Document Structure Extraction using Prior based High Resolution Hierarchical Semantic Segmentation (Supplementary)

Mausoom Sarkar[1], Milan Aggarwal[1⋆], Arneh Jain[2⋆], Hiresh Gupta[2⋆], and Balaji Krishnamurthy[1]

[1] Media and Data Science Research Labs, Adobe
[2] Adobe Experience Cloud

1 Form Structure Extraction: Re-flow Application

Document structure extraction has been performed for digitising documents to make them re-flowable which is useful in web based services. Organisations across various domains, such as finance, administration, healthcare etc., which have been using paper forms or flat PDF forms would want to digitize them by converting them into an appropriate digitised version (such as an HTML). Once these forms are made re-flowable, they can be used on many devices with different form factors so that whole form layout can be rendered dynamically as shown in figure 1. This availability across devices automatically increases the ease of doing business or provide services since people can interact with them easily. Form digitisation also enables other capabilities such as better handling of data filled in digitised version, applying validation checks on data filled in fields, consistent form design control, auto-filling similar fragments in a form etc.

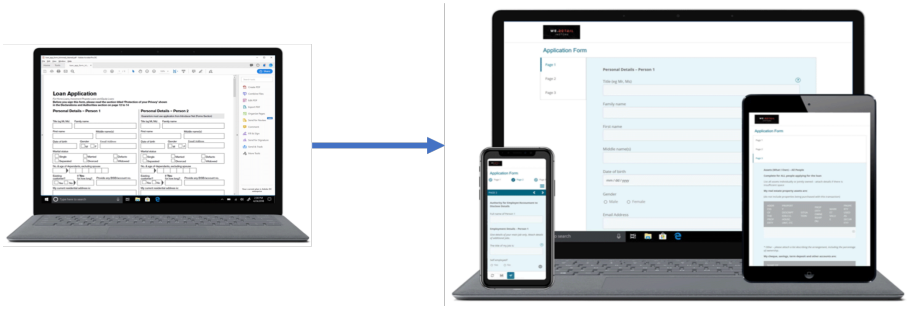

Fig. 1: Form digitisation to re-flow a form (left) on multiple devices (right).

⋆ equal contribution

2 Different complexity in form structure compared to regular documents

While most regular documents comprise of few large blobs that are well separated and easily distinguishable, forms are much more dense, complex and contains closely spaced structures at different levels of hierarchy. This has been depicted in figure 2.

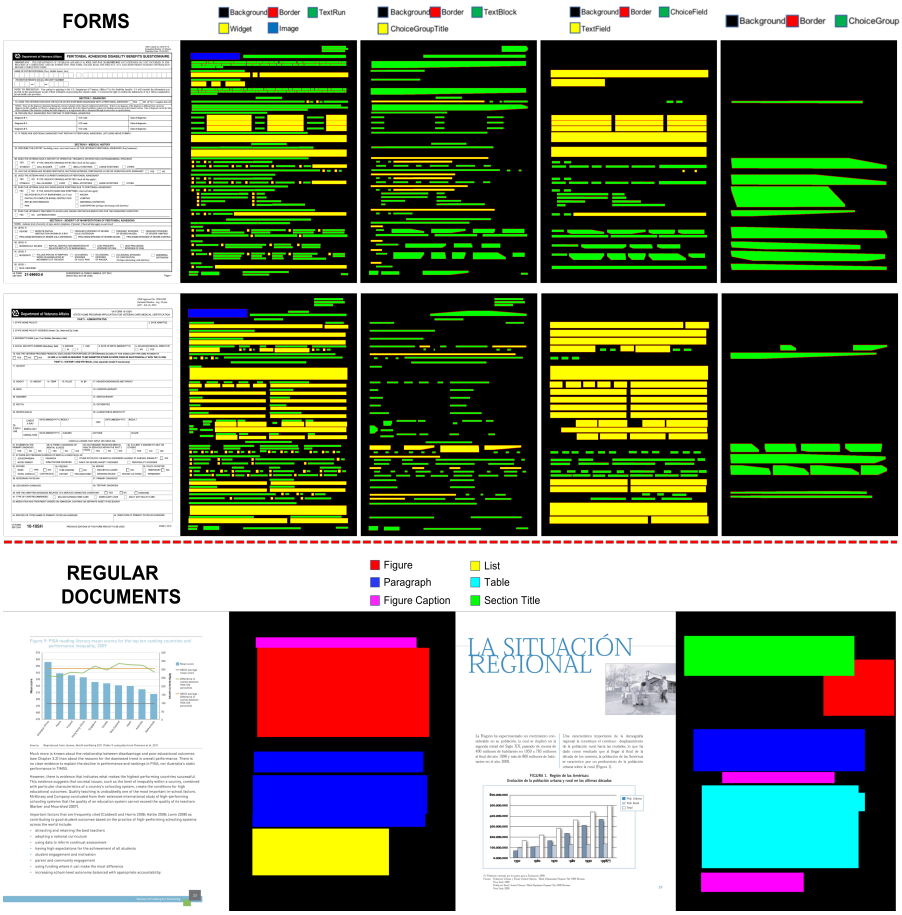

Fig. 2: Visualisation for highlighting the difference in structure complexity between forms (top 2 rows) and regular documents (bottom row). For forms, first image in the row is the form image followed by hierarchical segmentation masks for different structures. For documents, we show two examples with document images followed by their segmentation mask.

