Appendix

Anonymous submission

A Explanation of Dynamic Masks

Setting the node corresponding to the category inferred by the neural network as the activation position p, the dynamic masks (DM) can identify the regions of an image that a neural network focuses on the most when making a classification decision. The mathematical proof is as follows.

Define: z is the region in input image x, and $f_p(z)$ is the activation value of the neural networks f at position p when the data of region z is input to network f. $I(z) = kf_p(z)$, where k is a constant, and $k > 0$. $I(z)$ is the amount of information that region z contributes to the activation of neural networks f at position p, $I(z) \in [0, 1]$. The mask m is trained by optimizing the following function L:

$$L(m, z) = [f_p(z) - f_p(mz)]^2 + \eta m \tag{1}$$

where z is all the regions of d_i, and m is the corresponding mask value on it, $m \in [0, 1]$.

There are two public cognition. When the corresponding regions on the input image do not intersect, it is considered that information I of the contribution of the two regions to activation f_p is irrelevant. Additionally, the greater contribution of the investigation region to the activation implies a greater the contribution to the information increment. Mathematically, z_1 and z_2 are the two regions of d_i, $i \in \{1, 2, ..., N\}$, and g is the upsampling function, which upsamples to the size of the input image.

If $g(z_1) \cap g(z_2) = \emptyset$, then

$$I(z_1 + z_2) = I(z_1) + I(z_2) \tag{2}$$

if $I(z_1) < I(z_2)$, then

$$0 \leq \frac{\partial I(mz_1)}{\partial m} < \frac{\partial I(mz_2)}{\partial m} \tag{3}$$

Let: z_1 and z_2 are any two disjoint regions of d_i; m_1, m_2 are the mask values on z_1, z_2. From Equations (2) and (3), the following Equation (4) can be proved, when $L(m, z)$ in Equation (1) achieves the minimum value.

$$(I(z_1) - I(z_2))(m_1 - m_2) \geq 0 \tag{4}$$

Reductio ad absurdum. If $L(m, z)$ in Equation (1) has achieved the minimum value, and $\exists z_1$, z_2 satisfy:

$$(I(z_1) - I(z_2))(m_1 - m_2) < 0 \tag{5}$$

Let: $z(d_i)$ is all areas on d_i, $z_0 = z(d_i) - z_1 - z_2$, and m_0 is the mask value of z_0. $g(z_1) \cap g(z_2) = \emptyset$, $g(z_1) \cap g(z_0) = \emptyset$, $g(z_2) \cap g(z_0) = \emptyset$. Due to symmetry, it may be assumed that $I(z_1) < I(z_2)$. From Equations (3) and (5), it can be inferred that $\frac{\partial I(mz_1)}{\partial m} < \frac{\partial I(mz_2)}{\partial m}$ and $m_1 > m_2$.

$$\begin{aligned}
&L(m, z) = L(z_1, m_1, z_2, m_2, z_0, m_0) \\
&= [f_p(z_1 + z_2 + z_0) - f_p(m_1 z_1 + m_2 z_2 + m_0 z_0)]^2 \\
&+ \eta(m_1 + m_2 + m_0)
\end{aligned} \tag{6}$$

$$\begin{aligned}
&L^{'}(m, z) = L(z_1, m_2, z_2, m_1, z_0, m_0) \\
&= [f_p(z_1 + z_2 + z_0) - f_p(m_2 z_1 + m_1 z_2 + m_0 z_0)]^2 \\
&+ \eta(m_2 + m_1 + m_0)
\end{aligned} \tag{7}$$

$$\begin{aligned}
&L^{'}(m, z) - L(m, z) \\
&= [2f_p(z_1 + z_2 + z_0) - f_p(m_1 z_1 + m_2 z_2 + m_0 z_0) - \\
&f_p(m_2 z_1 + m_1 z_2 + m_0 z_0)][f_p(m_1 z_1 + m_2 z_2 + m_0 z_0) \\
&- f_p(m_2 z_1 + m_1 z_2 + m_0 z_0)] \\
&= \{[I(m_1 z_1) - I(m_2 z_1)] - [I(m_1 z_2) - I(m_2 z_2)]\}\{[I(z_1) \\
&- I(m_1 z_1)] + [I(z_2) - I(m_1 z_2)] + [I(z_1) - I(m_2 z_1)] \\
&+ [I(z_2) - I(m_2 z_2)] + 2[I(z_0) - I(m_0 z_0)]\ \}/k^2 \\
&= \left\{\int_{m_2}^{m_1} [\frac{\partial I(mz_1)}{\partial m} - \frac{\partial I(mz_2)}{\partial m}] dm\right\} [2\int_{m_0}^{1} \frac{\partial I(mz_0)}{k^2 \partial m} dm \\
&+ \int_{m_1}^{1} \frac{\partial I(mz_1) + \partial I(mz_2)}{k^2 \partial m} dm \\
&+ \int_{m_2}^{1} \frac{\partial I(mz_1) + \partial I(mz_2)}{k^2 \partial m} dm] < 0
\end{aligned} \tag{8}$$

