Appendix: Adding Seemingly Uninformative Labels Helps in Low Data Regimes

A. The CSAW-S Dataset

As one of our contributions, we release the CSAW-S dataset, which can be found at https://github.com/ChrisMats/CSAW-S. CSAW-S is a curated dataset containing mammography images with annotations of breast cancer and breast anatomy from experts and non-experts. It can be used to replicate our study, repurposed for other semantic segmentation tasks, or used in conjunction with other breast cancer datasets (*e.g.* DDSM (Lee et al., 2017), INbreast (Bowyer et al., 2000)) to form a large repository of mammograms with tumor annotations.

A.1. Collection

The CSAW-S dataset contains mammography screening images from 172 different cases of breast cancer. It is split into a training/validation set containing 312 images from 150 patients and a test set of 26 images from 23 different patients. The data was split to ensure that roughly the same distribution of classes appears in the training and test splits. In total, 338 high-resolution grayscale images of both MLO (Mediolateral-Oblique) and CC (Cranial-Caudal) views appear in the dataset. The screening images composing the dataset were selected from CSAW, a large corpus of screenings gathered from Hologic devices in Stockholm between 2008 and 2015 (Dembrower et al., 2019). The dataset is annotated with 12 classes (see Figure 1) including the *cancer* class (the expert task), ten classes representing breast anatomy (complementary classes), and the *background* class. The cancer annotations are provided by three radiology experts. The choice of which classes to annotate was guided by a discussion with the radiologist experts with a semantic segmentation task in mind (labeling prominent anatomy for quality control and studying possible correlation with cancer). Non-experts provided annotations for the complementary classes in the training set. Experts provided annotations for the complementary classes in the test set but not the training set. The non-experts had no prior experience with mammography images, but received a short training session. The complementary classes are highly imbalanced, with some appearing very infrequently (*e.g.* lymph nodes and calcifications, see Table 1).

A.2. Preprocessing

The image files from the mammography device are in the DICOM image format. The metadata of each file contains