Table 1: Effect of varying strip size and overlap height during inference on MIoU

Structure →	Text	Widget	Text	ChoiceGroup	Text	Choice	Choice
Strip Size/Overlap ↓	Run		Block	Title	Field	Field	Group
600/200	92.7	87.3	90.5	80.8	88.8	84.0	83.0
600/300	92.6	87.1	90.5	80.4	88.7	84.1	82.9
840/360	92.5	87.3	90.6	80.7	88.8	83.2	82.3
600/0	91.3	83.5	89.7	77.6	85.1	79.1	79.3

3 Varying Strip Size and Overlap Height

Table 1 shows effect of varying strip size(SS) and overlap height(O) at the inference stage for model trained at SS=600 and O=200. Comparing 600/200 with 600/300, there is not much change, possibly because overlap 200 is sufficient. Increasing height(840/360 specification) did not provide any improvements. However on decreasing overlap to 0(600/0), it can be observed that there is a drastic drop in MIoU which shows overlap is important.

4 Comparison of Time and Number of Parameters

Time taken during inference on an image(averaged over 1000 images) by our Highres network on a single gpu (for our main configuration Strip Size(SS)=600 and Overlap b/w strips(O)=200) is 0.45s while DLV3+ takes 0.079s. Our Lowresnet that operates at the same resolution(792) as DLV3+ takes 0.107s. Although, HighResNet takes more time, it provides a strong boost in MIoU and F1 as discussed in the paper. In terms of **parameters**, DLV3+ comprises of 59.46M parameters while our HighResNet model has 94.34M parameters.

5 Prediction Visualisations

Some visualisations comparing predictions made by different ablation methods and baselines with our main model have been shown in figure 3, 4.

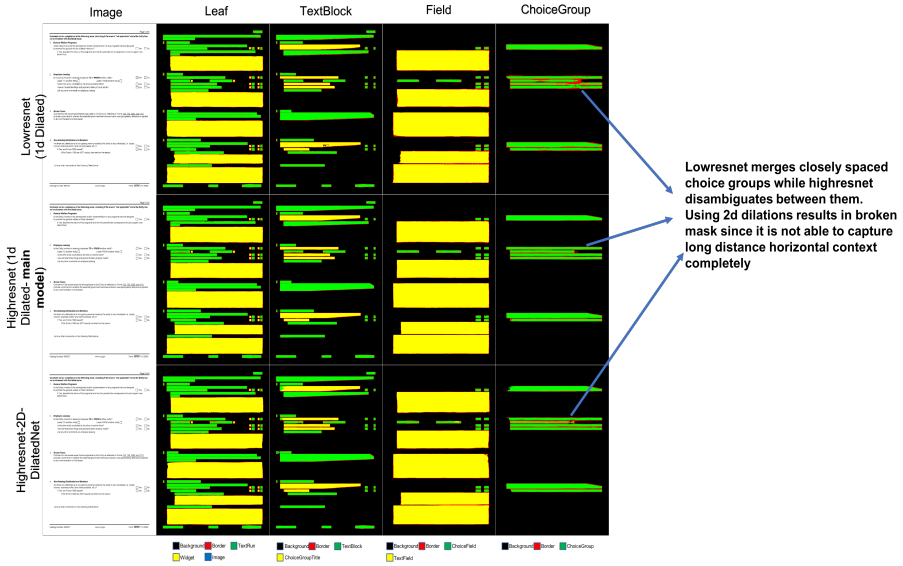

Fig. 3: Visualisation comparing predictions of lowresnet, 2D-DilatedNet(highresnet with 2d dilated convs) and Highresnet(with 1d dilated convs). As shown highresnet with 1d dilated conv gives better predictions. Zoom in to better view different parts of the form.

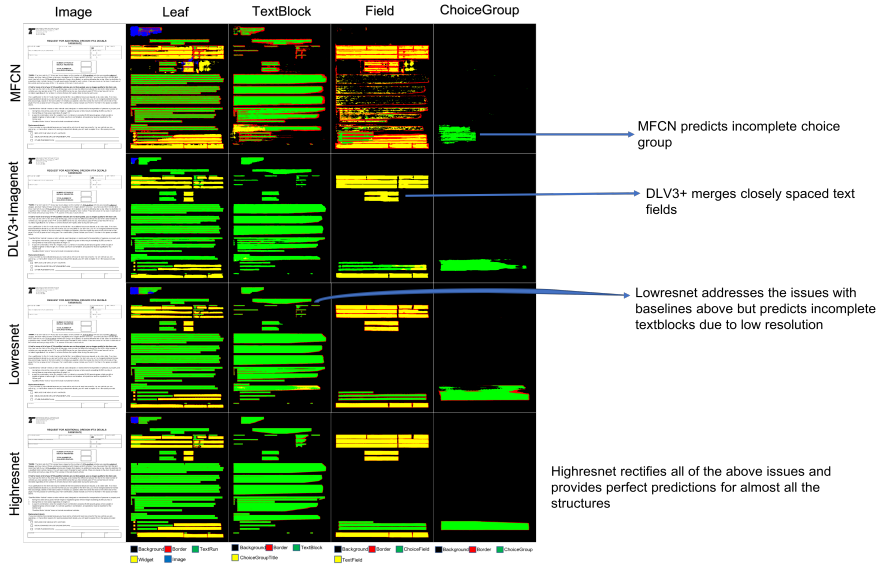

Fig. 4: Visualisation comparing predictions of the baseline methods MFCN, DLV3+(with Imagenet) with our Lowresnet and Highresnet. Zoom in to better view different parts of the form.