$L^{'}(m, z) < L(m, z)$, which contradicts that L has achieved a minimum. Therefore, Equation (4) holds.

As shown in Equation (4), optimization Equation (1) can make the mask achieve the following properties. Regions with higher decision contributions have higher mask values, which results in more image information being retained. Conversely, regions with lower decision contributions have lower mask values and result in less image information being retained. Therefore, the DM can analyze the importance of each pixel in the image to the classification of the neural network.

Algorithm 1 Dynamic Masks Learning

Input: Image X_0, Neural Network $f(x)$, Activation Position p, Upsampling Function $g(x)$, Loss Function L, Benchmark Vectors $\{d_i\}_{i=1}^{N}$, Auxiliary Vectors $\{c_j^k(i)\}_{i=1,j=1,k=0}^{N,T,K_j^i}$.
Output: Saliency Maps M_b.
Parameter: Weights $\{\lambda_i\}_{i=1}^{N}$, $\{\lambda_i^{(k,j)}\}_{i=1,j=1,k=0}^{N,T,K_j^i}$, Training Epochs C, $\{C_i^{(k,j)}\}_{i=1,j=1,k=0}^{N,T,K_j^i}$, Threshold γ, Learning Rate η, $\{\eta_i^{(k,j)}\}_{i=1,j=1,k=0}^{N,T,K_j^i}$.

$A \leftarrow f_p(X_0)$
for $i = 1$ **to** N **do**
Initialize d_i each element is 0.5
for $j = 1$ **to** C **do**
$M_i \leftarrow g(d_i)$
$A_i \leftarrow f_p(M_i \cdot X_0)$
$L_c \leftarrow L(A, A_i)$
$L_d \leftarrow ||d_i||_1$
$L_t \leftarrow L_c + \lambda_i L_d$
$\theta_{d_i} \leftarrow \theta_{d_i} - \eta \frac{\partial L_t}{\partial \theta_{d_i}}$
end for
for $j = 1$ **to** T **do**
$c_j^0(i) \leftarrow d_i$
$A_i^{(0,j)} \leftarrow f_p(g(d_i) \cdot X_0)$
for $k = 1$ **to** K_j^i **do**
for $s = 1$ **to** $C_i^{(k,j)}$ **do**
$M_i^{(k,j)} \leftarrow g(c_j^k(i) \cdot g(c_j^{k-1}(i)))$
$A_i^{(k,j)} \leftarrow f_p(M_i^{(k,j)} \cdot X_0)$
$L_c^{(k,j)} \leftarrow L(A_i^{(k-1,j)}, A_i^{(k,j)})$
$L_d^{(k,j)} \leftarrow ||c_j^k(i)||_1$
$L_t^{(k,j)} \leftarrow L_c^{(k,j)} + \lambda_i^{(k,j)} L_d^{(k,j)}$
$\theta_{c_j^k(i)} \leftarrow \theta_{c_j^k(i)} - \eta_i^{(k,j)} \frac{\partial L_t^{(k,j)}}{\partial \theta_{c_j^k(i)}}$
end for
end for
end for
end for
Initialize M_c to zero mask
for $i = 1$ **to** N **do**
for $j = 1$ **to** T **do**
for $k = 0$ **to** K_j^i **do**
$M_c \leftarrow M_c + g(c_j^k(i))$
end for
end for
end for
$M_b \leftarrow (M_c - \gamma) \cdot \{M_c \geq \gamma\}$
Normalize M_b
return M_b

B Pseudo Code

In order to give a description of hierarchical dynamic masks (HDM), the pseudo codes for the workflow of the DM and the scheme for generating and combining the hierarchical masks of HDM are shown in Algorithms 1 and 2, respectively.