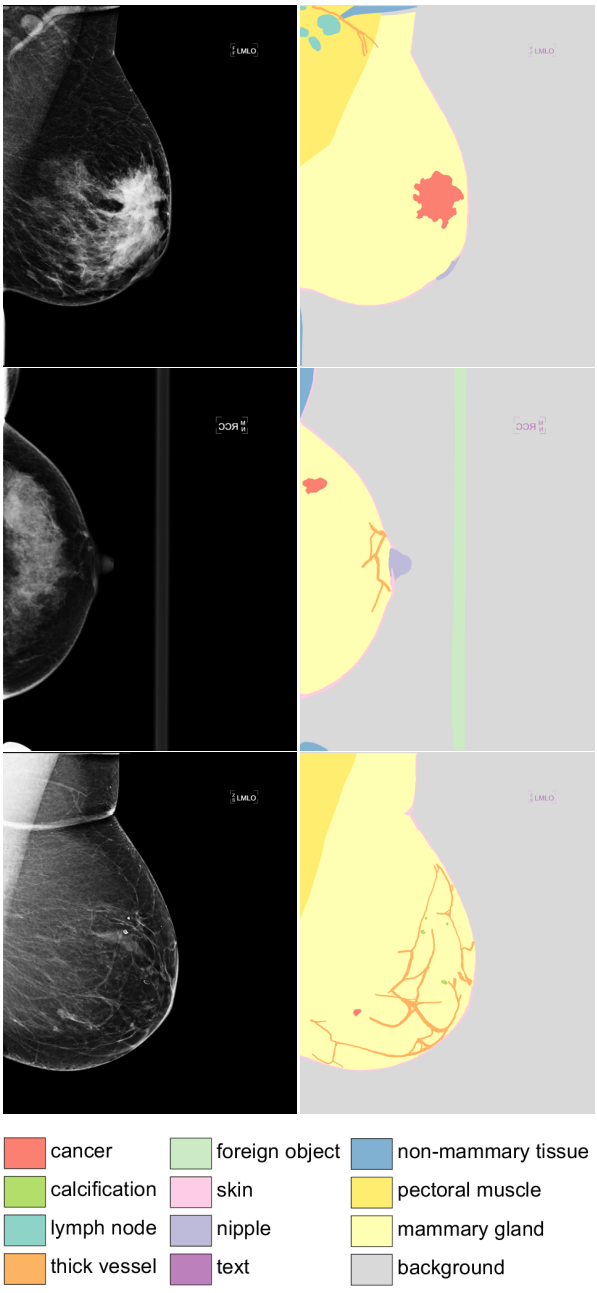

Figure 1. Three randomly selected examples from the CSAW-S dataset, which contains screening mammography images from 172 cases of breast cancer. Expert radiologist labels for cancer and *complementary labels* of breast anatomy made by non-experts are provided for all 342 images. The non-expert labels are in some cases imprecise or inaccurate (*e.g.* skin folds in top & bottom rows, or the improperly labeled object in lower-left of the middle row).

Table 1. Class ratios in the CSAW-S training set by pixel count and by image frequency reveal highly imbalanced classes.

PIXEL COUNT	(%)	IMAGE FREQUENCY	(%)
background	66.501	background	100.0
mammary gland	28.522	mammary gland	100.0
pectoral muscle	2.879	cancer	100.0
thick vessels	0.616	skin	100.0
skin	0.615	nipple	86.2
cancer	0.292	text	80.1
non-mammary tissue	0.271	thick vessels	77.9
foreign object	0.124	non-mammary tissue	55.4
nipple	0.082	calcifications	52.2
lymph nodes	0.063	pectoral muscle	50.0
text	0.031	lymph nodes	21.8
calcifications	0.003	foreign object	18.6

intensity windows (center and width) that determine the proper range of displayed pixel intensities accounting for differences in acquisition such as exposure, compression, etc. As a preprocessing step, we normalize each image using the DICOM window center and width metadata to re-scale the intensity range of the images. The re-scaling is done linearly and pixels outside the defined range are clipped. We used this approach because, as reported in (Clunie, 2003), the window values that are chosen by the operator or device result in consistent appearance for display. Many images exhibited inverted contrast (*i.e.* the background was white). To rectify this, we corrected the images by inspecting the DICOM photometric interpretation attribute which determines whether the minimum pixel value is black or white. Finally, we convert the DICOM files to 8-bit PNG images.

A.3. Expert and Non-expert Annotations

Cancer annotations for the training/validation set were provided by EXPERT 1. The complementary labels were sourced from seven non-experts with no medical training, one annotator per image. Each non-expert annotated between 15 and 109 images.

Along with each test image we provide cancer annotations from three expert radiologists: EXPERT 1, EXPERT 2 and EXPERT 3. Complementary labels of the breast anatomy are provided by both experts and non-experts for the test set. EXPERT 1 and EXPERT 2 provide the complementary labels (no complementary labels are provided by EXPERT 3 but they may be available in the future). Complementary labels are also provided by three non-expert annotators. The test set complementary labels were not utilized in our study, but are provided for other researchers interested in semantic segmentation of medical images.

Class labels of each image in the dataset are provided in the form of independent full pixel-wise binary masks for each class. These annotations were generated using QuPath (Bankhead et al., 2017), a tool for annotating large medical

