University of Nebraska-Lincoln
Progress report

Proposal# 129426
Oct 12th, 2019

Progress Report on Second Iteration

Yi Liu

1. Figure/Graph Extraction for Beyond Words Collection

In the first iteration of this project, we proposed a two-step approach to extract figure/graph and generate metadata for the Beyond Words collection. The first step, an FCN (U-NeXt) combining ResNeXt and U-Net was built and trained to segment and classify graphic snippets on newspaper pages based on ground truth extracted from Beyond Words. Besides, the ResNeXt part of the model was transferred from pre-trained ImageNet ResNeXt-101 to reduce training parameters. Based on dhSegment, using transfer learning is able to boost training effectiveness, and preserve a good performance. The second step, a text segmentation, and recognition model retrieved textual content in the graphic snippets (i.e. extracted graphic snippets from the first step). Specifically, EAST text detection was applied to find text regions for an OCR process to retrieve words within graphic snippets. And the retrieved word was encoded into metadata for further usages, such as search queries.

In the second iteration, we focus on *evaluating and improving the segmentation step using the U-NeXt model*. The U-NeXt model is an extension of dhSegment model. The dhSegment model used pre-trained ResNet while our U-NeXt used pre-trained ResNeXt model. Note that ResNeXt is an improved version of ResNet. In the study of dhSegment on the Beyond Words collection, the classification accuracy was 88% and the mean intersection over union (mIoU) was 26%. The U-NeXt is expected to have a better performance than dhSegment. Further, the EAST text detection largely depends on the performance of the segmentation step. Hence, improving segmentation step is a key component of this project.

Note that two metrics are used to evaluate the performance of this project. First, the classification accuracy is a pixel-wise accuracy. It computes the percentage of correctly labeled pixels to total numbers of pixels for each class. Second, the mIoU evaluates whether the predicated region accurately covers the true region in the ground-truth overall classes.

2. Datasets

Two datasets were used to train and evaluate the segmentation step using the U-NeXt model, the Beyond Words collection and the European Newspapers collection. In the Beyond Words collection (BW), some graphic regions appeared on a page are missing in the ground-truth. And the marked region in the ground-truth does not tightly map to the actual shape of the graphic region. This lack of reliable ground-truth in the BW collection led us to pre-training our model on a more comprehensively labeled dataset called the European Newspapers collection (ENP). By doing so, during training, some local minimum, created by the aforementioned issues in the Beyond Words collection, could be avoided. Specifically, the ENP contains 480 images in total, in which, there are 384 images in training set and 96 images in validation set. And the BW contains 1,532 images in total, in which, there are 1,226 images in training set and 306 images in the validation set.

Further, the similarity shared by ENP and BW collections is the crucial reason why the ENP can be used for pre-training. First, both ENP and BW collections are document images that are digitized from newspapers. Hence, they share a similar content layout and density. Second, the ground-truth of the ENP marked five classes: (1) background, (2) text, (3) figure, (4) layout separator, and (5) table; while the ground-truth of the BW marked background and five detailed types of figures. Hence, the learned knowledge from the pre-trained model on figures using the ENP provides a good reference for U-NeXt to identify the figure region. Then, the fine-tuning using the BW could focus on detailed figure type differentiation than identify the figure region.

3. Experimental Results

In this experiment, early stopping is not applied since the expectation of the performance is unknown. Here we report on two sets of results: from the pre-training experiment and from the fine-tuning experiment:

- The pre-training experiment involved training and testing on the ENP dataset (up to 700 iterations).
- The fine-tuning experiment involved four different approaches.
  - The first approach trained and tested on the BW dataset without using the ENP-trained classifier. This is meant to serve as a baseline design.
  - The second approach used the above ENP-trained classifier as the beginning classifier, and training and testing it on the BW dataset (up to 80 iterations). We added this design because using a pre-trained classifier for a similar task could help the fine-tuning experiment address the issue of lack of ground truth data mentioned in the previous section. This second approach is a variant of the first approach.
  - The third approach replaced the deconvolutional layer with a resizing layer in the deep learning model, and training and testing on the BW dataset. Since the deconvolutional layer is known to suffer from the "checkerboard" issue [Distill 2016], the resizing layer is seen as a potential improvement technique. This third approach is thus a variant of the first approach.
  - The fourth approach performed a two-class segmentation, instead of six classes on the BW dataset for both training and testing. This is because the training dataset is biased where there is a predominantly large number of background pixels compared to other classes of pixels[1]. By collapsing all the object pixels into one class, we hope to reduce the imbalance in the number of pixels in each class during training. This fourth approach is thus also a variant of the first approach.