Algorithm 2 Hierarchical Generation and Combination

Input: Image X_0, Dynamic Masks $Q(x)$, Dynamic Masks Number S, Training Epochs C, Iterations N, Weight Parameters $\{v_i\}_{i=1}^{S}$, Neural Network $f(x)$, Activation Position p.
Output: Mix Saliency Maps M_h.
Parameter: Weight λ, Learning Rate η.

$A \leftarrow f_p(X_0)$
for $i = 1$ **to** S **do**
$M_i \leftarrow Q(X_{i-1})$
Initialize M_c to zero mask
for $j = 1$ **to** i **do**
$M_c \leftarrow M_c + M_j$
end for
Normalize M_c
$X_i \leftarrow (1 - M_c) \cdot X_0$
end for
Initialize M_h to zero mask
for $j = 1$ **to** C **do**
Initialize M_s to zero mask
$W \leftarrow 0$
for $i = 1$ **to** S **do**
$w_i \leftarrow 0$
for $j = i$ **to** S **do**
$w_i \leftarrow w_i + v_j^2$
end for
$W \leftarrow W + w_i$
$M_s \leftarrow M_s + w_i M_i$
end for
$M_s \leftarrow M_s / W$
$A_s \leftarrow f_p(M_s \cdot X_0)$
$L_c \leftarrow L(A, A_s)$
$L_d \leftarrow ||M_s||_1$
$L_t \leftarrow L_c + \lambda L_d$
for $i = 1$ **to** S **do**
$\theta_{v_i} \leftarrow \theta_{v_i} - \eta \frac{\partial L_t}{\partial \theta_{v_i}}$
end for
$M_h \leftarrow M_s$
end for
return M_h

C Visualization

In this section, we provide more visualization of saliency maps for HDM. We compare the visualization of the saliency maps of many methods on CUB-200-2011 [Wah *et al.*, 2011] and iChallenge-PM [Fu *et al.*, 2019], as shown in Figures 1, 2, 3, and 4. We show the visualization of HDH's saliency maps for CUB-200-2011 and iChallenge-PM in Figures 5 and 6.

References

[Fu *et al.*, 2019] Huazhu Fu, Fei Li, José Ignacio Orlando, Hrvoje Bogunovic, Xu Sun, Jingan Liao, Yanwu Xu, Shaochong Zhang, and Xiulan Zhang. Palm: Pathologic myopia challenge. *IEEE Dataport*, 2019.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

