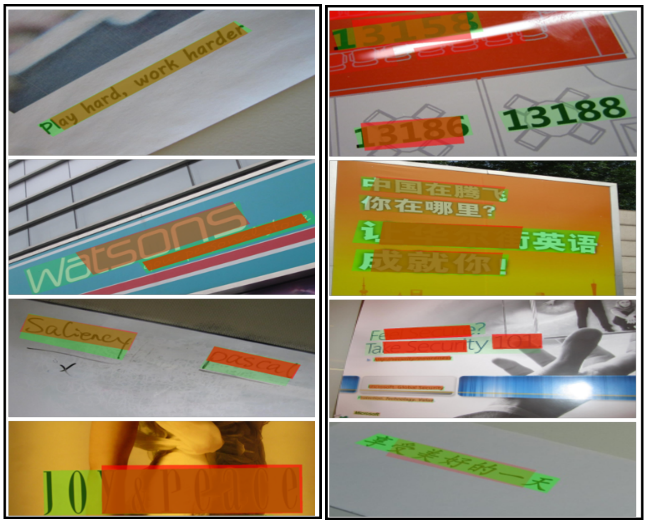
Appendix

Pseudo code of OM & MO solution. Algorithm 1 summarizes the joint Word&Text-Line annotation evaluating procedure. Note that not all details are covered by 1, which can be found on the source code.

Experiments on Non-Latin text. The experiments on the main paper only conducted on the word-level Latin datasets. Actually, it may be more effective for detecting long text lines. Thus we further supplemented an experiment on well-known MSRA-TD500 [29]. Because we cannot retrieve the others' previous detection results of this dataset, we train East [32] and an improved version of Mask R-CNN with additional data selected from RCTW-17 [26] to test the TIoU metric. The results are shown in Table 3, and some of the qualitative results are shown in Figure 9. Note that, during the evaluation, we calculate the exact value of the area of the polygon instead of using approximate area calculation in [29]. It can be seen from Table 3 that the recall, precision, and Hmean all drop significantly, which are mainly because there exists many defective detections shown in Figure 9. TIoU metric also highlights the difference between object detection and text detection task, showing there still exists a large room for detection methods to improve.

Table 3. Comparison of metrics on the TD500. i: IoU. t: TIoU.

Methods	R_i	P_i	F_i	R_t	P_t	F_t
East [32]	0.615	0.49	0.546	0.411	0.369	0.389
Mask R-CNN++ [5]	0.832	0.837	0.834	0.638	0.679	0.658



(a) Improved Mask R-CNN. (b) East.

Figure 9. Visualization results of MSRA-TD500. Green: Ground truth. Red: Detection result. Orange: Overlapping region. Note that previous metrics would regard all these detections with 100% recall and precision.

Algorithm 1 Joint Word&Text-Line Evaluation.

```
Input:
  $S$ - Dataset to be evaluated.
  $W$ - Word-Level annotations.
  $W_i$: the $i$-th annotation of $W$.
  $T$ - Text-Line annotations.
  $T_j$: the $j$-th annotation of $T$.
  $D$ - Detection results.
  $M$ - Matching indicator. All zero in the beginning.
Evaluation Procedure:
  (a) Creating $T$ from $W$ of $S$ manually.
    (i) $T$ ignores all "don't care" instances of $W$.
    (ii) Normally, each annotation of $T$ contains at least
      two instances of $W$.
  (b) Using "don't care" instances of $W$ to distinguish
  "don't care" detections of $D$.
    (i) If $\frac{Area(W_i \cap D_j)}{Area(D_j)} > 0.5$, $D_j$ is marked as "don't
      care".
  (c) Creating matching indexes between $T$ and $W$.
    (i) This step can be done during step (a). Deciding
      which $W_i$ belongs to which $T_j$.
    (ii) If not (i), then if $\frac{Area(T_j \cap W_i)}{Area(W_i)} > 0.5$, simply
      marking $W_i$ belongs to $T_j$.
  (d) Evaluating $D$ on $T$ in advance.
    If $M_{D_i}$ is zero and $D_i$ is not marked as "don't care':
      If IoU of $D_i$ and $T_j$ is $> 0.5$:
        Accumulating TIoU precision
        $M_{D_i} = 1$
        For $W_1$ ... $W_k$ that belong to $T_j$:
          If $\frac{Area(W_k \cap D_i)}{Area(W_k)} < 0.5$:
            leave $W_k$ to the step (f)
          else:
            Accumulating TIoU recall using Eq. 23
            Marking $W_k$ as "don't care"
        end for
      end if
    end if
  (e) Using new "don't care" instances of $W$ to distinguish "don't care" detections of $D$.
    (i) Same as step (b)
  (f) Evaluating $D$ on $W$.
    If $M_{D_i}$, $M_{W_j}$ are zero and $D_i$, $W_j \neq$ "don't care':
      If IoU of $D_i$ and $W_j$ is $> 0.5$:
        Accumulating TIoU precision
        Accumulating TIoU recall
        $M_{D_i} = 1$
        $M_{W_j} = 1$
      end if
    end if
Output:
  Final TIoU precision.
  Final TIoU Recall.
  Using Eq. 16 to calculate final TIoU Hmean.
```