Methodology....

Fusion Sub-network

- Frequency domain sub-network (physical level): l_0
- Pixel domain sub-network (semantic level): $\{l_1, l_2, l_3, l_4\}$
- via attention mechanism, and the enhanced image representation is computed as follows:

  - $F(l_i) = v^T \tanh(W_f l_i + b_f), i \in [0,4]$, $\alpha_i = \frac{\exp(F(l_i))}{\sum_i \exp(F(l_i))}$, $u = \sum_i \alpha_i l_i$

Methodology.....

Fusion Sub-network

- $F(l_i) = v^T \tanh(W_f l_i + b_f), i \in [0,4]$, $\alpha_i = \frac{\exp(F(l_i))}{\sum_i \exp(F(l_i))}$, $u = \sum_i \alpha_i l_i$
  - W_f: weight matrix, b_f: bias term, v^T: transposed weight vector
  - F: score function, a_i: normalized weight of i-th feature vector
- Obtained high-level representation of image u at both physical and semantic levels.
- Use a fully-connected layer with softmax activation to project to two classes, gain the **probability distribution:** $p = \text{softmax}(W_c u + b_c)$