. In this work, uncertainty estimation is evaluated under varying level of domain shift: in cross dataset setting, but also when adapting model trained on data from simulation. It was shown that an ensemble, consisting of models using different backbones and/or augmentation methods yield very competitive performance, and greatly improve model calibration under domain shift setting. Further, it was shown, that such ensemble of models can be utilized in the downstream task of domain adaptation, where it was used to improve the model accuracy in the self-training setting.**

***Index Terms*—uncertainty estimation, semantic segmentation, domain adaptation, self-training, ensemble of models**

I. Introduction

In recent years visual recognition has witnessed an impressive progress on many benchmarks. However, application of deep learning methods for agents operating in the real-world, e.g. autonomous driving has been limited. A significant challenge is that current vision models lack robustness [1]. It was shown that current CNN-based models are sensitive to novel type of noise [2], changes in context [3] or novel weather conditions [4]. Those examples show that CNNs are sensitive to distributional shift: when the test-time distribution of data differs from the training distribution. Additionally, current models seems to be biased towards texture information [5], largely ignoring shape information. This again, can be very dangerous for the real-world deployment, for example in case of sensors noise [6].

This work has been partially supported by Statutory Funds of Electronics, Telecommunications and Informatics Faculty, Gdańsk University of Technology.

[§]Equal contribution

What is more, current models tend to be overconfident in their outputs [7]. The problem is even more evident for the distributional shift [8]. For models operating in the real-world it is of great importance to be robust to such distributional changes, because it for many applications it is not possible to collect a large and diverse enough dataset that contains all possible situations that may occur during deployment (e.g. new weather or lighting conditions, different types of distortions).

A task of special importance for agents operating in the real-world is reliable uncertainty estimation, which can be beneficial in many ways. During deployment agent could warn that its prediction is not reliable (medicine), or could effectively integrate predictions from different modalities (autonomous driving) [9]. Uncertainty estimation could be also used for pseudo-labelling of unlabelled data, to further improve model accuracy in the target domain in self-training setting [10].

In this work we focus on studying uncertainty estimation for semantic segmentation, which is a very important task with large potential of applications. Further, our study focuses on distributional shift which is of great importance for real-world applications We study uncertainty calibration in different settings:

- when model trained on the simulation is tested on real-world data (large distributional shift)
- cross-dataset evaluation (mild distributional shift)

Further, we utilized a state-of-the art method for model calibration, namely ensemble of models [8], [11], to improve the model calibration. This allow to greatly improve calibration of predictive uncertainty, especially under domain shift. Finally, we show the effect of using ensemble of models on downstream task - domain adaptation, for which we utilize a popular self-training approach [10], [12], [13]. Our study is aimed at the reality-check for uncertainty estimation and domain adaptation methods. Especially studying performance for the varying domain shift for the aforementioned methods is important empirical work for real-world applications. We have focused on autonomous driving applications, due to the availability of large annotated datasets from both simulation and real-world, and potential applications. Our contributions

are as follows:

- We study how the uncertainty estimation for semantic segmentation is affected by varying level of distributional shift. Further an ensemble of models approach is evaluated in the same setting.
- We show that simple color transformations can be as effective as style-transfer data augmentation for increasing models' robustness.
- We show how ensemble of models can be utilized in the self-training approach to further improve model adaptation to the target domain.

II. Related work

Robustness. Evaluating models in out-of-distribution (OOD) setting, when the test-time dataset is from different distribution than training data is important for real-world applications [8], [14], [15]. This is because machine learning models might provide wrong predictions when presented with for example noisy data, different lighting or weather conditions [6]. To improve models' robustness in visual recognition several methods based on data augmentation were proposed, of which style-transfer data augmentation is very popular [5], [16]. In our work, style-transfer data augmentation was utilized, but we also noticed that simply applying color-jittering during training can be beneficial for the cross-dataset evaluation, which confirms a recent finding that very simple naturalistic augmentation can be very effective [17].

