Learning Semantic Segmentation from Multiple Datasets with Label Shifts

Dongwan Kim[1], Yi-Hsuan Tsai[3], Yumin Suh[4], Masoud Faraki[4], Sparsh Garg[4], Manmohan Chandraker[4,5], and Bohyung Han[1,2]

[1]ECE & [2]IPAI, Seoul National University
[3]Phiar Technologies [4]NEC Labs America [5]UC San Diego

Appendix A More Examples of Multi-Label Predictions

In Figures A-1, A-2, and A-3, we present additional qualitative results from BDD, IDD, and Mapillary datasets, respectively. These figures extend Figure 4 from the main paper, and aim to show that similar behavior can be observed in BDD and IDD.

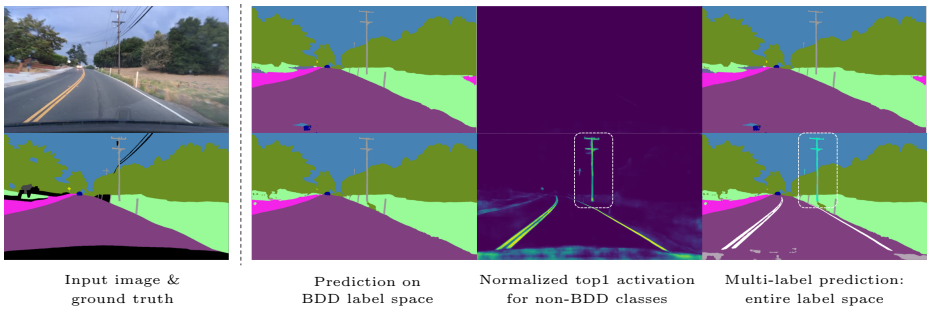

Fig. A-1. Multi-class predictions of the **BDD** dataset. The first row corresponds to an HRNet-W48 model trained with the CE loss, while the second row corresponds to our C-R BCE model. While both models make strong predictions on the BDD label space (column 2), only the C-R BCE model has high (normalized) activations for non-BDD classes in regions with label conflict (column 3). For example, Lane markings (color: white), while not labeled in BDD, are predicted via multi-class prediction, as well as the "utility pole" class (color: green) from the Mapillary dataset.

Interestingly, on Mapillary we notice that even the CE model can output strong activations for "lane marking", which was not the case when evaluating on datasets that do not label this class (Cityscapes: Figure 4, BDD: Figure A-1, and IDD: Figure A-2). Based on this observation, we argue that the CE model insidiously learns a connection between the domain and the label space. This could be seen as a form of overfitting. On the contrary, the C-R BCE model evidently does a much better job at generalizing the label space across domains.

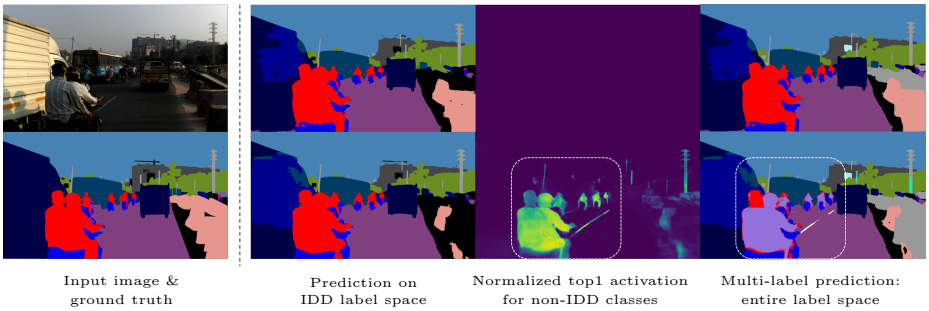

Fig. A-2. Multi-class predictions of the **IDD** dataset. Most "riders" are predicted as "motorcyclists" (color: light purple) in the multi-class prediction.

