Supplementary Material

Anonymous ACL submission

1 Qualitative Results

In this section, we present few success and failure cases for SHNet model on variety of image-expression pairs.

In Figure 1, we present results where SHNet successfully grounded the referring expression in the image. SHNet is able to identify fine grained distinctive information about the referent from the referring expression, and utilize it to correctly localize the referent in complex visual scenes in (c), (d), (f) and (j). Specifically in (c), (d) and (j), SHNet is able to identify the correct person from large group of people based on the combination of person's attribute ("dark hair"), attributes of person's clothing ("green sleeves", "no shirt" etc) and its location with respect to other objects in the image ("by the wall"). Additionally, SHNet localizes objects which are out of focus and are partially visible, ex: (b), (e), (g) and (h). We would like to point out that in these cases, rather than merely picking the most prominent objects, our network effectively incorporates the information from textual expression in visual domain to identify the less prominent correct object. In (a) and (i), the referring expressions refer to unstructured regions in image, our network predicts these regions with refined boundaries. In (k) and (l) of Figure 1, the referred objects occupy extremely small region in the image space and SHNet is able to accurately locate them.

In Figure 2, we present some failure cases of our approach. Our approach mostly fails in cases when either the referring expression or the visual scene is ambiguous in (a), (c) and (e), the visual scene is heavily cluttered in (b) and (d), or when common sense reasoning is required like (f). For example: the expression in (a), "chair at the end of table on the left" is itself ambiguous and non-specific, as there are two chairs at the end of table on left side. Similarly, in (b) their are multiple keyboards with a mouse on top and our method predicts one of the keyboards on the left with a partial black mouse on the top. In (d), the plant branch on the left is barely visible and also a lot of clutter is present. It is noteworthy, that in each case, SHNet predicts a well segmented and refined output and the class predictions are also correct (an umbrella, a chair, a bottle, a keyboard etc.).