3.1 Pre-training Experiment

Figure 1 shows the training performance of the pre-training experiment reaches 91.30% on pixel-wise accuracy and 57.19% on mIoU. And the testing performance is 81.90% on pixel-wise accuracy and 48.18% on mIoU. From the result, the convergence is observed (i.e., the tendency of accuracy gets close to 100% percent). The observed convergence indicates the parameters are getting trained to fit the task; hence, the model is ready for fine-tuning.

3.2 Fine-tuning Experiment 1: without pre-trained ENP classifier

Figure 2 shows the performance of the experiment without using the pre-trained ENP classifier reaching 89.08% on training pixel-wise accuracy, 50.43% on training mIoU, 80.11% on testing pixel-wise accuracy and 38.00% on testing mIoU. The experiment lasts 80 epochs, and the convergence is observed on both training and testing curves. However, the testing curve shows instability, that, although the tendency towards higher testing accuracy, the testing accuracy varies high and low rapidly during the experiment. And Table 2 (row 1 - 4) shows the class-wise testing performance on accuracies and mIoUs. The class-wise stats show that the classifier failed to recognize classes of editorial cartoons, illustrations, and maps. These

[1] There are 88.21% pixels in background class, but for the rest of classes, only 0.71% in editorial cartoon class, 2.89% in comics/cartoon class, 1.38% in illustration class, 6.64% in photograph class, and 0.18% in map class.

three classes happen to be the top three rarest classes. Hence, the misrecognition issue is likely caused by the rareness of corresponding classes. However, overall, the performance of the classifier is promising since both training and testing accuracies reached 80% within only 80 training epochs.

3.3 Fine-tuning Experiment 2: using pre-trained ENP classifier

Figure 3 shows the performance of the experiment using the pre-trained ENP classifier reaching 89.41% on pixel-wise training accuracy, 41.21% on training mIoU, 85.53% on testing accuracy and 38.57% on testing mIoU. Though the performance indicators above might look promising, upon further investigation, the classifier trained during the fine-tuning experiment attempted to classify as many pixels as background pixels after training convergence. Table 2 (row 5 - 8) shows the class-wise stats. We see that, after the convergence, all training and testing stats for non-background classes are zero. Hence, the performance stats are better than the first fine-tuning experiment numerically, but the actual performance is worse since none of the objective class was recognized. As previously mentioned, the background pixel is the majority over all pixels of the BW dataset. Such imbalance could create a “deep” local minimum. We suspect that the classifier fell into the “deep” local minimum. And it could not “jump” out from the minimum. In fact, the large fluctuations at the beginning epochs are indirect evidence. It shows that the classifier tried but failed to “jump” out from the minimum. However, the advantage of using pre-trained ENP classifier is the faster converging speed. Therefore, by taking such advantage, the computational resources could be saved comparing to others.

3.4 Fine-tuning Experiment 3: using resizing layer

Figure 4 shows that, for testing performance, the pixel-wise accuracy reached 86.69% and the mIoU reached 37.84%. The performance did not show clear improvement, because the pixel-wise testing accuracy is higher while the testing mIoU lower than the experiment 3.2. More, similarly, in the class-wise performance, shown in Table 2 (row 9 - 12), we also found that pixel-wise accuracy and mIoU of the editorial cartoon, illustration, and map classes are zeros. However, the curve in Figure 4 did not show the instability like experiment 3.2. Hence, the instability likely came from the “checkerboard” issue since the resizing layer was introduced to solve the issue. Therefore, from the perspective of stability, using the resizing layer has better performance than experiment 3.2.

3.5 Fine-tuning Experiment 4: combined two-class segmentation

Training a classifier to learn information from rare classes is very hard. Hence, combining five non-background classes into one class could decrease the complexity of the task, which could lead to improvements. In fact, pixels in non-background classes only 11.79% of the entire training dataset in total. And in this experiment, Figure 5 shows the combined class segmentation outperformed all other fine-tuning experiments. That is, for training performance, the pixel-wise accuracy was 91.76% and the mIoU was 71.44%; and, for testing performance, the pixel-wise accuracy was 88.89% and the mIoU was 64.97%.