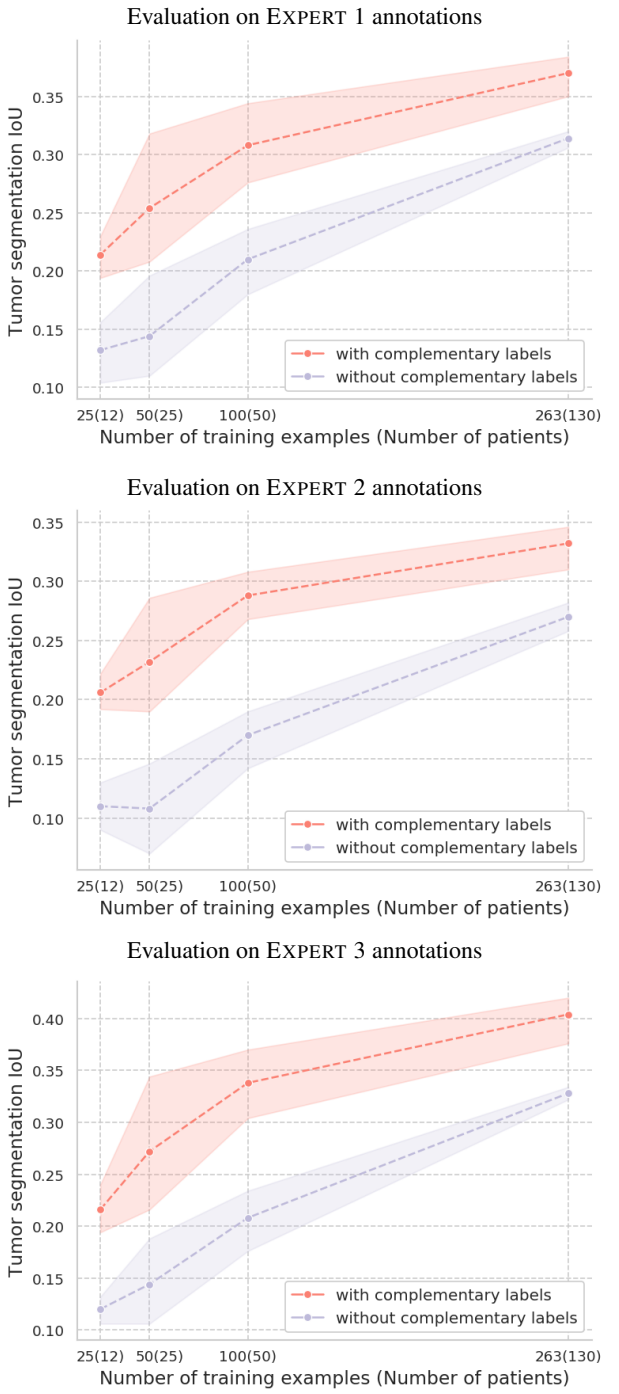

Figure 2. Results on CSAW-S when evaluating using annotations from different experts: EXPERT 1 (top) , EXPERT 2 (middle) and EXPERT 3 (bottom). The training data was annotated by EXPERT 1, biasing the models towards this expert. Evidently, training with complementary labels results in higher IoU scores for all cases. This difference is magnified when evaluating the annotations provided by EXPERT 2 and EXPERT 3, indicating an increased robustness to annotator bias when complementary labels are used. Interestingly, all models performed better when evaluated on EXPERT 3's annotations.

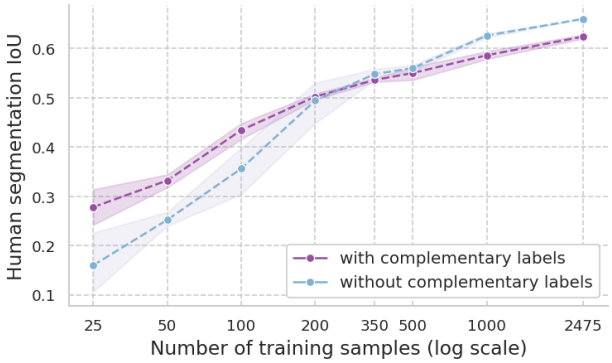

Figure 3. Results after merging the confusing classes *person* and *rider* on the CITYSCAPES dataset into a unified label *human* (the complementary classes are reduced from 32 to 31). Compared to Figure 5 in the main article, there is an absolute increase in IoU scores in both models, but the overall trend remains, and it is clear that additional complementary labels still help.

