We propose an end-to-end approach for phrase grounding in images. Unlike prior methods that typically attempt to ground each phrase independently by buildi