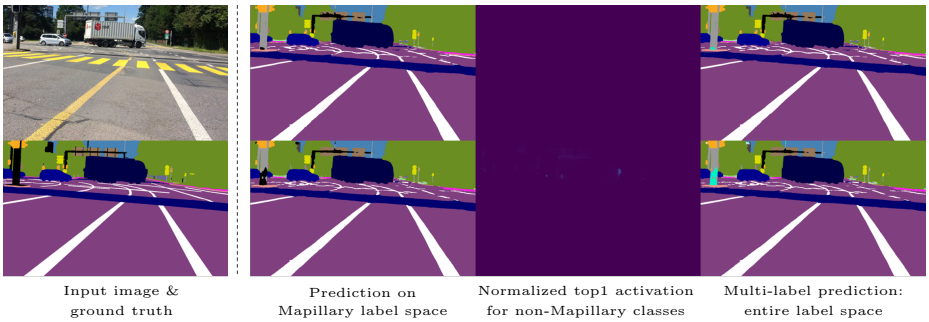

Fig. A-3. Multi-class predictions of the **Mapillary** dataset. There is almost no noticeable difference between the predictions of the CE and C-R BCE model. This figure shows that the CE model is capable of predicting fine-grained labels. However, it will not do so for domains that did not have these fine-grained labels in the first place.

Appendix B Predicted Multi-Labels

Table A-1 presents the list of multi-labels that are predicted by our model and used by the Class-Relational BCE loss. The "Primary Label" column lists the names of the original categories, while the "Secondary Label" column lists the corresponding multi-label, *i.e.*a pixel labeled as "bike lane" will also receive supervision for the "road" class. As mentioned in the paper, these relationships are not symmetric, *i.e.*a pixel of "road" will not necessarily lead to supervision of "bike lane".

The list of primary and secondary labels in Table A-1 supports our hypothesis from the main paper. As expected, the "bicyclist" and "motorcyclist" classes are both mapped to "rider" as well, which means that any supervision on those classes will also help learn stronger representations for the "rider" class. In the CE setting, the model would only receive negative supervision (label of 0) for the "rider" class from Mapillary data, and in the Null BCE setting, the model would receive no supervision for the "rider" class from Mapillary data. However, in our

Table A-1. Predicted multiple labels for Class-Relational BCE

Primary Label	Secondary Label
bike lane	road
catch basin	road
crosswalk - plain	road
lane marking - general	road
parking	road
pothole	road
service lane	road
junction box	obs-str-bar-fallback
mailbox	obs-str-bar-fallback
phone booth	obs-str-bar-fallback
traffic sign (back)	obs-str-bar-fallback
trash can	obs-str-bar-fallback
curb cut	sidewalk
pedestrian area	sidewalk
bicyclist	rider
motorcyclist	rider
ground animal	person
other rider	person
banner	billboard
phone booth	building
caravan	car
on rails	train
trailer	truck

C-R BCE setting, Mapillary data of the "bicyclist" and "motorcyclist" classes would also provide positive supervision for the "rider" class, so the model can learn a more robust representations for "rider".

Appendix C Importance of our BCE Loss

In the main paper, we outline the issues encountered when using the cross-entropy (CE) loss function, and propose to use the binary cross-entropy (BCE) loss as a replacement to alleviate these issues. Here we further discuss the importance of the BCE loss in our method and how tightly this new loss function is coupled with the second-stage training using class-relational BCE loss.

In Section 3.3 of the main paper, the class relationship is more meaningful when the underlying model is trained with the BCE loss. The reason is that, BCE loss operates on a per-class basis, and thus the output scores can be high for multiple classes. This property is useful when using class relationships as some categories may correlate to each other during model training. On the contrary, using the CE loss based on Softmax only outputs the probability of a certain class, describing how probable that class is the ground-truth, *relative to all other classes*, and thus it cannot exploit the class relationships as the BCE loss does.

Furthermore, the BCE loss enables us to train with multiple ground-truth classes on a single pixel. This property is especially useful when training with multiple datasets that have different label spaces, since classes may not be fully disjoint to each other, *i.e.* they may actually refer to the same class or one class may be a subset of another. Examples of such cases are described under the "Discussions" paragraph in Section 3.3.

Appendix D Results on Seen Datasets

Table A-2. mIoU comparisons for the seen datasets in the C+I+B+M setting with HRNet-W18 and HRNet-W48.

