Supplementary Materials for "Relax Image-Specific Prompt Requirement in SAM: A Single Generic Prompt for Segmenting Camouflaged Objects"

Anonymous submission

Algorithm

Algorithm 1: Algorithm of our GenSAM.

```
Input: Sample $x_i^t$ from target Domain $D_t = \{x_i^t\}^{n_t}$; a
       image-to-caption model BLIP2, our devised
       spatial CLIP and segmentation model SAM
       and a generic task prompt $P_g$.
Output: Segmentation result $mask_{\text{iter}^*}$.
for $iter$ = 1 to Iter do
    Feed $x_i^t$ into BLIP2 to generate image caption $C_i$,
      transfer prompt $P_g$ into three chains of thoughts
      $Q_i^1$ to $Q_i^3$.
    for $j$ = 1 to 3 do
        Image $x_i^t$, caption $C_i$, the $j$-th chain of
          thought $Q_i^j$ are fed into BLIP2 together to
          get foreground and background keyword
          $fk_i^j$ and $bk_i^j$ respectively. Then $fk_i^j$ and
          $bk_i^j$ are fed into spatial CLIP to get
          foreground and background heatmap
          $F_{fk}^j(x_i^t)$ and $F_{bk}^j(x_i^t)$.
    end
    Heatmaps generated for foreground and
      background from various chains of thought are
      averaged respectively to get $F_{fk}(x_i^t)$ and
      $F_{bk}(x_i^t)$. Then, final heatmap in the $iter$-th iter
      is $H_i$. the points on $H_i$ that are higher than
      threshold are selected as image-specific visual
      prompt to guide SAM to segment $x_i^t$, and get
      $mask_{iter}$. $H_i$ is seen as a visual prompt to
      weighting on $x_i$ in the next iteration
end
The selected iteration iter$^*$ is determined by selecting
  the iteration's result that closely resembles the
  average mask across all iterations, and the final
  segmentation mask is $mask_{\text{iter}^*}$.
```

Difference between GenSAM and ClIP Surgery

CLIP surgery is a method used to generate corresponding visual results by inputting specific object names. However, our GenSAM not only generates precise localization for the corresponding objects but also guides SAM to perform effective segmentation. Our GenSAM incorporates a k-k-v self-attention mechanism, which effectively replaces the v-v-v self-attention mechanism used in CLIP Surgery. Additionally, it utilizes foreground and background consensus methods to obtain heatmaps for foreground and background separately, instead of representing the background heatmap through the cancellation of corresponding heatmaps with empty statements. This approach results in improved localization performance.

More Visualizations of GenSAM

The visualization result of COD10K is shown in Fig. 1. For each sample, the first row shows the weighted image in each iteration, the second row shows the consensus heatmap and the third row shows the segmentation results. The red and blue points in the first and third rows respectively show the positive and negative point prompts. Furthermore, the visualization result of dataset CHAMELEON and CAMO is shown is Fig. 2.