Uncertainty estimation. One of the problems with modern neural networks is that they are poorly calibrated and tend to be overconfident in the predictions [7]. Different techniques exists for improving estimates of predictive uncertainty. A classical approach is called temperature scaling, where the model confidences are scaled using post-hoc procedure on the held-out validation set [18]. A popular approximate Bayesian approach is a dropout-based model, where the predictive uncertainty is computed based on the multiple outputs of the model on given image (with dropout enabled) [19]. Another sampling based approach uses agreement between ensemble of models as a measure of model uncertainty [20]. Interestingly using ensembles was shown to yield the best results on uncertainty estimation under the distributional shift [8], [11]. For the ensembles the common setup is to use the neural networks trained using different random weights initialization, to induce diversity between models [21]. This is because it was shown that networks pretrained on the same dataset stay in the same basin in the loss landscape, and thus reduce variation in the models [22]. However, we found that semantic segmentation models trained on evaluated datasets using random initialization perform rather poor. As such, we show that it is possible to create efficient model ensemble using models with different backbones and data augmentations.

Domain adaptation. While it is a standard to evaluate machine learning models on i.i.d., for the real-world deployment the data may come from different distribution than training data. As such many methods for domain adaptation were proposed which uses unlabelled data from the target domain to improve the accuracy of the model. Popular approaches includes matching image statistics between domains [23], learning shape-based representation [12], self-learning [12], self-supervision [24] or using data from simulation [25], [26]. Using simulated data is in particular interesting, since it simulation allows to generate numerous and diverse training examples. At the same time, difference in data distribution between source and target domain is very challenging for the real-world problems and sometimes using the labelled data can actually hurt performance [17]. As such it is important to evaluate models' performance under varying level of distributional shift.

In our work we make use of self-learning method which works in two stages. First, given a trained model, confident predictions are gathered for the target domain, which are also called pseudo-labels. In the next stage, the pseudo-labels are used to finetune the model, which allow for domain adaptation. The potential problem with self-learning is that gathered pseudo-labels might contain erroneous predictions. As such, we propose to use an ensemble approach to gather the pseudo-labels, as ensembles are known to have both good accuracy and uncertainty estimation, which are crucial for the efficient pseudo-labeling stage.

Similar to our work, in [27] it was shown that ensemble of models is efficient for improving uncertainty estimation in medical image segmentation. We additionally show the effect of ensembles under distributional shift and its utility for downstream task of domain adaptation. Ensemble predictions on unlabelled dataset were also used as soft targets for direct training supervision in knowledge distillation framework [28], [29]. Here we use an alternative approach where the least confident predictions are discarded during training.

Similar to us, [12] use style-transfer data augmentation to train a base model, which is further adapted to the target domain using self-training. We show that simpler data augmentations can also be very efficient, and that ensemble of models makes the finetuning stage more efficient which makes our work complementary.

III. Methodology

A. Semantic Segmentation

Semantic segmentation can be viewed as a pixel-wise classification problem where the goal is to assign to each pixel a predicted category $c \in \{1, ..., C\}$. As it is now common in visual recognition area, semantic segmentation models are mostly based on Convolutional Neural Networks (CNNs), for example FCN [30]. As it is a classification problem a standard cross-entropy loss can be used to optimize the model weights over the training images:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}(y_i = c)log(p(\hat{y_i} = c)) \qquad (1)$$

where $(y_i = c) \in \{0, 1\}$ indicates whether class c is correct class for pixel i and $\hat{y_c}$ is a predicted probability for class c at pixel i, and N is the number of pixels. For each pixel

model returns a logit vector $z_i \in R_c$. Further a $softmax$ function is applied $p_i = softmax(z_i)$, which returns a list of predicted class probabilities for given pixel. Class with the highest probability is used as the predicted class with associated probability score.

Over years many different architectures were developed and in our work we have used DeepLabV3+ [31], which is commonly used in the community. Furthermore, one can use different backbones (e.g. large ResNet-101 or a lightweight MobileNet) accordingly to the requirements.

For evaluation two metrics are used. Pixel accuracy simply measures how many % of pixels are correctly predicted. Another popular metric is mean IoU (intersection over union). IoU metric is computed for each class and then the mean value (mIoU) is reported.