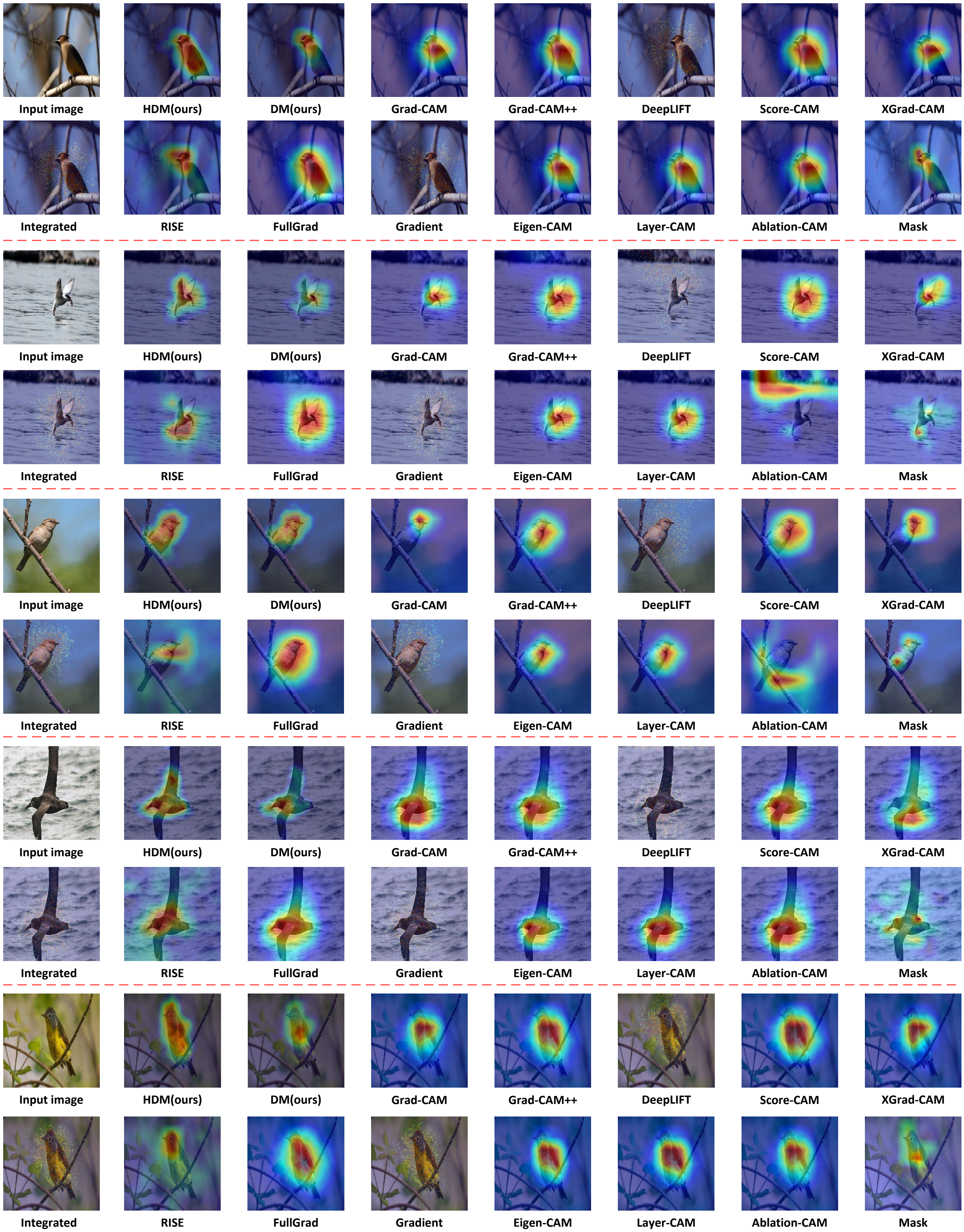

Figure 1: Visual comparison results of saliency maps for the bird images of the CUB-200-2011 dataset.