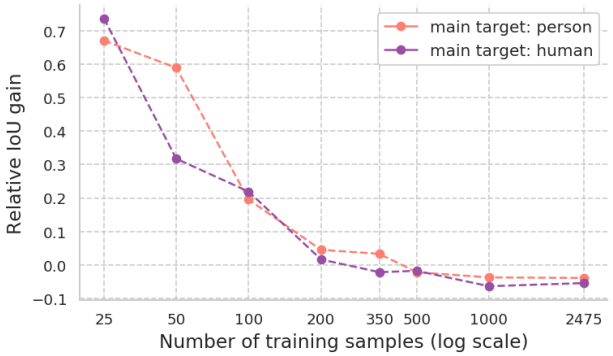

Figure 4. Relative IoU gains between the with- and without-complementary label models on CITYSCAPES when the main target is *person* (orange) and *human* (purple). The *human* class contains both *person* and the confusing label, *rider*.

images. Annotators were instructed to completely mark each class. In case of overlaps, all overlapping objects were labeled. In this work, we do not address multi-label classification (where each pixel can be associated with multiple classes). Therefore, the 11 binary masks for each image were combined to form a single multi-class mask as follows. Each pixel was assigned a single label by inspecting the 11 masks. Unannotated pixels were designated as background. Pixels with only one matching annotation receives the class label of the corresponding masks. In case of overlap, the labels were determined by the following priority order (with lower values having higher priority): (1) *cancer*, (2) *calcification*, (3) *lymph node*, (4) *thick vessel*, (5) *foreign object*, (6) *skin*, (7) *nipple*, (8) *text*, (9) *non-mammary tissue*, (10) *pectoral muscle*, (11) *mammary gland* and (12) *background*. The twelve-class masks generated by this process were used to train the networks.

A.4. Class Imbalance

The multi-class annotations for the CSAW-S dataset, produced as described above, are highly imbalanced (see Table 1). Classes such as *cancer*, *skin*, *mammary gland*, and *background* appear in every image, but with vast differences in area (*e.g. cancer* only accounts for 0.29% of pixels in the dataset while *mammary gland* accounts for 28.3%). Some classes are even more rare, such as *lymph nodes* and *calcifications* which are present in 22.2% and 51.2% of images (respectively), but only account for 0.064% and 0.004% of total pixels. These two particular classes are extremely rare in terms of image support but are also useful signs for diagnosing cancer. Therefore, these rare but important classes present an interesting segmentation challenge.

A.5. Generating Training Examples

Throughout our experiments on the CSAW-S dataset, we generate training samples by randomly extracting 512×512 patches from the full resolution mammography images. This is because (1) the high resolution of the mammograms cause memory issues during the training procedure, and (2) the small training set sizes necessitate the use of heavy augmentations, which also causes memory issues. To ensure more balanced class representation in the training data, we generate training examples for every image by sampling 10 center-cropped patches from locations belonging to each of the 12 classes (uniform sampling). We perform this data generation strategy offline once for each of the 5 experimental repetitions since it is an expensive procedure, in terms of memory and computation. The with- and without-complementary label models share the same training example crops for each experimental repetition.

During training we employ an extensive set of augmentations including rotations and elastic transformation in addition to standard random flips, random crops of 448×448, random brightness and random contrast augmentations on the 512×512 patches.

A.6. Evaluating Against Different Experts

The CSAW-S training set contains cancer annotations only from EXPERT 1. This biases networks trained on this data towards the opinion of this expert. The test set includes tumor annotations from EXPERT 1 and two additional expert radiologists. As seen in the main article, agreement between the experts is relatively low ($\approx$0.67), therefore we expect the IoU scores to differ when evaluating on annotations from the other experts. As we can see in Figure 2, the complementary labels result in increased performance over the baseline, regardless of which expert is considered. Furthermore, we can see that the generalization gap between the models trained with and without complementary labels increases as