B. Uncertainty Estimation

An output of semantic segmentation network is a predicted class c for each pixel, associated with confidence value p. Ideally such classifier would be well calibrated - correct predictions should be associated with high confidence and poor predictions contrary. One of the ways to measure model calibration is to compute an Expected Calibration Error (ECE) [7]. To compute ECE score pixel-wise predictions are partitioned into M equally-sized bins based on the confidence value and the ECE score is computed as the difference between the average confidence and the average accuracy in each bin, weighted by the number of predictions in each bin:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{2}$$

where B_m is the set of indices that falls into the *mth* bin. Intuitively, when a well-calibrated segmentation network outputs a 90% confidence value for some set of pixels, it should be correct in 90% of the cases. The lower the value, the better calibration is obtained (0 means perfect calibration).

C. Ensemble of models

To improve model calibration, we utilize model ensemble method which was shown to provide the best results among other methods, especially under distributional shift [8]. In ensemble approach it is common to train models using randomly initialized networks to induce diversity between models [21]. However, we found that semantic segmentation models trained on GTA or Cityscapes dataset, without pretraining performs rather poor. As a result we use ImageNet pretraining, however in order to achieve diversity between models different backbones are used and / or augmentation methods. It was shown in the literature that using 5 models can already provide very good results [8], and because of the computational budget, we use 5 models in our experiments. Namely given M independently trained models, a final semantic segmentation p_E for the image x can be computed as the average of all models predictions:

Fig. 1: Different augmentation strategies applied to sample images from the GTA dataset. First column - color transformations, second column - style transfer.

$$p_E(x) = \frac{1}{M} \sum_{i=1}^{M} p_m(x) \tag{3}$$

where p_m is the prediction of the *mth* model in the ensemble.

D. Data augmentation

In order to improve models' adaptation to the distributional shift a style-transfer data augmentation was utilized, which was shown to improve models' robustness [5], [16]. As the source of style images, Kaggle's *Painter By Numbers*[1] dataset was used, similar as in [5] and during training a stylized image is sampled with probability $p = 0.5$, otherwise original image is used. In order to generate stylized dataset we have used method presented in [32].

We also hypothesized that using simple color transformations could be also beneficial, in the domain adaptation setting as it would make the model more invariant to the texture information. As such, as an alternative to the style-transfer, a following color jittering transformations from the Tensorflow API[2] were also used during training: random changes in brightness, contrast, saturation and hue of the images. Details are described in the implementation details section.

Using different augmentation strategies could be also beneficial in the ensemble, as the models trained with different augmentation might learn different representations. Fig. 1 shows examples of augmented images.

IV. Experiments

A. Datasets

For our experiments, popular semantic segmentation datasets were used. All of them contain dense pixel-level semantic annotations for the same 20 classes (including ignore class - usually background).

GTA [33] is a dataset for which data were collected in the simulated environment, i.e. a modern computer game. It

[1]https://www.kaggle.com/c/painter-by-numbers/

[2]https://www.tensorflow.org/api_docs/python/tf/image

consists of 22466 training and 2500 validation images and is commonly used to evaluate simulation to real transfer. **Cityscapes** [34] is a popular autonomous driving dataset for which data was collected in 27 cities in Germany, consisting of 2975 training images and 500 validation images. Although Cityscapes is a diverse dataset, a potential limitation is the fact that the data was collected mostly during daytime in good weather conditions. **Berkeley Deep Drive (BDD)** dataset [35] provides data collected in diverse weather conditions (e.g. rain, snow), scene types(city, highway, countryside) and also images recorded during nigh-time. Pixel-level annotations are provided for 10000 training and 1000 validation images. Finally, also the Foggy Cityscapes [36] dataset is used, which contains an original Cityscapes images refined with synthetic fog effect. Tests are performed on this dataset to provide additional measure of robustness of trained models to distortions unseen during training.

In our experiments we focus on domain adaptation from simulation to real data (GTA $\rightarrow$ Cityscapes) and cross-datasets evaluation (Cityscapes $\rightarrow$ BDD).

B. Implementation details

For all experiments DeepLabv3+ [31] network was used with different backbones (ResNet-101, Xception41, Xception65) pretrained on ImageNet. Specifically, all models have been trained on 2 GPUs for 100.000 steps with batch size of 16. As in the original paper, a polynomial decay learning rate is used with initial learning rate = 0.01 and parameter *power* set to 0.9.