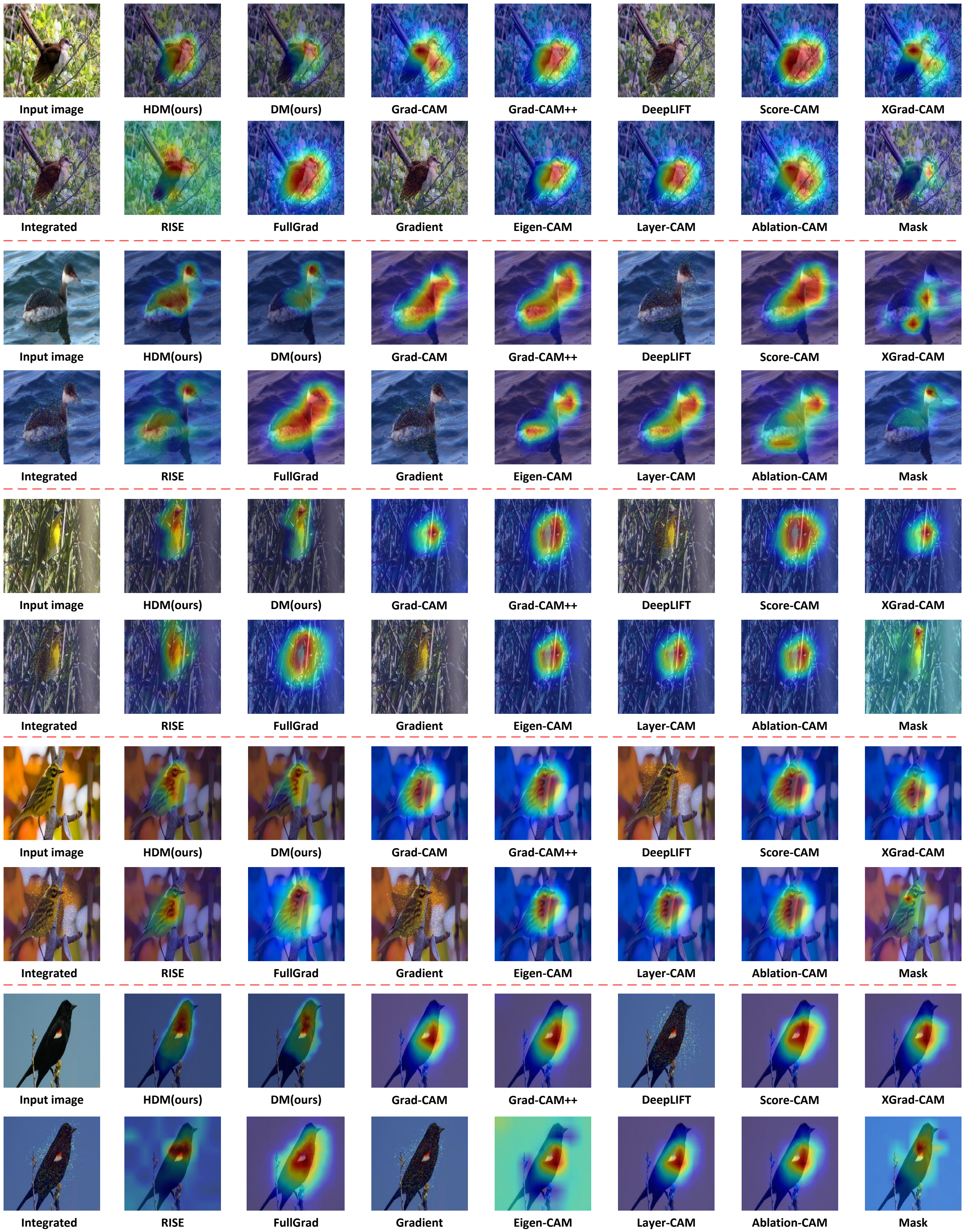

Figure 2: Visual comparison results of saliency maps for the bird images of the CUB-200-2011 dataset.

