Figure A1: Lexicon-free output predictions on non-alphanumeric-text images by the proposed Recursive Recurrent Nets with Attention Modeling (R\(^2\)AM) framework. By directly operating on images without alphanumeric characters, we can see our model produces output characters that are best fit to the underlying character-level language model implicitly learned from the training data.

Table A1: Left: network architectures for Base CNN and the proposed untied recursive CNNs. Right: the bar chart shows the corresponding performance for networks with different depths on Synth90k dataset. In this experiment we gradually increase the depth of the baseline CHAR model in [17] from 5 conv layers until we reach the performance plateau at 8 conv layers (denoted as Base CNN as our strong baseline). However, we can further boost the performance by using the proposed untied recursive CNNs. Notice that our recursive CNNs have the same number of parameters as Base CNN but achieve significantly better accuracy.