Data augmentations are consistent with official implementation[3], specifically random scaling (in range 0.5 to 2.0) and left-right flipping was applied during training procedure. All images has been rescaled to the size 512 x 1024 pixels. The color jittering data augmentation was applied using TensorflowAPI with following transformations: random brightness (adjust factor in range[0, 0.25)), random contrast (contrast factor in range [0.5, 1.5)), random saturation (saturation factor in range [1.0, 3.0)) and random hue(hue offset in range [0, 0.25)). Chosen hyperparameters were experimentally validated to provide visually diverse images.

All models are evaluated on the validation sets (as test sets ground-truth data is not publicly available). During fine-tuning stage models are trained for 25.000 steps as we noticed that the training loss has converged around 20.000 steps for all the models. When reporting the results CJ model stands for a model trained using color jittering transformations, while SIN stands is a model trained using style-transfer, as in [5].

C. Baseline models

First, DeepLabV3+ model with ResNet backbone is trained on both GTA and Cityscapes datasets and further evaluated (Table I. Several observations can be made. First, there is a very big drop in accuracy when the models are evaluated under domain shift, and the gap is larger for sim to real

[3]https://github.com/tensorflow/models/tree/master/research/deeplab

TABLE I: Performance of **DeepLabv3 using ResNet-101** backbone under different evaluation settings.

Model name	mIoU	pix. acc	ECE	mIoU	pix. acc	ECE
	GTA $\rightarrow$ GTA			**GTA $\rightarrow$ Cityscapes**		
Baseline	80,8	96,6	0.16	25,4	60,1	23.36
CJ	80,6	96,4	0.21	40,4	83,7	6.5
SIN	77,2	95,9	0.21	40	83,9	5.06
	Cityscapes $\rightarrow$ Cityscapes			**Cityscapes $\rightarrow$ BDD**		
Baseline	74,1	95,5	1.49	42,8	83,9	9.55
CJ	74,4	95,4	1.36	49,1	89,4	5.07
SIN	71,4	95	1.18	49,3	89,8	4.56

adaptation (GTA to Cityscapes) when comparing with cross-dataset evaluation (Cityscapes to BDD). Further it can be observed that texture based data augmentation (Color jitter and style-transfer), only slightly affects the performance on the source domain, but they show a really impressive performance in the domain adaptation setting. For GTA to Cityscapes the mIoU has increased from **25.4 to 40.4** and similarly for Cityscapes to BDD mIoU has increased from **42.8 to 49.1**. Nevertheless, the domain gap is still quite large, model trained on Cityscapes dataset achieves mIoU of 74.1, compared to 40.4 achieved by model trained on GTA dataset.

Surprisingly, applying simple color transformations works as well as using an advanced technique of style-transfer, which is consistent with very recent finding [37]. Looking at the model calibration, one can notice that all of the models are almost perfectly calibrated when evaluated on the source domain, however when evaluated under domain shift the ECE metric has greatly increased, e.g. for model trained on Cityscapes dataset the metrics has increased from **1.49 to 9.55** when evaluated on the BDD dataset instead of Cityscapes. Consistent, with recent finding it was shown that using texture based data augmentation improves model calibration under domain shift [16], with SIN model obtaining slightly better results than using color transformations.

D. Model calibration

To improve model calibration, an ensemble of models method was used, which utilizes three different backbones (ResNet-101, Xception41, Xception65) and two different augmentation methods (color jitter and style-transfer). We also experimented with PNAS architecture, which is known to achieve great accuracy, however the performance was not satisfactory, as no pretrained model is currently available for that model. Table II shows performance in cross-dataset setting for the Xception models. Comparing to Table I one can see that Xception models perform slightly better than models using ResNet-101 as the backbone.

Table III shows ensemble performance when 3 and 5 models are used in the computation. For $M = 3$, we have used models trained using color jittering transformations (ResNet-101 and Xception backbones) and for $M = 5$ two additional models trained using style transfer augmentation were used (ResNet-101 and Xception41 backbones). While results with no domain shift are comparable, under the domain shift the

TABLE II: Xception models performance under cross-dataset setting.