Table 1 Average performance of fine-tuning experiments

	Without Pre-trained ENP Classifier		Using Pre-trained ENP Classifier		Using Resizing Layers		Combined Two-class Segmentation	
	Train	Test	Train	Test	Train	Test	Train	Test
Accuracy	89.08%	80.11%	89.42%	85.53%	88.90%	86.69%	91.76%	88.89%
mIoU	50.43%	38.00%	41.21%	38.57%	51.31%	37.84%	71.44%	64.97%

Table 2 Class-wise statistics of fine-tuning experiments

			Background	Editorial Cartoon	Comics/ Cartoon	Illustration	Photograph	Map
Without Pre-trained ENP Classifier	Train	Accuracy	92.70%	0.00%	10.66%	0.00%	92.11%	0.00%
		mIoU	90.81%	0.00%	7.00%	0.00%	54.46%	0.00%
	Test	Accuracy	84.43%	0.00%	44.82%	0.00%	72.38%	0.00%
		mIoU	79.99%	0.00%	24.97%	0.00%	52.09%	0.00%
Using Pre-trained ENP Classifier	Train	Accuracy	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		mIoU	89.42%	0.00%	0.00%	0.00%	0.00%	0.00%
	Test	Accuracy	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		mIoU	85.53%	0.00%	0.00%	0.00%	0.00%	0.00%
Using Resizing Layers	Train	Accuracy	90.87%	6.32%	41.85%	3.26%	88.22%	0.00%
		mIoU	89.46%	5.16%	29.07%	2.61%	47.90%	0.00%
	Test	Accuracy	97.83%	0.00%	4.19%	0.00%	40.60%	0.00%
		mIoU	87.38%	0.00%	0.20%	0.00%	34.24%	0.00%

			Background	Non-Background
Combined Two-class Segmentation	Train	Accuracy	91.02%	90.45%
		mIoU	90.22%	52.66%
	Test	Accuracy	92.82%	68.18%
		mIoU	86.64%	43.29%

4. Conclusion

In this second iteration of the figure/graph extraction task, we tested our proposed U-NeXt model during the first iteration of exploration. The pre-training stage used the ENP collection. Though the pre-training performance was promising, it was not very strong. In addition, the fine-tuning stage with several experiments used the BW collection as well as other improvement techniques reported in machine learning. The fine-tuning experiments showed evidence that the issue in BW collection affected the performance.

Further, according to the visualized extraction result, we found two widespread issues in the BW ground truth. First, the missing component issue appears to be quite widespread in the BW ground truth data. For example, shown in Figure 5, a large portion of a photograph in the document is missing from the ground truth, but is captured by our U-NeXt classifier. Second, there are inaccurate rectangular regions. For instance, shown in Figure 6, the ground truth region includes incorrectly a large portion of the text content. In future work, these issues are a good starting point for improving the BW ground truth.

However, we found a very interesting occurrence where the U-NeXt classifier tried to fit the exact shape of the figure/graph region. For example, shown in Figure 7, the classifier prediction tried to fit the exact shape of the eagle on the right-hand side of the newspaper page. We speculate that the light background of the figure/graph region might have confused the classifier. And this may suggest that the actual performance of the U-NeXt may be better than the statistical evaluation used in our experiments (i.e., pixel-wise accuracies and mIoU).

Hence, we propose two ways to continuously improve the performance of this figure/graph extraction task.

- First, splitting the figure/graph extraction task to a pipeline of two tasks: (1) extraction of graphics from the background and textual content and (2) classification of the extracted graphics to detailed graphic types. Such arrangement would reduce the complexity of the task.
- Second, there is still room to improve the U-NeXt model for the extraction task directly. For example, the resizing layer can improve the performance of our experiment.

At this stage, it is hard to say which of the above solutions would yield better results. They all have advantages. The model using a pre-trained ENP classifier converges faster; The resizing layer avoids the "checkerboard" issue and improves stability. And the combined class segmentation can decrease the task difficulty while the direct six-class segmentation can avoid introducing complexity from pipelining two tasks.

Furthermore, we found that, because the classifier tries to fit the exact shape of the graphical content, the actual classification performance may be higher than the statistical evaluation indicated. However, comparing to the issues from the U-NeXt model, the major problem is that the BW ground truth *has two widespread quality issues to be fixed*. We believe that performance improvement will be observed if the ground truth issues can be removed.