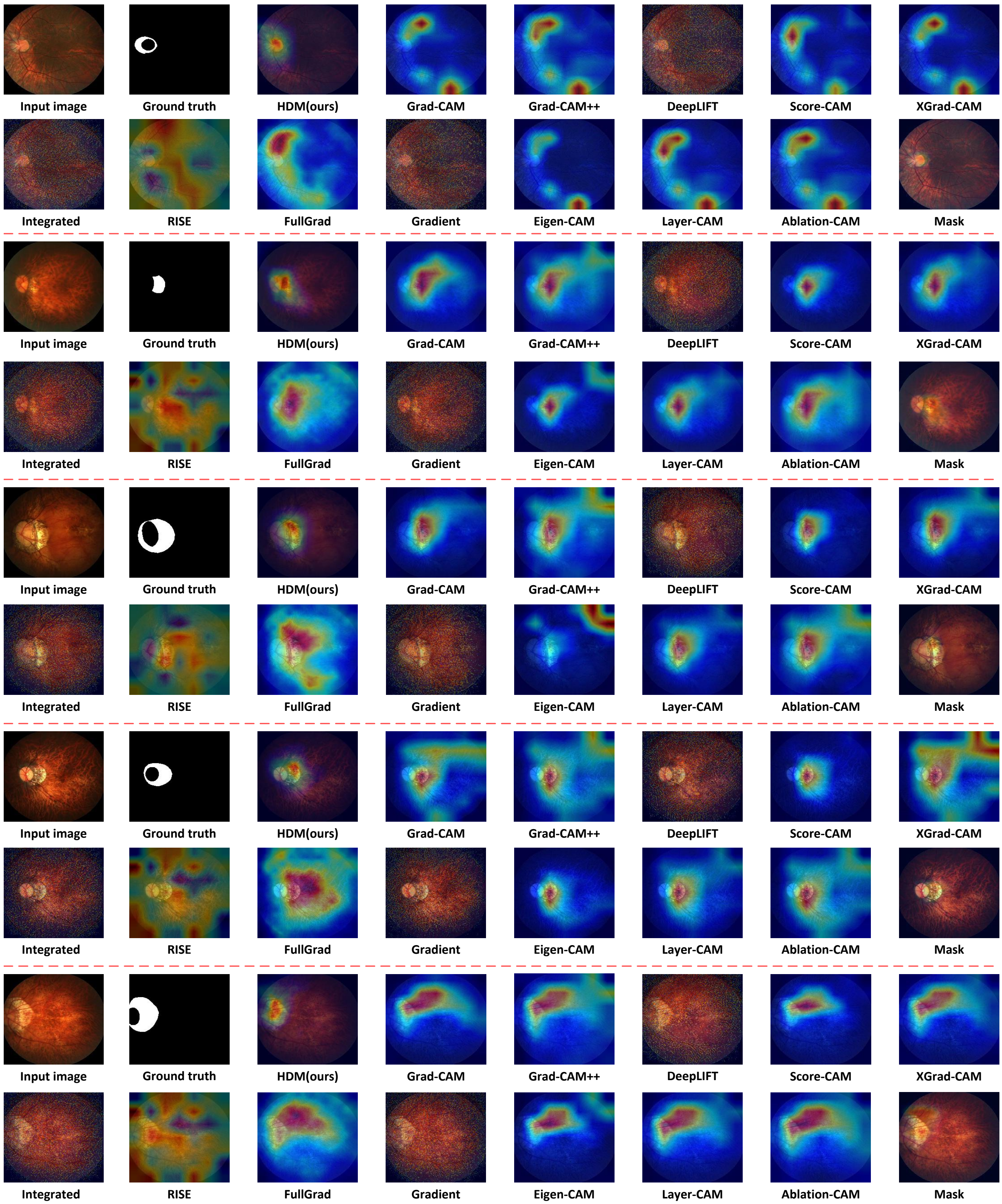

Figure 3: Visual comparison results of saliency maps for the fundus retina images of the iChallenge-PM dataset.

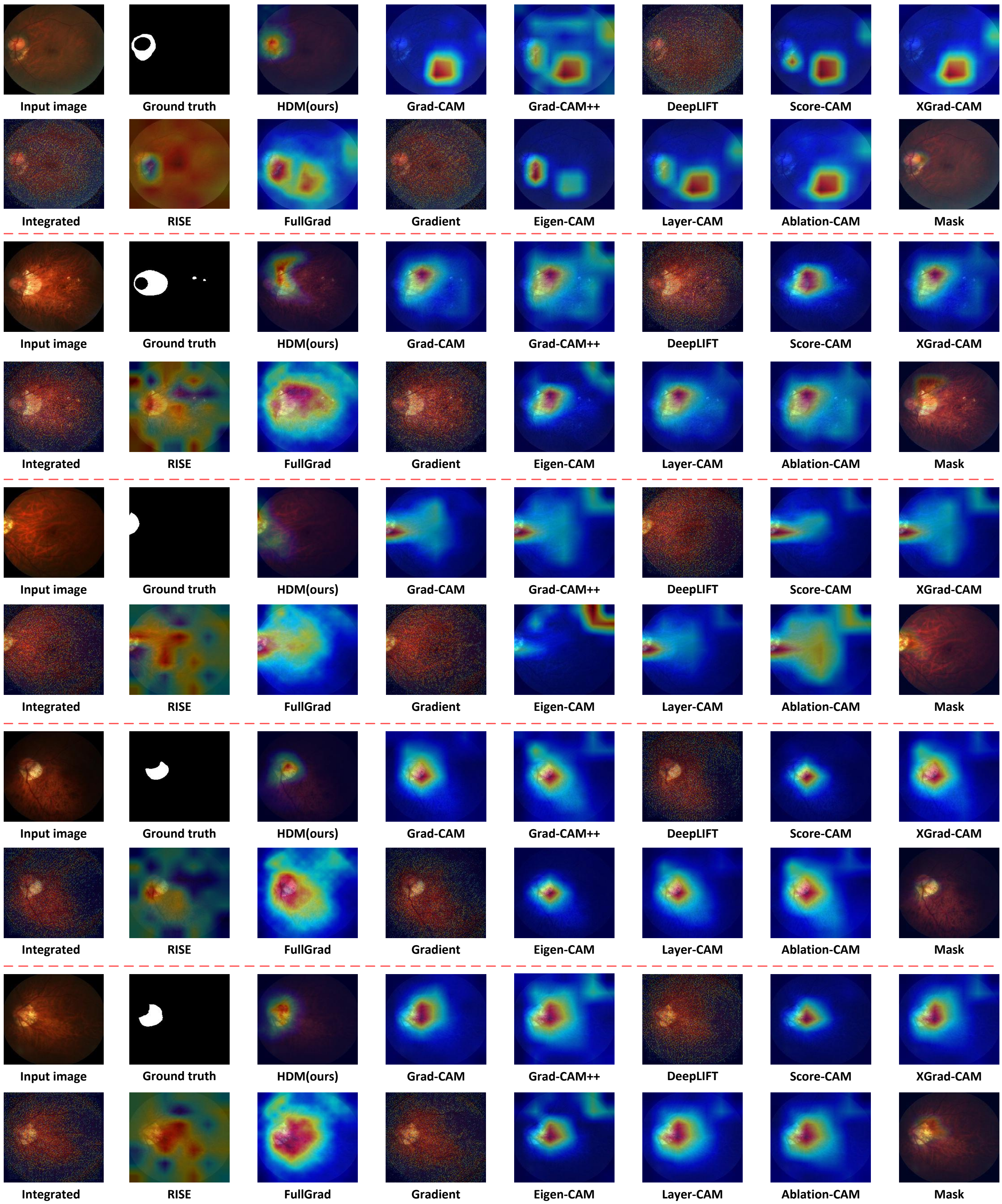

Figure 4: Visual comparison results of saliency maps for the fundus retina images of the iChallenge-PM dataset.