Name	mIOU	pix. acc	ECE
GTA → Cityscapes			
Xception41 (CJ)	41.8	82.7	7.3
Xception41 (SIN)	43.7	86.3	4.05
Xception65 (CJ)	41.3	81.97	7.47
Cityscapes → BDD			
Xception41 (CJ)	52.6	90.3	4.5
Xception41 (SIN)	51.1	90.9	3.74
Xception65 (CJ)	52.4	90.4	5.09

TABLE III: Ensemble of models performance.

Model name	mIoU	pix. acc	ECE	mIoU	pix. acc	ECE
	GTA → GTA			**GTA → Cityscapes**		
Baseline (M=3)	x	96.7	1.49	x	69.81	4.2
Augmented (M=3)	81.9	96.8	0.81	43.2	84.7	2.45
Augmented (M=5)	81.4	96.7	1.02	44.5	86.27	1.09
	Cityscapes → Cityscapes			**Cityscapes → BDD**		
Baseline (M=3)	x	x	x	x	x	x
Augmented (M=3)	77.2	96.0	0.36	55.7	91.3	1.99
Augmented (M=5)	77.0	96.0	0.29	56.2	91.7	1.09

obtained results are better when using 5 models. The mIoU has increased from 43.2 to 44.5 and from 55.7 to 56.2 on Cityscapes and BDD datasets respectively. Similarly the ECE is significantly reduced on both datasets. Also it is very important to notice that the ensemble performance is better than its strongest member, i.e. for the Cityscapes to BDD transfer the strongest single model obtains mIoU of 52.6 (Xception41 - CJ) while the ensemble accuracy has greatly improved to 56.2. Similarly, the ECE has greatly improved for the ensemble under domain shift: for GTA to Cityscapes transfer the ensemble ECE is 1.09, while the best results from single model is 4.05 (Xception41 - SIN). Additionally fig. (2) shows calibration plot. Comparing model calibration of our highest-capacity model (Xception65) with the calibration of the ensemble. Overall, it was confirmed that our ensemble improves both accuracy and uncertainty calibration, especially under domain shift.

One of the potential usages of well-calibrated uncertainty estimation is a self-training. For that purpose, we first estimate the precision / recall points for different confidence thresholds t (fig. 3). Namely, such curve is an approximation of how many % of pixels can be automatically annotated with what precision. Overall, it can be noticed that much higher recall values are obtained for the ensemble. For example, at precision of 95% for the Xception65 model the recall is around 56.5%, while for the the ensemble it has increased to 71.2%. This shows that ensembles are a very powerful technique. A complimentary work to ours, shows that ensembles can be used to efficiently label new dataset [38]. Ensemble was used to coarsely annotate new dataset with high accuracy, and then human annotators were employed to refine initial predictions.

E. Domain adaptation

As it was shown, ensemble of models allowed to improve model precision in the domain adaptation setting, and also

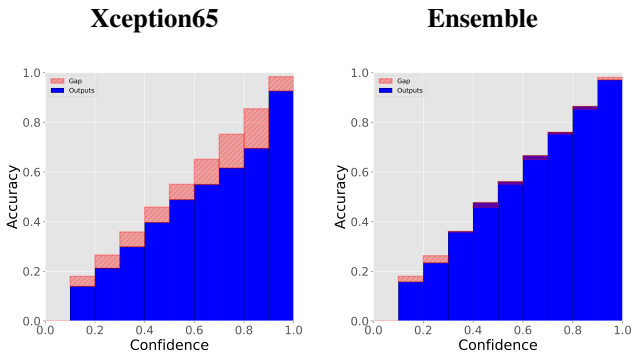

Fig. 2: Calibration plots for Xception65 model and model ensemble (M=5) evaluated on the GTA → Cityscapes adaptation. Note great calibration for the ensemble of models.