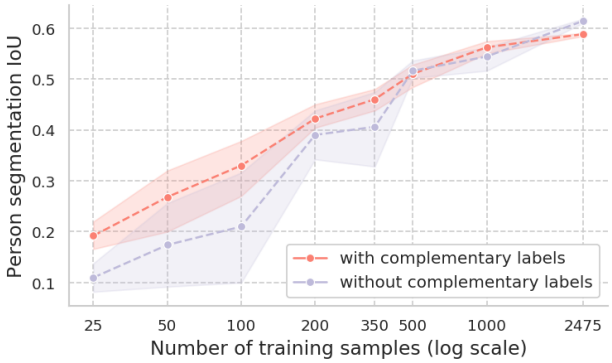

Figure 5. Results when imposing a domain shift between the training and test data by ordering the training examples. Compared with Figure 5 from the main text, we note that the domain shift causes an absolute drop in performance for both models, but the performance gap remains steady.

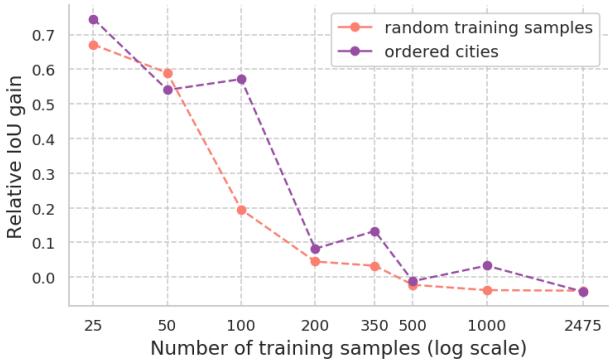

Figure 6. Relative IoU gains (between models trained with and without complementary labels) when imposing a domain shift between the training and test data (purple), and when there is no domain shift (orange). Both cases show models trained with complementary labels. Note that, although the purple curve is noisier, there seems to be little effect on IoU performance gap between the two setups, indicating robustness to domain shifts.

we evaluate on EXPERT 2 and EXPERT 3. This indicates that the model with complementary labels is more robust to annotator bias. Interestingly, all models performed better when evaluated on EXPERT 3's annotations.

B. Sanity Checks For the Training Procedure

We ran a series of sanity checks to satisfy any concerns regarding the standard training protocol used in our experiments. We report that our finding – a significant generalization gap between models with and without complementary labels – remains consistent regardless of the training technique used.

In detail, we tested the following variations to our standard training procedure:

- We replaced the DeepLabv3 segmentation model with FCN-32 (Long et al., 2015). We found no change in the generalization gap.
- We replaced GroupNorm with BatchNorm. We found no change in the generalization gap.
- Instead of using IMAGENET pretrained models, we randomly initialized the weights. We found no change in the generalization gap, although there was a significant drop in performance for both cases.
- We froze the normalization statistics *1)* throughout the training process and *2)* for the first 60% of the training iterations and fine tune for the rest 40%. We found that freezing the normalization statistics did not help the for CSAW-S dataset but for the CITYSCAPES and PASCAL VOC there were significantly large improvements, especially towards the extreme low data regime. As has been reported before (Raghu et al., 2019; He et al., 2019), this is because the statistics dramatically change as we move from the natural domain to the medical one.
- We extended the default augmentations with random scale and resizing as well as elastic transforms. As expected, augmentations resulted in increased generalization for every case. Nonetheless, we found that heavy augmentations favour the models trained with complementary labels slightly more.
- We evaluated the importance of complementary labels when the main target is trivial (e.g. class *sky* and *ego vehicle* in CITYSCAPES). We consider a class as trivial when it has many of the following traits which make it easily distinguishable: little texture variation, regular shape, clear edges, and large pixel-count per image. For these cases, we find high IoU scores even when limited examples are present. We also found that the most important factor is the normalization method, and gains from the complementary labels only persist in the extreme low data regime (less than 10-20 training examples).

Finally, we note that we tuned our training settings for the case where only expert annotations are available (no complementary labels) and set them as default to ensure that the gains from the complementary labels are valid. We did not further tune the settings for models with complementary labels, so this likely leads to sub-optimal settings and an under-reporting of the actual generalization gap.