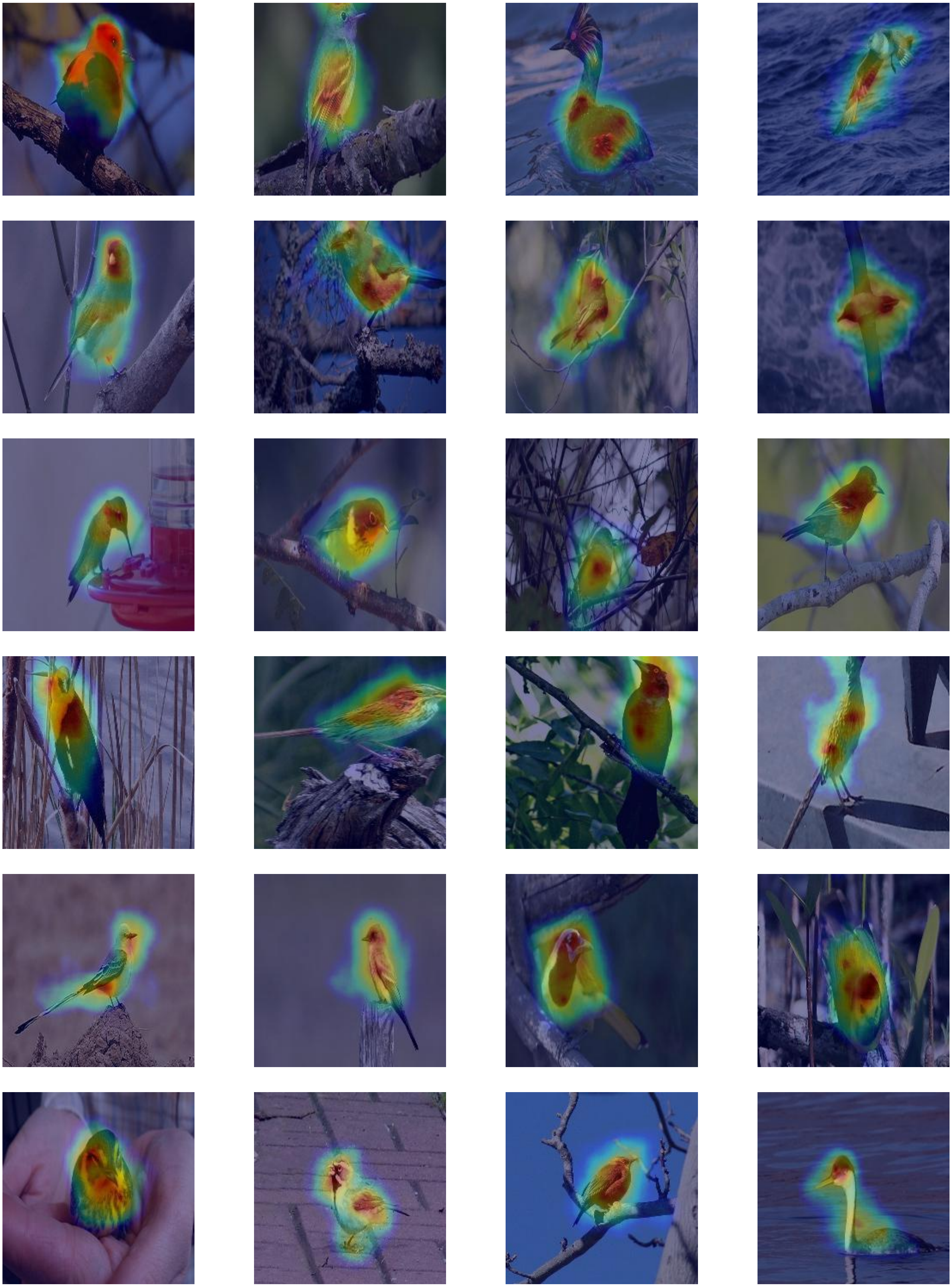

Figure 5: Visualization of HDM's saliency maps in CUB-200-2011 images. All saliency maps are normalized to range [0,1] and visualized using JET colormap. The saliency map's color gradient from blue to red indicates the neural network's increasing attention probability.