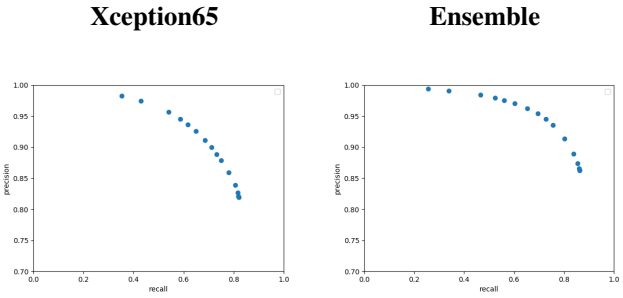

Fig. 3: Precision / recall points evaluated at different confidence threshold in range [0.1, 0.995] on GTA → Cityscapes transfer.

greatly improved uncertainty estimation which can be efficiently utilized in the self-training setting. Firstly, semantic segmentation model is used to obtains pseudo-labels on the target datasets, using some threshold t. In the literature, threshold of value 0.9 is commonly used [12], and the same value is used in our experiments. For the ensemble variant, such threshold allows to annotate 70.1% of the pixels with 92.6% accuracy. Fig. 4 shows obtained pseudo-labels. In general, it can be noticed that the ensemble's labels are less noisy, and the object boundaries are more tightly aligned around the object of interest.

After the pseudo-labels were obtained for target datasets, they were used for model finetuning. In this section results for different models are presented:

1) ResNet-101 using standard data augmentation.
2) ResNet-101 trained using additional color jittering data augmentation.
3) Previous model finetuned on target datasets using pseudo-labels obtained by that model (*CJ + fine* in the tables)
4) ResNet-101 finetuned on target datasets using pseudo-labels obtained by model ensemble (*CJ + ens* in the tables)

Table IV shows final results, including per-class evaluation. Firstly, consistent with other works, the self-training approach

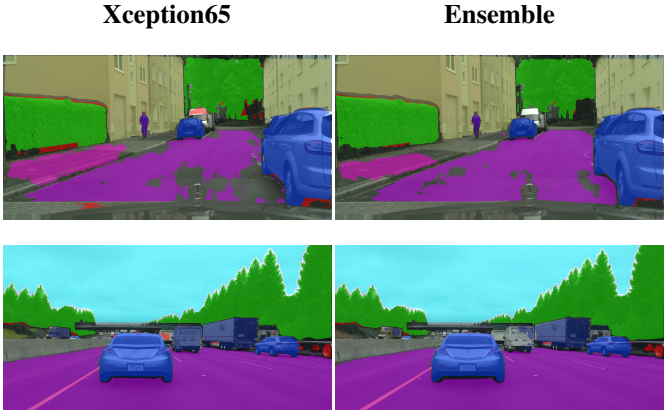

Fig. 4: Examples of pseudo-labels obtained obtained on GTA → Cityscapes transfer (first row), and on Cityscapes → BDD transfer (second row).

improves the model accuracy (from 40.4 to 41.2 and from 49.1 to 51.4 on Cityscapes and BDD datasets respectively). When the pseudo-labels are collected using an ensemble approach, a model accuracy has further greatly increased.

Important problem with ensemble of model is that one has to train and evaluate multiple models which is very costly. However, recently introduced BatchEnsemble method significantly reduced the computational and memory costs [39]. Similarly, it was shown that training one neural network with multi-input multi-output (MIMO) configuration can be efficient strategy to improve models' robustness. Yet, applying those ideas to high-level task of semantic segmentation is important future work.

V. Conclusions

In this work calibration of model predictive uncertainty under different, realistic for real-world application setting was studied. It was shown that ensemble of models allowed to significantly improved the uncertainty estimation, especially under domain shift. Notably, the performance gains are consistent when the domain gap is large (simulation to real transfer). Our ensemble consists of models using different backbones and/or data augmentations. Interestingly, it was also shown, that simple color transformations can achieve similar performance to much more sophisticated style-transfer augmentation and both types of data augmentation are crucial in the domain adaptation setting.

Further, the ensemble of models was utilized for domain adaptation using self-training method. The improved uncertainty calibration and model accuracy allowed to significantly improve the finetuning stage, the mIoU has increased from 41.2 to 44.0, and from 51.4 to 54.2 on Cityscapes and BDD datasets respectively.

References

[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[2] S. Dodge and L. Karam, "A study and comparison of human and deep learning recognition performance under visual distortions," in *2017 26th international conference on computer communication and networks (ICCCN)*, pp. 1–7, IEEE, 2017.

[3] A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, and B. Katz, "Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models," in *Advances in Neural Information Processing Systems*, pp. 9453–9463, 2019.