C. Easily Confused Classes

The class *person* is easily confused with the class *rider* in the CITYSCAPES dataset. We investigated whether gains from the complementary labels for the *person* segmentation task are due to the explicit modeling of the *rider* class. In other words, are the observed effects attributed to complementary labels actually due to the modeling of easily

Table 2. Class ratios in the PASCAL VOC dataset by pixel count and by image frequency.

PIXEL COUNT	(%)	IMAGE COUNT	(%)
background	64.7	ambiguous	99.9
ambiguous	5.9	background	99.5
person	5.2	person	29.9
cat	3.1	chair	9.9
bus	2.0	cat	8.9
dog	1.9	car	8.5
train	1.8	dog	8.3
car	1.6	bird	7.2
sofa	1.5	sofa	6.4
motorbike	1.3	aeroplane	6.0
dining table	1.3	bottle	5.7
chair	1.2	train	5.7
horse	1.1	tv/monitor	5.7
bird	1.0	dining table	5.6
tv/monitor	1.0	motorbike	5.5
sheep	1.0	potted-plant	5.5
cow	0.9	boat	5.3
aeroplane	0.9	bus	5.3
boat	0.7	horse	4.6
potted-plant	0.7	bicycle	4.4
bottle	0.7	cow	4.4
bicycle	0.3	sheep	4.3

confused classes? Intuitively, explicitly modeling the *rider* class, which would otherwise be part of the background (in the without-complementary labels case), should help reduce false positives. To investigate, we merged the annotations of the classes *person* and *rider* into a unified label *human* and repeated the experiments. Although the absolute IoU scores for both models improved (Figure 3), the relative IoU gains from the complementary labels showed no appreciable change (Figure 4). Thus, we infer that although the explicit modeling of confusing classes is helpful (see label importance in the main text), it is not the only reason that including complementary labels are beneficial in low data regimes.

D. Robustness to Domain Shifts

As described in the main article at the end of Section 5.3, we tested if adding complementary labels improves model robustness to domain shifts. Our goal in this experiment is to test if complementary labels help when there is a domain shift between the training distribution and the test distribution. For example, if medical images acquired from certain devices appear in the training set, and images from a different set of devices appear in the test set.

We set up an experiment in which domain shifts were artificially imposed in the training data as follows. An ordered training set is created by shuffling images individually from each city, and then placing them in a random order grouped by city. For example, this may result in a training set with images from Stuttgart, then Aachen, Hamburg, etc. Then, we repeat the experiments for *person* segmentation. The models are trained using subsets of the randomly ordered and shuffled training sets, with the same schedule as our main experiments $N = \{25, 50, 100, 200, 350, 500, 1000, 2475\}$. We repeat this 5 times for each N. The result of this procedure is that models trained with $N = 25$ or $N = 50$ will only see data from a single city, but will be tested on a set containing all cities (each city contains between 77 and 259 images). This represents our imposed domain shift. As more data is added, more cities will appear in the training set.

We find that adding complementary labels improve performance in the presence of domain shifts (Figure 5). Although the absolute IoU performance is lower when the domain shift is imposed than than when there is no domain shift, the performance gap between models with and without complementary labels holds and widens to some extent (Figure 6). Furthermore, while the curve for the model trained with

Table 3. Class ratios in the CITYSCAPES dataset by pixel count and by image frequency.