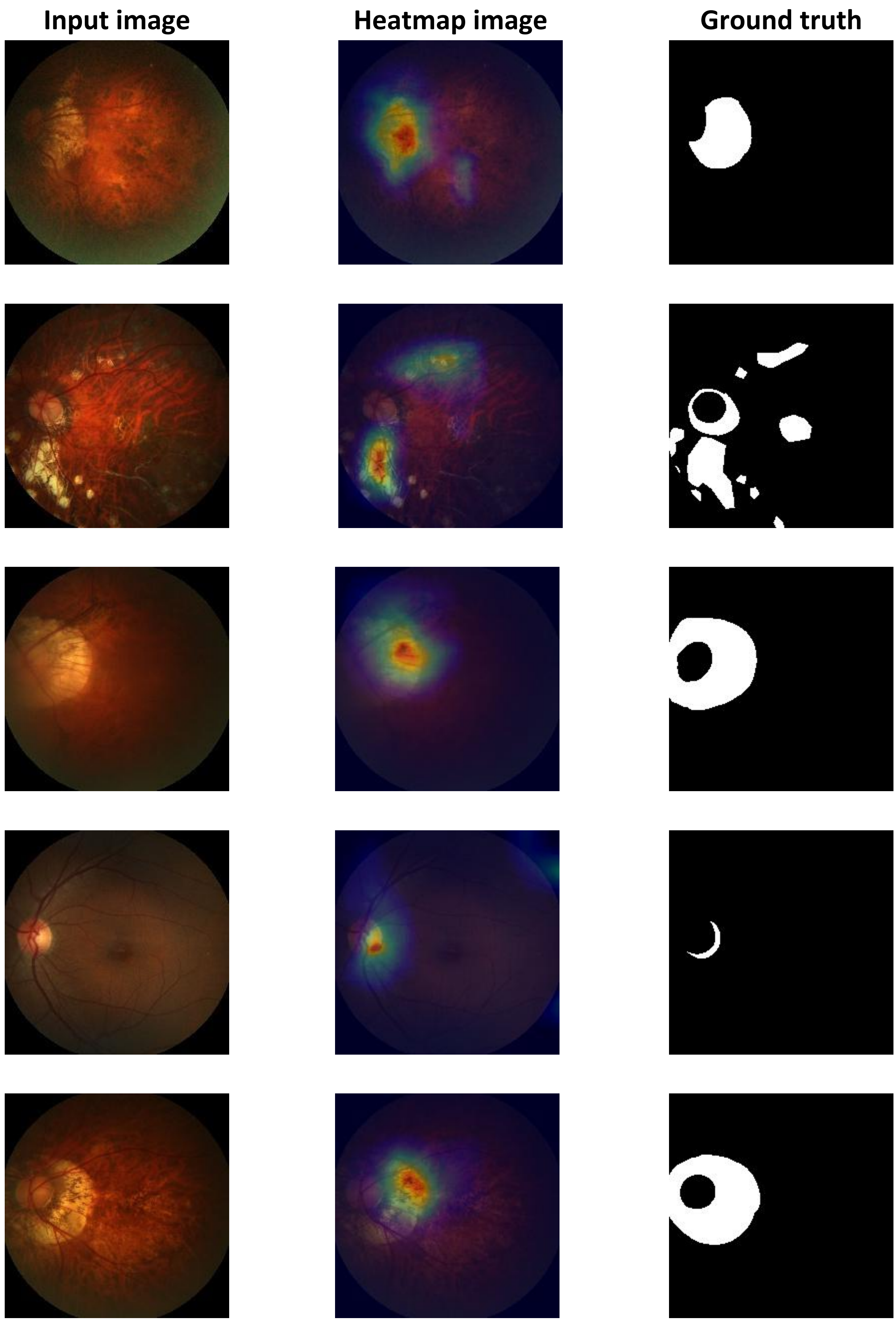

Figure 6: Visualization of HDM's saliency maps in iChallenge-PM images. All saliency maps are normalized to range [0,1] and visualized using JET colormap. The saliency map's color gradient from blue to red indicates the neural network's increasing attention probability.