[4] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel, "Benchmarking robustness in object detection: Autonomous driving when winter is coming," *CoRR*, vol. abs/1907.07484, 2019.

[5] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019.

[6] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *arXiv preprint arXiv:1706.04599*, 2017.

[8] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, pp. 13991–14002, 2019.

[9] D. Feng, Y. Cao, L. Rosenbaum, F. Timm, and K. Dietmayer, "Leveraging uncertainties for deep multi-modal object detection in autonomous driving," *arXiv preprint arXiv:2002.00216*, 2020.

[10] Q. Xie, M. Luong, E. H. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 10684–10695, IEEE, 2020.

[11] F. K. Gustafsson, M. Danelljan, and T. B. Schön, "Evaluating scalable bayesian deep learning methods for robust computer vision," *CoRR*, vol. abs/1906.01620, 2019.

[12] M. Kim and H. Byun, "Learning texture invariant representation for domain adaptation of semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12975–12984, 2020.

[13] Y. Zou, Z. Yu, B. V. K. V. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III* (V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds.), vol. 11207 of *Lecture Notes in Computer Science*, pp. 297–313, Springer, 2018.

[14] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," 2016.

[15] C. Kamann and C. Rother, "Benchmarking the robustness of semantic segmentation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8828–8838, 2020.

[16] S. Cygert and A. Czyżewski, "Toward robust pedestrian detection with data augmentation," *IEEE Access*, vol. 8, pp. 136674–136683, 2020.

[17] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," *CoRR*, vol. abs/1804.09170, 2018.

[18] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[19] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, vol. 48, pp. 1050–1059.

[20] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA* (I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, eds.), pp. 6402–6413, 2017.

TABLE IV: Domain adaption results for our models with per-class evaluation.

Name	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	**mIoU**
Gta to Cityscapes																				
Baseline	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x x	x	25.4
CJ	80.3	28.9	80.9	30.9	22.5	25.8	37.0	17.5	83.8	31.0	76.6	**58.4**	**19.6**	83.0	28.7	24.7	0.0	**27.4**	**11.0**	40.4
CJ + fine	86.1	36.4	83.1	24.9	28.7	27.8	**39.6**	19.4	85.7	38.4	79.5	56.9	13.0	86.5	31.0	23.6	0.	22.6	0.	41.2
CJ + ens	**88.6**	**43.2**	**85.0**	**36.3**	**33.8**	**30.7**	37.4	**21.9**	**86.8**	**44.9**	**83.9**	57.5	14.5	**87.3**	**37.2**	**32.2**	0.0	15.0	0.0	**44.0**
Cityscapes to BDD																				
Baseline	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	42.8
CJ	91.8	54.5	79.9	**19.8**	27.1	**41.9**	**43.3**	43.8	82.5	39.1	91.4	58.2	29.7	85.2	27.7	25.5	0.	**49.1**	42.6	**49.1**
CJ + fine	93.2	60.4	**81.4**	18.7	36.6	37.4	40.5	44.2	83.0	42.0	91.7	62.2	43.7	85.1	36.4	23.6	0.	47.6	48.7	51.4
CJ + ens	**94.4**	**62.5**	81.0	17.5	**37.7**	38.6	38.6	**45.5**	**85.0**	**43.2**	**92.2**	**63.2**	**46.8**	**87.1**	**42.6**	**54.7**	0.0	44.9	**53.4**	**54.2**