- Therefore, in the next iteration of this projection, work should also be done on the BW ground truth to fix the quality issues as well as the imbalance class issue. Specifically, increase the number of pixels for rare classes.

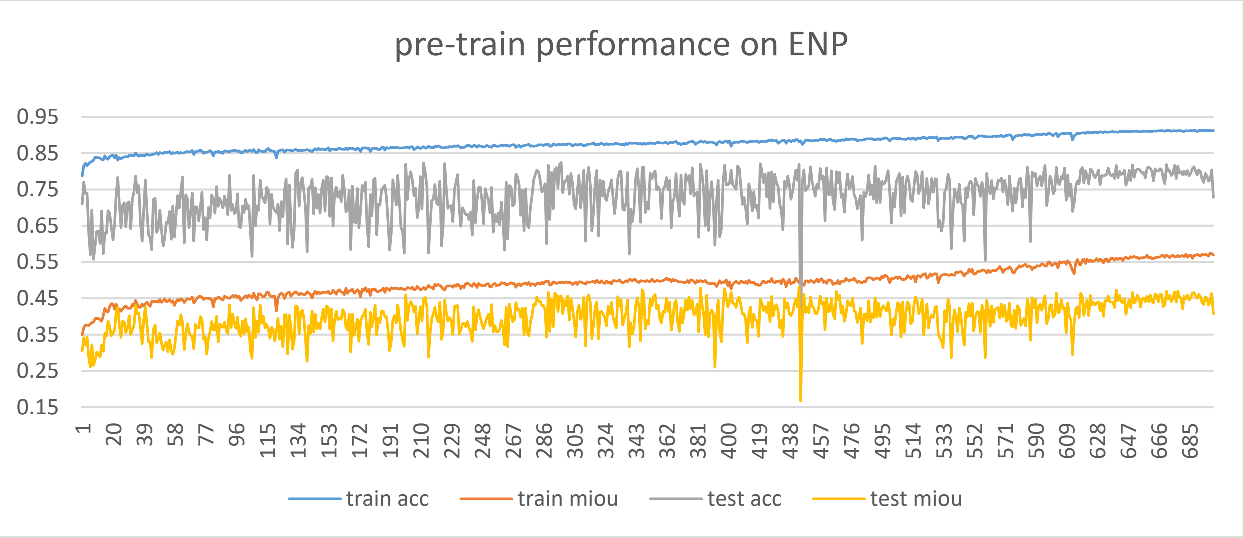


Figure 1 Pre-train ENP classifier performance.

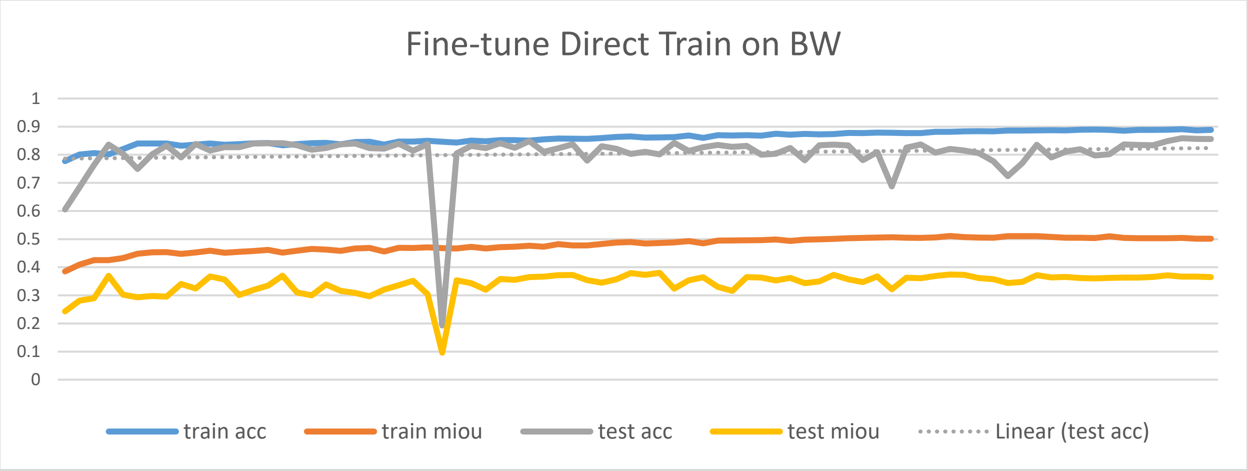


Figure 2 Fine-tuning experiment 1 - the baseline.

