Supplementary Material for "Masked Vision-Language Transformers for Scene Text Recognition"

Figure 1: Visualization of fine-tuned MVLT and MVLT*. For each example, we show the STR results of MVLT, MVLT*, and the ground-truth. The wrong predictions are shown in red. The images are from test datasets.

Figure 2: Visualization of pretrained MVLT and MVLT*. For each example, we show the masked image (left), the image reconstruction result of $decoder_2$ of the pretrained MVLT (mid-left), the image reconstruction result of $decoder_2$ of the pretrained MVLT* (mid-right), and the ground truth (right). The images are from test datasets.