Method	Arch.	Cityscapes	IDD	BDD	Mapillary	Mean
Multi-dataset (CE)	HRNet W18	71.8	63.1	60.4	37.2	58.1
UniSeg (Null BCE)		71.7	63.2	60.2	36.1	57.8
UniSeg (C-R BCE)		71.5	63.0	60.0	35.4	57.5
Multi-dataset (CE)	HRNet W48	80.4	70.4	67.1	46.4	66.1
UniSeg (Null BCE)		80.9	71.1	68.0	45.5	66.4
UniSeg (C-R BCE)		81.0	70.9	67.4	43.0	65.6

We present the results of the seen datasets for the "C-I-B-M (All)" setting in Table A-2. As seen in this table, our Null BCE and C-R BCE models maintain competitive performance to the CE baseline even when evaluating on the seen datasets of their original label space in individual datasets, but not on the unified label space from all the datasets. Note that, for datasets like Mapillary that contain fine-grained categories, the performance drops more as the CE model pays more attention to fine-grained information. However, the benefit of our model is orthogonal to training a good fine-grained model, since our goal is to train multiple datasets together and generalize to unseen datasets in the unified label space.

Appendix E Qualitative Results on Unseen Datasets

In Figure A-4, A-5, A-6, we show qualitative results for semantic segmentation on the unseen datasets, *i.e.*, WildDash, KITTI, and CamVid. Note that, although these input images are not from the training datasets, our model is still able to provide accurate segmentation predictions compared to the ground truths. All predictions are made on the label space of the input image.

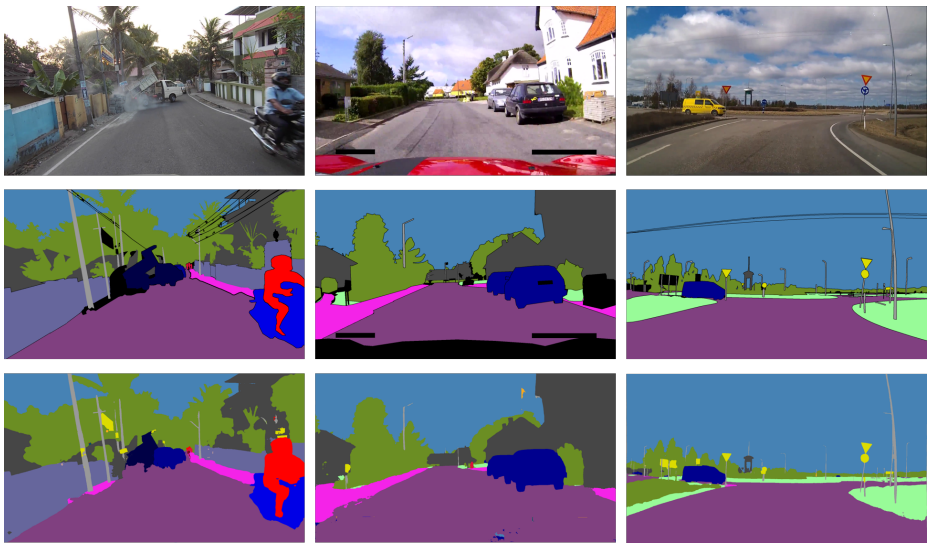

Fig. A-4. Additional WildDash visualizations. From top to bottom: Original image, Ground Truth, UniSeg model prediction.

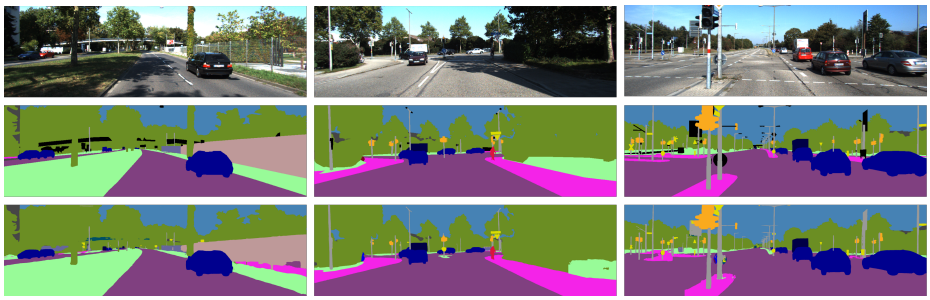

Fig. A-5. Additional KITTI visualizations. From top to bottom: Original image, Ground Truth, UniSeg model prediction.