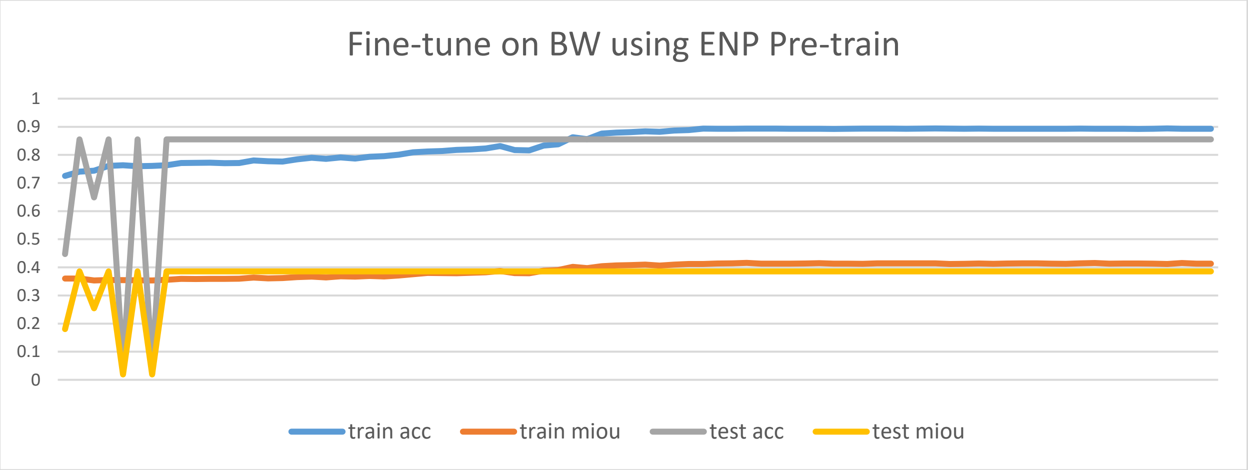


Figure 3 Fine-tuning experiment 2 - using pre-trained ENP classifier.

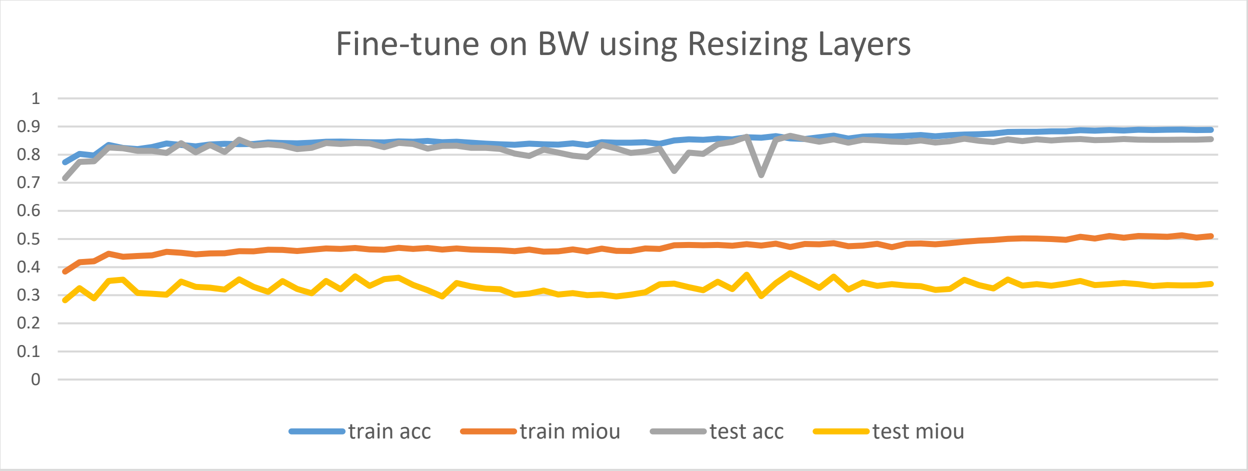


Figure 4 Fine-tuning experiment 3 - using resizing layers.