PIXEL COUNT	(%)	IMAGE COUNT	(%)
road	36.61	ego vehicle	100.0
building	20.06	pole	98.8
vegetation	14.35	static	98.7
car	6.69	road	98.6
sidewalk	5.41	building	98.5
sky	3.04	vegetation	97.0
ego vehicle	2.43	car	94.8
static	1.42	traffic sign	94.4
ground	1.29	sidewalk	94.0
person	1.26	sky	86.3
pole	1.18	person	78.5
terrain	1.05	traffic light	55.8
fence	0.77	terrain	54.9
parking	0.63	bicycle	53.7
wall	0.58	dynamic	44.7
traffic sign	0.52	fence	42.1
bicycle	0.4	ground	34.7
bridge	0.3	rider	34.2
dynamic	0.29	wall	31.6
train	0.27	parking	24.4
truck	0.23	rect. border	22.2
bus	0.22	unlabeled	19.7
rect. border	0.22	motorcycle	17.1
traffic light	0.19	truck	11.9
rail track	0.19	bus	8.6
rider	0.14	out of roi	8.6
motorcycle	0.1	bridge	7.5
tunnel	0.05	polegroup	7.2
caravan	0.04	train	4.4
out of roi	0.03	rail track	3.5
trailer	0.02	trailer	2.5
unlabeled	0.01	caravan	1.9
guard rail	0.01	tunnel	0.8
polegroup	0.01	guard rail	0.6

Figure 7. Segmentation results on CITYSCAPES show complementary labels improve performance in low-data settings. From top to bottom: the full image, fine annotations and predictions from networks trained with only the *person* labels (*blue*) and predictions from a network trained with both *person* labels and complementary labels (*red*) using $N = \{25, 100, 350, 2475\}$ training examples. In the last row, complementary labels begin to hurt performance (the crossover in Figure 5 in the main text).

complementary labels retains the same shape as in Figure 5 from the main text, the curve for the model trained without complementary labels shows more variance, and is sensitive to steps where new cities are added (Figure 5).

E. Label Frequency on CITYSCAPES and PASCAL VOC

In Table 2 and Table 3, we report the number of pixels and the number of images that contain each class for CITYSCAPES and PASCAL VOC. Note that the CITYSCAPES dataset has severe class imbalance whereas PASCAL VOC is more balanced – the most and least frequent classes differ by at most one order of magnitude at most (excluding the background and ambiguous class). The greater pixel count imbalance in CITYSCAPES corroborates our qualitative observation that objects in CITYSCAPES are more likely to appear at different scales than in PASCAL VOC. For complex classes like *person*, this can make the segmentation problem significantly harder and may partially explain the performance difference between the two datasets.

Figure 8. Segmentation results on PASCAL VOC show complementary labels improve performance in low-data settings. From top to bottom: the full image, annotations and predictions from networks trained with only the *person* labels (*blue*) and predictions from a network trained with both *person* labels and complementary labels (*red*) using $N = \{200, 500, 1464\}$ training examples.

F. Segmentation Results on CITYSCAPES and PASCAL VOC

In Figure 5 and Figure 6 of the main text, we showed quantitatively that adding complementary labels results in increased IoU scores for CITYSCAPES and PASCAL VOC in low data regimes. In Figure 7, we visualize results for CITYSCAPES and confirm that segmentations from models trained with complementary labels result in more accurate segmentation masks in low data regimes where segmentation are generally poor. As we increase the number of training examples, we see diminishing returns when adding complementary labels. When a large number of examples are included in the training procedure (last row of Figure 7), complementary labels begin to hurt performance.

In Figure 8, we visualize the model predictions on PASCAL VOC and confirm that the segmentation results also reflect the trend in IoU we reported in the main text: when limited samples are present in the training set, the addition of complementary labels results in better segmentations.

References

Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., McQuaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., et al. Qupath: Open source software for digital pathology image analysis. *Scientific reports*, 7(1):1–7, 2017.

Bowyer, K., Kopans, D., Moore, R., Sallam, M., Chang, K., and Woods, K. The digital database for screening mammography. In *Proceedings of the 5th international workshop on digital mammography*, pp. 212–218, 2000.

Clunie, D. A. Dicom implementations for digital radiography. *RSNA*, 2003:163–172, 2003.

Dembrower, K., Lindholm, P., and Strand, F. A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks—the cohort of screen-aged women (csaw). *Journal of digital imaging*, pp. 1–6, 2019.

He, K., Girshick, R., and Dollar, P. Rethinking imagenet pre-training. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., and Rubin, D. L. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4:170177, 2017.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, pp. 3342–3352, 2019.