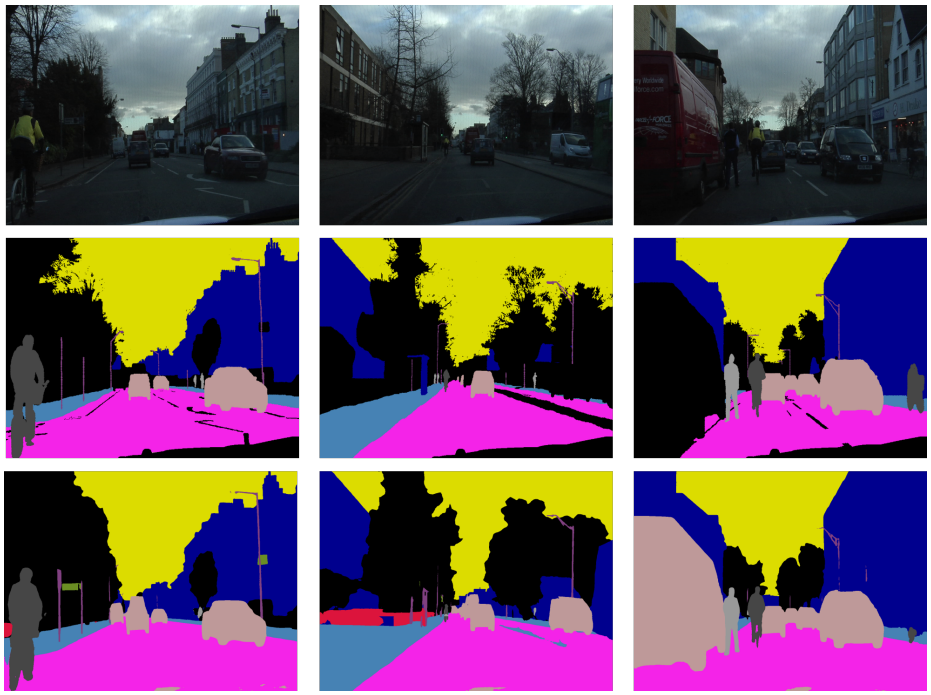

Fig. A-6. Additional CamVid visualizations. From top to bottom: Original image, Ground Truth, UniSeg model prediction.

Appendix F Derivation of Gradients

We derive the gradients for the cross-entropy loss (Eq. (2) in the main paper) and the binary cross entropy loss (Eq. (4) in the main paper). First, let us simplify our formulation and refer to just a single output location, (h, w), of the segmentation map from the i^{th} dataset.

$$\mathcal{L}_{seg}^{ce} = -\sum_{k=1}^{K_u} Y^{(k)} \log(P^{(k)}). \tag{a-1}$$

Given an ground truth y, the one-hot label $Y^{(k)} = 1$ when $k = y$ and $Y^{(k)} = 0$ when $k \neq y$. Thus, this equation can be simplified as:

$$\mathcal{L}_{seg}^{ce} = -\log(P^{(y)}). \tag{a-2}$$

With the cross-entropy loss, we use a softmax function over the channel dimension on the output, O. Hence, substituting $P^{(y)}$ with the softmax over O obtains:

$$\mathcal{L}_{seg}^{ce} = -\log\left(\frac{e^{O^{(y)}}}{\sum_k e^{O^{(k)}}}\right) = \log\left(\sum_k e^{O^{(k)}}\right) - O^{(y)}. \tag{a-3}$$

Now, taking the gradient of $\mathcal{L}_{seg}^{ce}$ with respect to the output of an arbitrary parameter θ:

$$\frac{\partial \mathcal{L}_{seg}^{ce}}{\partial \theta} = \sum_k \frac{\partial \mathcal{L}_{seg}^{ce}}{\partial O^{(k)}} \frac{\partial O^{(k)}}{\partial \theta}. \tag{a-4}$$

In the paper, we leave $\frac{\partial O^{(k)}}{\partial \theta}$ as is, since it is irrelevant for this derivation. Instead, we focus on the $\frac{\partial \mathcal{L}_{seg}^{ce}}{\partial O^{(k)}}$ term.