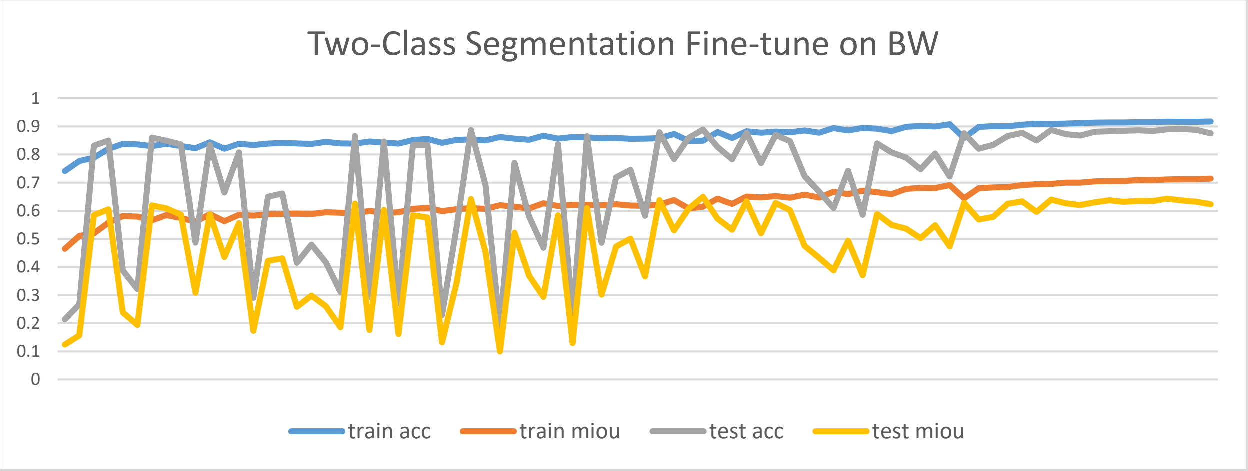
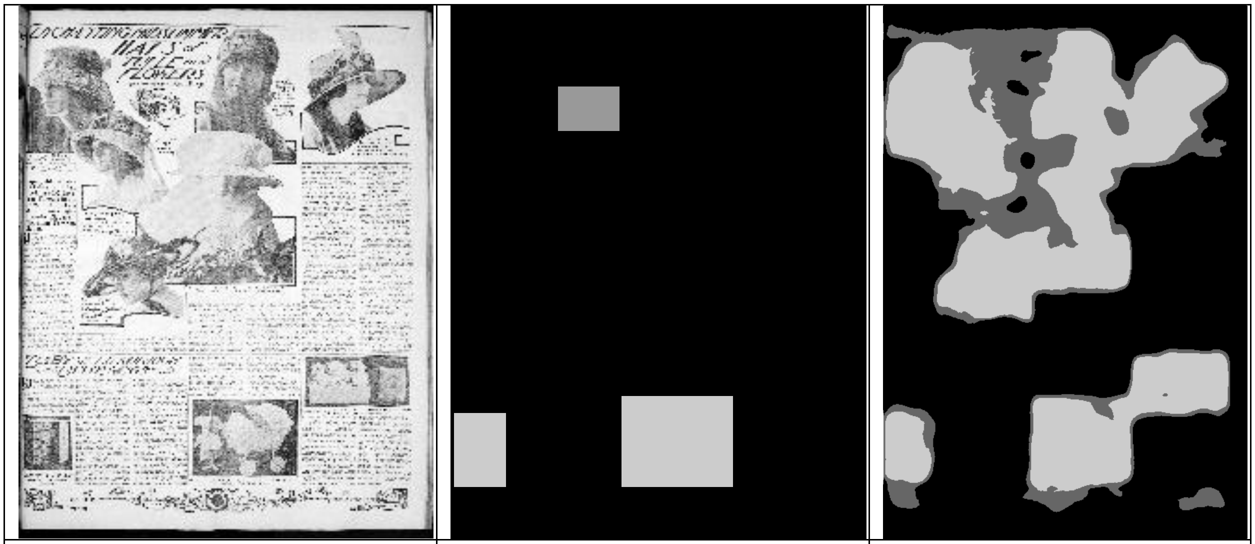
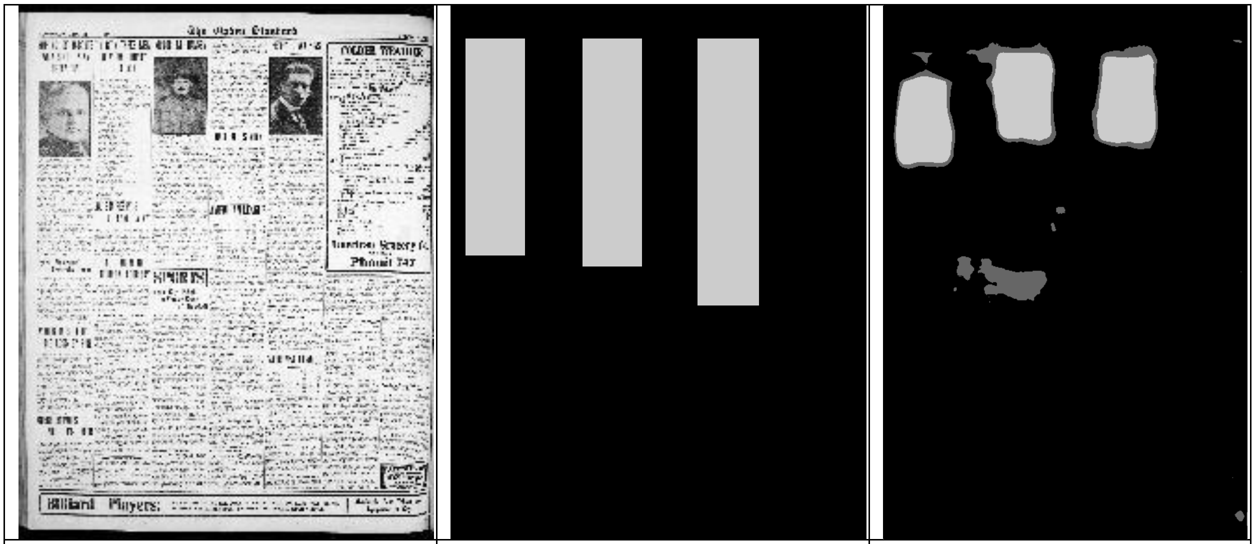


