Few-Shot Learning via Saliency-guided Hallucination of Samples
(Supplementary Material)

Hongguang Zhang\textsuperscript{1,2} Jing Zhang\textsuperscript{1,2} Piotr Koniusz\textsuperscript{2,1}
\textsuperscript{1}Australian National University, \textsuperscript{2}Data61/CSIRO
firstname.lastname@{anu.edu.au\textsuperscript{1}, data61.csiro.au\textsuperscript{2}}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{salience_maps.png}
\caption{Examples of saliency maps on the Open MIC dataset. The MNL detector was used.}
\end{figure}

1. Saliency Maps on the Open MIC dataset

In Figure 1, we present saliency maps for some exhibit instances from the Open MIC dataset. Many exhibits can be filtered out reliably. However, saliency maps for composite scenes containing numerous exhibits are the ones most likely to fail. In the future, we will investigate how to improve the use of such unreliable saliency maps for such exhibits. Note that our results on exhibitions containing such composite scenes still benefit from our approach—our mixing network can reduce the noise from saliency maps.

2. Evaluations for 224 × 224 pixel images

We employ 84 × 84 image in our experiments for fair comparison with other state-of-the-art methods presented in our paper. However, it is easy to use large size images in our network without its modifications due to the ability of second-order representations to aggregate variable number of feature vectors into a fixed-size matrix (our relationship descriptors are stacked matrices). Here we apply 224 × 224 image to demonstrate the benefits from larger image size.


Below we present the diagrams of two baseline networks used in our paper. The baseline 1 in Figure 2 is the original pipeline ‘w/o Sal. Seg.’, which is trained without saliency segmentation or data hallucination—it is very similar to the SoSN pipeline \textsuperscript{7}. Figure 3 demonstrates the baseline 2 ‘w/o Hal.’, which employs saliency network to segment the foregrounds and backgrounds but does not hallucinate the data (no mixing of a foreground with numerous different backgrounds is allowed).

In our paper, the reported results are obtained by using baseline 2 ‘w/o Hal.’ pipeline as teacher in TriR regulariza-

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
Model & Fine Tune & 5-way Acc. & 5-shot \\
\hline
Matching Nets \textsuperscript{[6]} & N & 43.56 ± 0.84 & 55.31 ± 0.73 \\
Meta Nets \textsuperscript{[2]} & N & 49.21 ± 0.96 & - \\
Meta-Learn Nets \textsuperscript{[3]} & N & 43.44 ± 0.77 & 60.60 ± 0.71 \\
Prototypical Net \textsuperscript{[4]} & N & 49.42 ± 0.78 & 68.20 ± 0.66 \\
MAML \textsuperscript{[1]} & Y & 48.70 ± 1.84 & 63.11 ± 0.92 \\
Relation Net \textsuperscript{[5]} & N & 51.36 ± 0.86 & 65.63 ± 0.72 \\
SoSN \textsuperscript{[7]} & N & 52.96 ± 0.83 & 68.63 ± 0.68 \\
\hline
\end{tabular}
\caption{Accuracy on the miniImageNet dataset given different size of images. See \textsuperscript{[5, 7]} for details of baselines. The asterisk (*) denotes the ‘sanity check’ results on our proposed pipeline given disabled both saliency segmentation and hallucination.}
\end{table}
**Figure 2:** The network architecture of baseline 1 'w/o Sal. Seg.'. It can be seen that once the Saliency Net and data hallucination strategies are disabled, the network pipeline are very similar to SoSN. Note that we write $R_1, \ldots, R_N$ for brevity rather than $R_{11}, \ldots, R_{NN}$ (as dictated by Eq. (4) of our main submission) as no hallucination takes place here e.g., we evaluate only $R_{ij}$ for $i = j$. Moreover, note that $\Phi_i$ are not generated by the foreground-background mechanism from Eq. (3) of our main submission. Instead, entire images are encoded.

**Figure 3:** The network architecture of baseline 2 'w/o Hallucination'. Note that although the data hallucination mechanism is disabled, we still apply saliency maps to segment foregrounds and backgrounds as we want the TriR loss to learn to account for the potential noise stemming from the foreground-background segmentation which is used in the main network. Moreover, we write $R_1, \ldots, R_N$ for brevity rather than $R_{11}, \ldots, R_{NN}$ (as dictated by Eq. (4) of our main submission) as no hallucination takes place here e.g., we evaluate only $R_{ij}$ for $i = j$. Note that $\Phi_i$ are generated by the foreground-background mechanism in Eq. (3) of our main paper (we abbreviate $\Phi_{ii}$ to $\Phi_i$).

**Table 2:** Evaluations on the miniImagenet dataset given different teacher networks for the TriR regularization.

<table>
<thead>
<tr>
<th>Model</th>
<th>Fine Tune</th>
<th>5-way Acc. 1-shot</th>
<th>5-way Acc. 5-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline 1 (opt. (i)) as a teacher network in TriR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SalNet Intra-class Hal.</td>
<td>N</td>
<td>36.11 ± 0.86</td>
<td>71.36 ± 0.67</td>
</tr>
<tr>
<td>SalNet Inter-class Hal.</td>
<td>N</td>
<td>57.24 ± 0.94</td>
<td>72.49 ± 0.65</td>
</tr>
<tr>
<td>baseline 2 (opt. (ii)) as a teacher network in TriR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SalNet Intra-class Hal.</td>
<td>N</td>
<td>36.57 ± 0.86</td>
<td>71.78 ± 0.69</td>
</tr>
<tr>
<td>SalNet Inter-class Hal.</td>
<td>N</td>
<td>57.45 ± 0.88</td>
<td>72.01 ± 0.67</td>
</tr>
</tbody>
</table>

**References**