$$\begin{aligned}
\frac{\partial \mathcal{L}_{seg}^{ce}}{\partial O^{(k)}} &= \frac{\partial}{\partial O^{(k)}} \log\left(\sum_{k'} e^{O^{(k')}}\right) - \frac{\partial}{\partial O^{(k)}} O^{(y)} \\
&= \frac{O^{(k)}}{\sum_{k'} e^{O^{(k')}}} - \frac{\partial}{\partial O^{(k)}} O^{(y)} \\
&= \frac{O^{(k)}}{\sum_{k'} e^{O^{(k')}}} - Y^{(k)} \\
&= P^{(k)} - Y^{(k)},
\end{aligned} \tag{a-5}$$

As seen in Eq. (a-5), when k points to a conflicting class in which two datasets provide different labels, we end up with the gradient conflict depicted in Eq. (2) of the main paper. Using a similar process with the aforementioned simplifications for the binary cross-entropy loss:

$$\mathcal{L}_{seg}^{bce} = -\sum_k^{K_i} Y^{(k)} \log(Q^{(k)}) + (1 - Y^{(k)}) \log(1 - Q^{(k)})), \tag{a-6}$$

where Q denotes the sigmoid-activated output, O. Since

$$\frac{\partial Q^{(k)}}{\partial O^{(k)}} = Q^{(k)}(1 - Q^{(k)}), \tag{a-7}$$

we calculate the gradient of the BCE loss with respect to an output, $O^{(c)}$:

$$\begin{aligned} \frac{\partial \mathcal{L}_{seg}^{ce}}{\partial O^{(c)}} &= -\frac{Y^{(c)}}{Q^{(c)}} Q^{(c)}(1 - Q^{(c)}) + \frac{(1 - Y^{(c)})}{1 - Q^{(c)}} Q^{(c)}(1 - Q^{(c)}) \\ &= Y^{(c)}(Q^{(c)} - 1) + Q^{(c)}(1 - Y^{(c)}). \end{aligned} \tag{a-8}$$

Again, depending on the ground truth class, the righthand term in Eq. (a-8) simplifies to either $Q^{(c)} - Y^{(c)}$ or $Y^{(c)} - Q^{(c)}$.

In Eq. (2) of the main paper, we omitted the sum over k for the gradient of the cross-entropy loss with respect to parameter θ. The fix is reflected in (a-4). Note that this change does not affect the conclusion we make from these equations.

Appendix G Union of Categories

Table A-3. Naïve Union of label space for Cityscapes, BDD, IDD, and Mapillary: there are 70 categories and we list them for individual datasets.

Cityscapes	BDD	IDD	Mapillary	Cityscapes	BDD	IDD	Mapillary
		autorickshaw				obs-str-bar-fallback	
			banner				on rails
			barrier				other rider
			bench				other vehicle
bicycle	bicycle	bicycle	bicycle			parking	parking
			bicyclist				pedestrian area
			bike lane	person	person	person	person
			bike rack				phone booth
		billboard	billboard	pole	pole	pole	pole
			bird				pothole
			boat			rail track	rail track
		bridge	bridge	rider	rider	rider	
building	building	building	building	road	road	road	road
bus	bus	bus	bus				sand
car	car	car	car				service lane
			car mount	sidewalk	sidewalk	sidewalk	sidewalk
			caravan	sky	sky	sky	sky
			catch basin				snow
			cctv camera				street light
			crosswalk - plain	terrain	terrain		terrain
		curb	curb	traffic light	traffic light	traffic light	traffic light
			curb cut	traffic sign	traffic sign	traffic sign	traffic sign
			ego vehicle				traffic sign (back)
fence	fence	fence	fence				traffic sign frame
			fire hydrant				trailer
			ground animal	train	train		
		guard rail	guard rail				trash can
			junction box	truck	truck	truck	truck
			lane marking - crosswalk			tunnel	
			lane marking - general			utility pole	
			mailbox	vegetation	vegetation	vegetation	vegetation
			manhole			vehicle fallback	
motorcycle	motorcycle	motorcycle	motorcycle	wall	wall	wall	wall
			motorcyclist				water
			mountain				wheeled slow