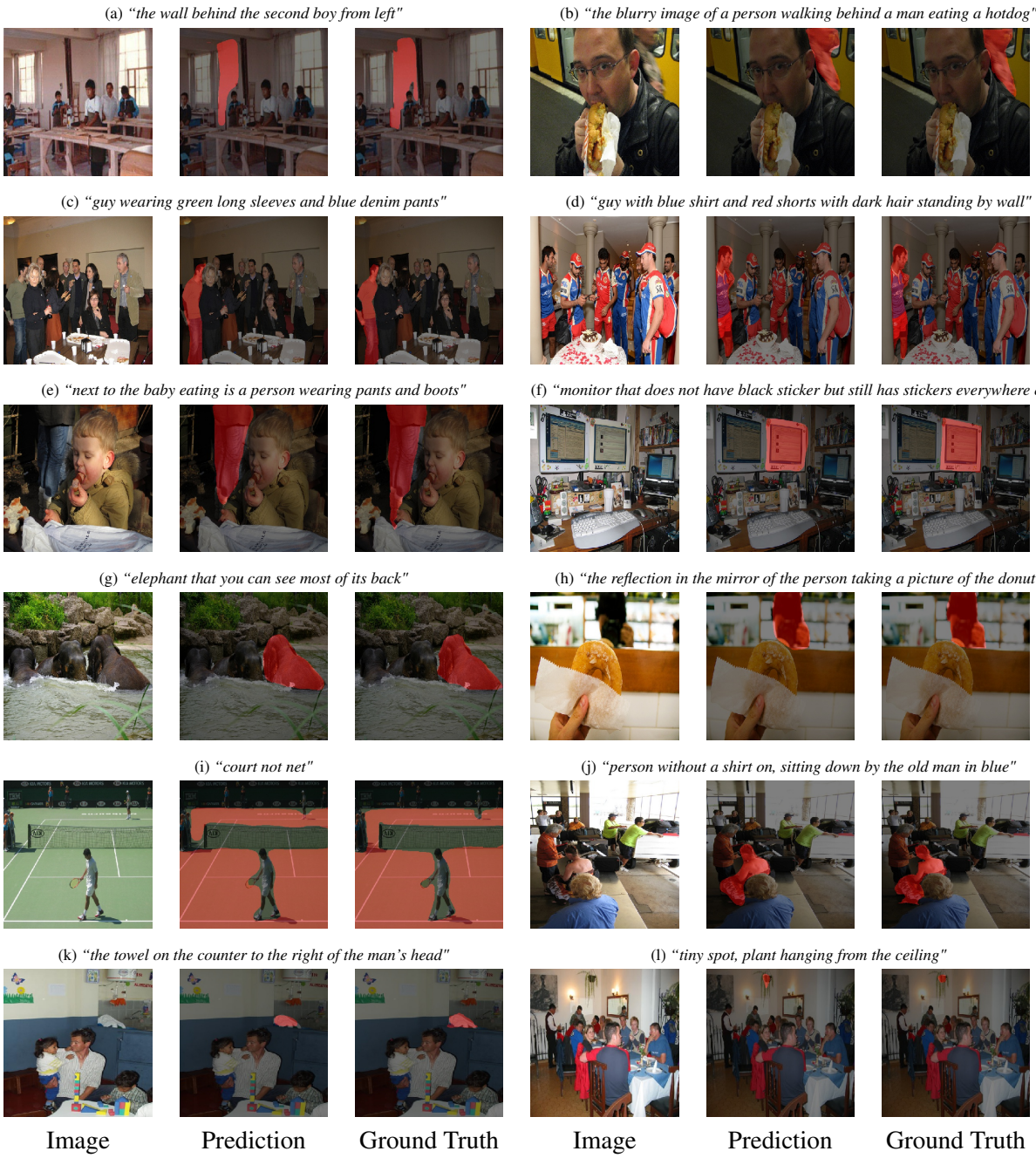

Figure 1: Qualitative examples where SHNet successfully localized the referred object.

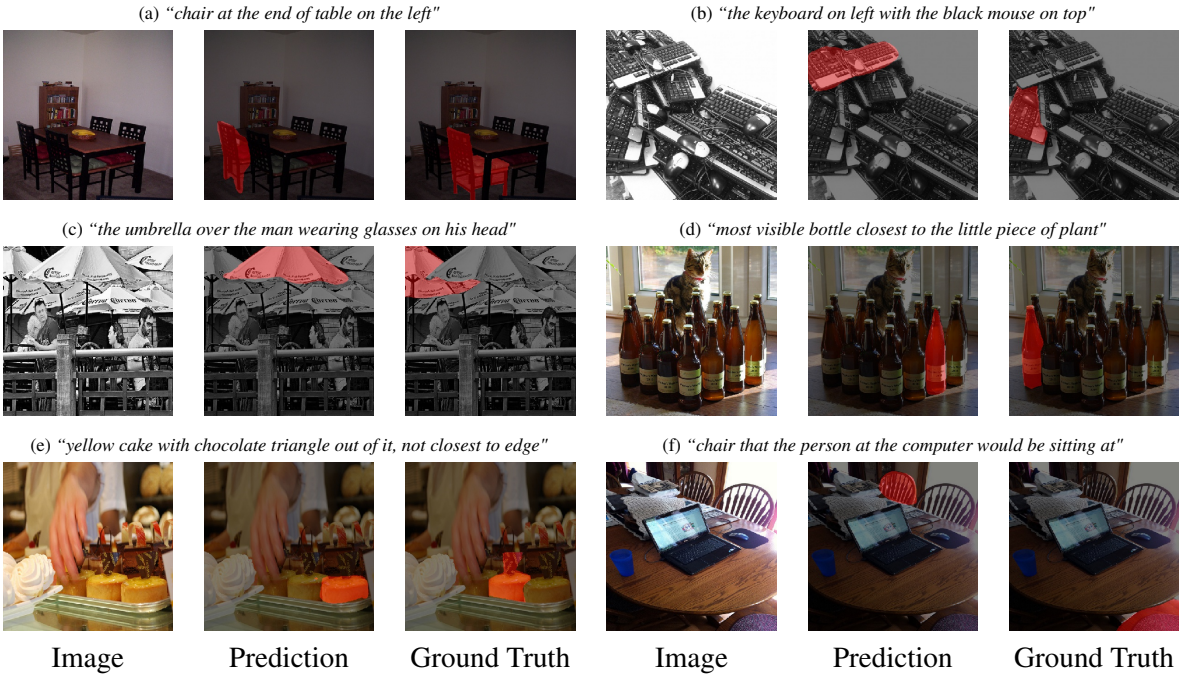

Figure 2: Qualitative examples where SHNet failed to localize the referred object.