Figure 5 Fine-tuning experiment 4 - combined two-class segmentation.



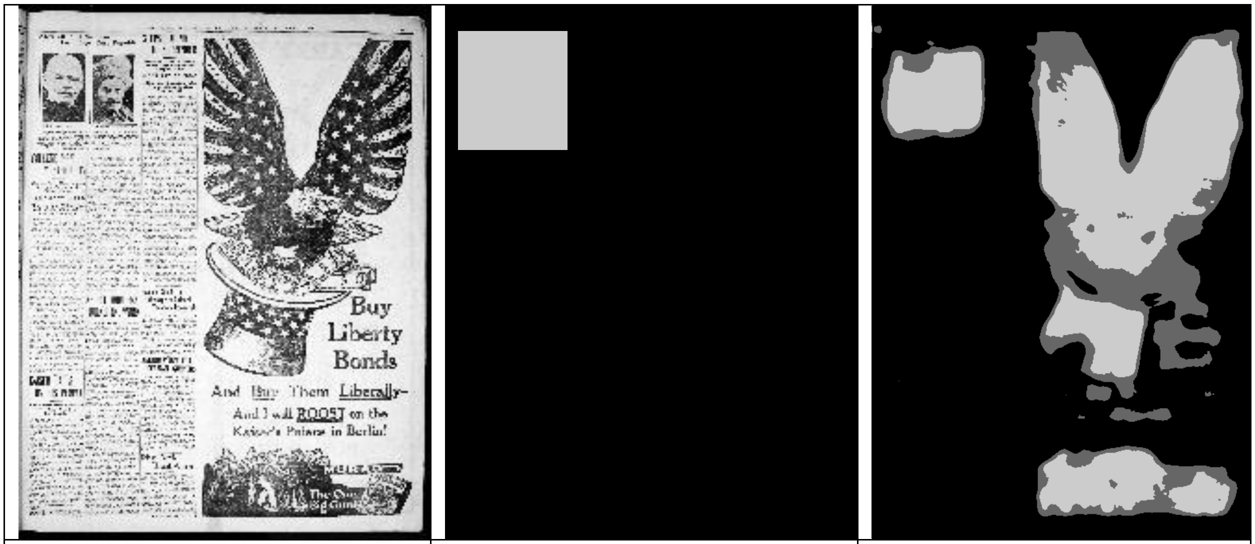
Newspaper page | Ground truth | U-NeXt Prediction

Figure 6 The missing component issue



Newspaper page | Ground truth | U-NeXt Prediction

Figure 7 The extra text content issue



Newspaper page | Ground truth | U-NeXt Prediction

Figure 8 Classifier tried to fit the exact shape of the graphic content