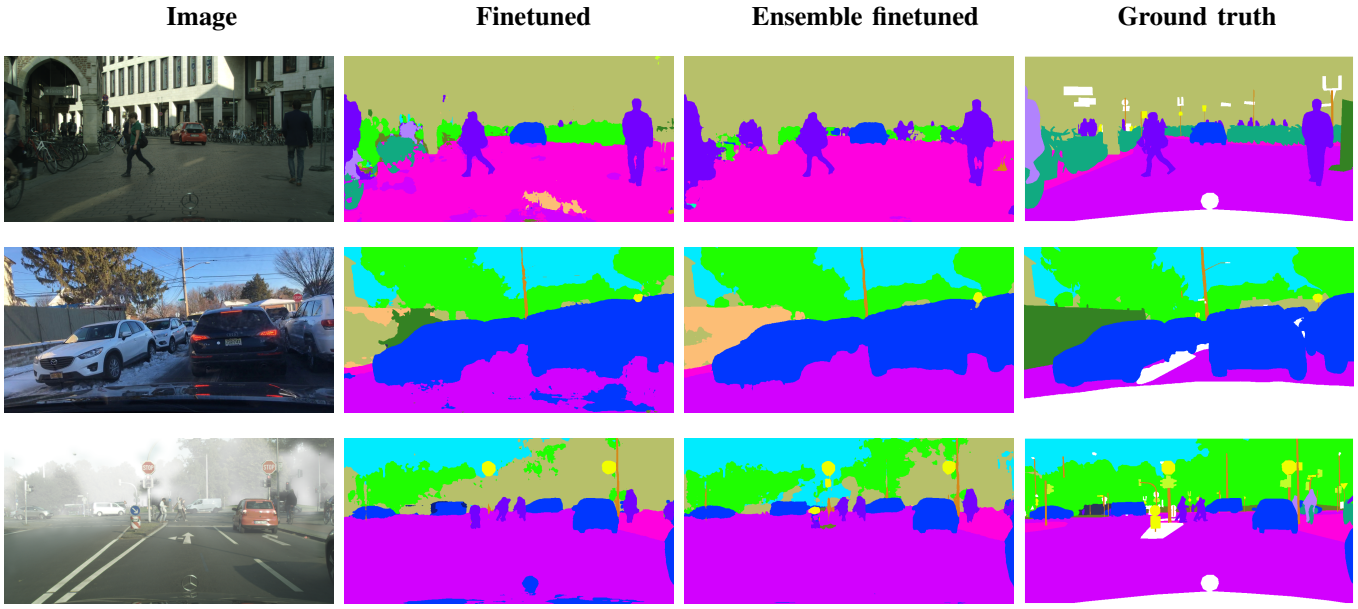

Fig. 5: Qualitative results of trained models on GTA → Cityscapes transfer (first row) and Cityscapes → BDD transfer (second row). Last row shows results on Foggy Cityscapes.

TABLE V: Evaluation on Foggy Cityscapes dataset at the highest intensity of fog.

Name	robust mIOU	robust IoU
CJ	37.4	80.8
CJ + fine	33.3	79.5
CJ + ens finetune	34.1	80.8
CJ + ens finetune	36	81.6

[21] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[22] B. Neyshabur, H. Sedghi, and C. Zhang, "What is being transferred in transfer learning?," 2020.

[23] A. Abramov, C. Bayer, and C. Heller, "Keep it simple: Image statistics matching for domain adaptation," *CoRR*, vol. abs/2005.12551, 2020.

[24] Y. Sun, E. Tzeng, T. Darrell, and A. A. Efros, "Unsupervised domain adaptation through self-supervision," *arXiv preprint arXiv:1909.11825*, 2019.

[25] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis, "Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks," 2019.

[26] B. Zhou, N. Kalra, and P. Krähenbühl, "Domain adaptation through task distillation," *arXiv preprint arXiv:2008.11911*, 2020.

[27] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 3868–3878, 2020.

[28] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006* (T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, eds.), pp. 535–541, ACM, 2006.

[29] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[30] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.

[31] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[32] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," *CoRR*, vol. abs/1703.06868, 2017.

[33] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *European conference on computer vision*, pp. 102–118, Springer, 2016.

[34] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[35] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 2633–2642, IEEE, 2020.

[36] C. Sakaridis, D. Dai, and L. V. Gool, "Semantic foggy scene understanding with synthetic data," vol. 126, pp. 973–992, 2018.

[37] K. L. Hermann, T. Chen, and S. Kornblith, "The origins and prevalence of texture bias in convolutional neural networks," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), 2020.

[38] G. Zhou, S. Dulloor, D. G. Andersen, and M. Kaminsky, "EDF: ensemble, distill, and fuse for easy video labeling," *CoRR*, vol. abs/1812.03626, 2018.

[39] Y. Wen, D. Tran, and J. Ba, "Batchensemble: an alternative approach to efficient ensemble and lifelong learning," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020.