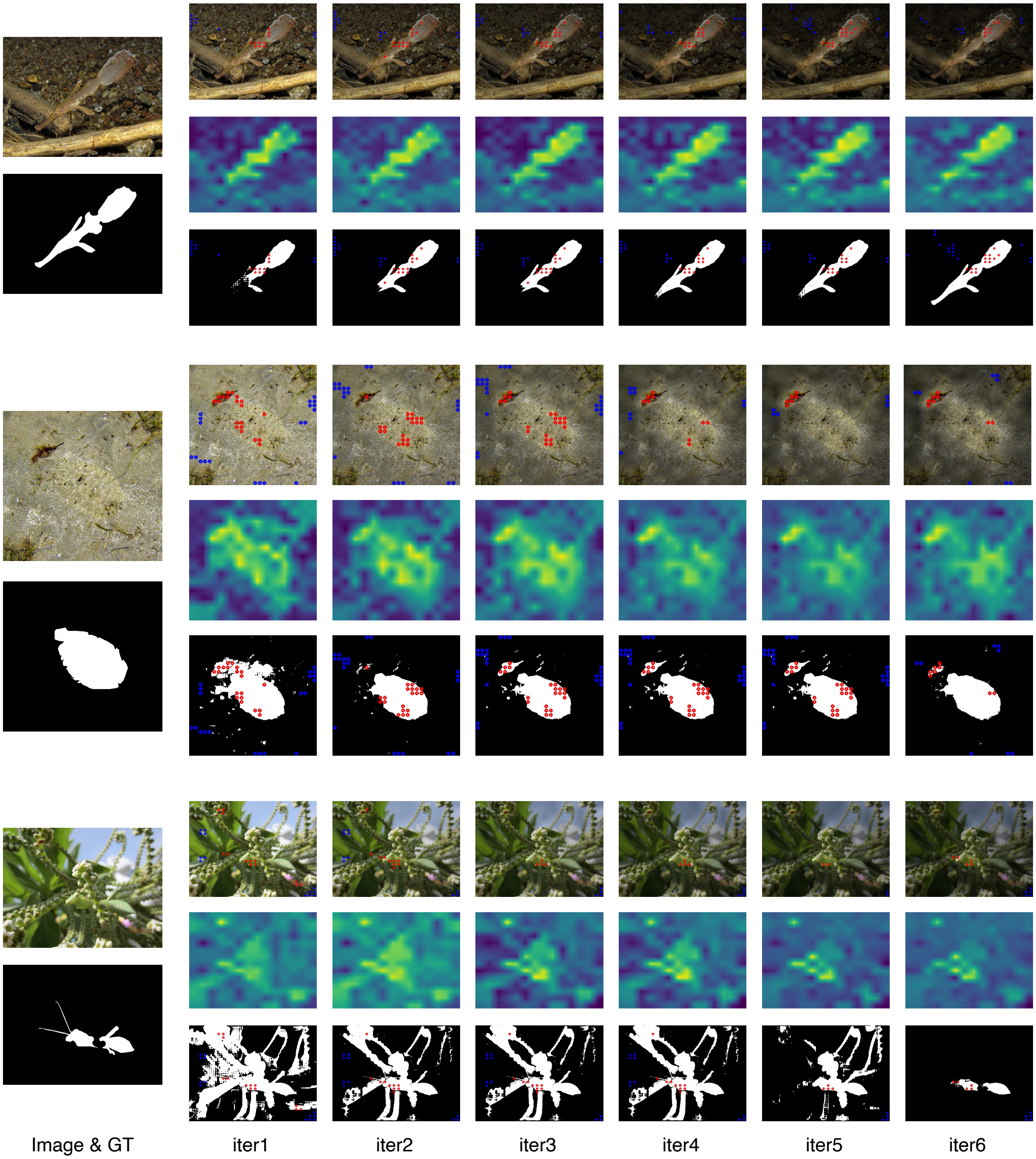

Figure 1: More visualizations on COD10K.

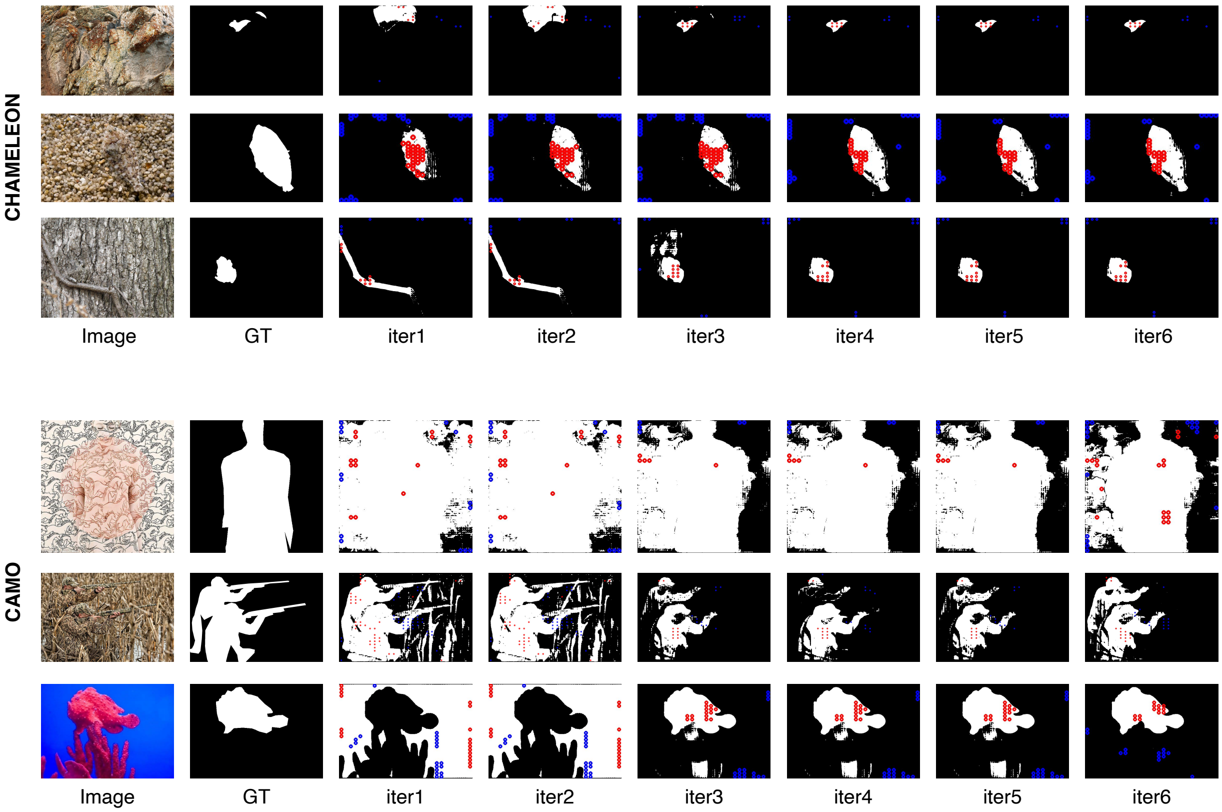

Figure 2: More visualizations on CHAMELEON and CAMO datasets.