北京航空航天大學
BEIHANG UNIVERSITY

机器学习报告

院（系）名称	Secrecy
学生姓名	HuangCaohui
学生班级	Secrecy
学生学号	Secrecy
题目	基于 SAM 的图像显著性预测
指导教师	Secrecy

目录

基于 SAM 的图像显著性预测

一、研究背景

当人类观察者观看图像时，有效的注意力机制将他们的目光吸引到视觉刺激具有明显变化的显着区域。大量的研究工作试图模拟这种选择性的视觉机制，因为计算显著性可以应用于广泛的应用，如图像重定向、对象识别、视频压缩、跟踪和其他依赖于数据的任务，例如图像字幕。

传统的显著性预测方法通过定义捕获低级线索（如颜色、对比度和纹理）或语义概念（如面部、人物和文本）的特征来遵循生物学证据。

由此，文章[1]提出了一种新颖的显著性预测架构，它结合了注意力卷积长短期记忆网络（Attentive ConvLSTM），该网络迭代地关注相关空间位置以细化显著性特征。

众所周知，当观察者查看计算机显示器上呈现的复杂场景时，有一种强烈的倾向，即围绕场景中心而不是周边。文章[1]给出了一种可训练的并且可以自动学习中心先验的高斯先验网络。

二、模型架构

2.1 SAM 模型

SAM (Saliency Attentive Model)模型，即显著性注意力模型，用于图像显著性预测，模型结构如图 2-1 所示。

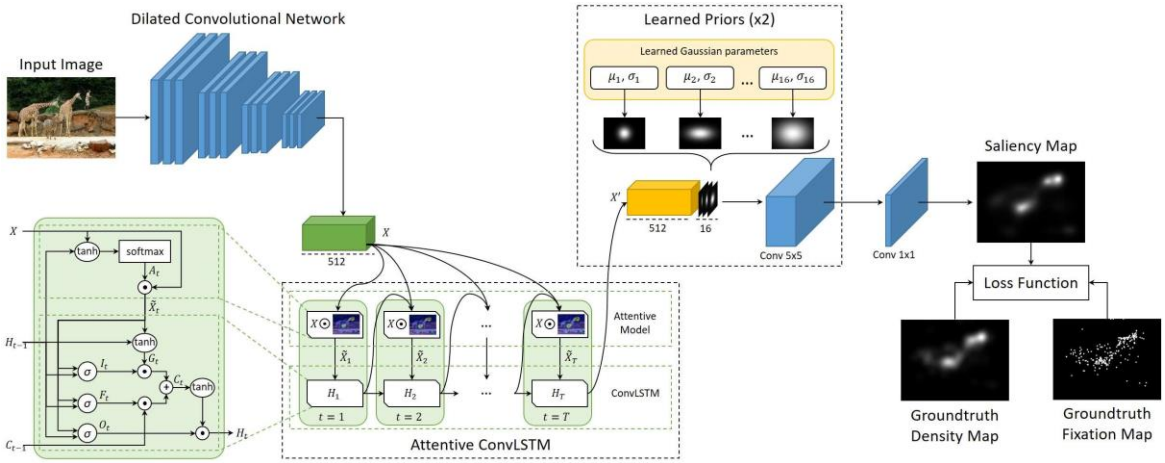

图 2-1 SAM 模型

模型主要由三部分构成：扩张卷积网络，卷积 LSTM 网络和高斯先验网络。模型首先将数据输入扩张卷积网络，获取多通道特征的输出 X 。再将输出 X 送入 ConvLSTM 网络，获取最后一个时刻 T 的网络输出 X'。再通过网络学习服从不同高斯分布的高斯先验知识，将网络输出 X' 与学习的到的高斯先验知识做空洞卷积，获取网络最后一级的输出。最后将高斯先验网络的输出与单通道的卷积核进行卷积、上采样并归一化获取最终的显著性预测图，通过与实际显著性图进行比较，获取网络的损失函数值再进行反向传播计算。-

2.2 DCN-VGG 网络

DCN-VGG（Dilated Convolutional Network-VGG）网络，即扩张卷积 VGG 网络，相对于原始的 VGG-16 网络，如图 2-2 中 C 列所示，主要移除了网络最后的池化层和三个全连接层，将倒数第二个池化层步长改为 1，同时将最后三层 512 的卷积网络换成空洞卷积，用于增大感受野，减小 VGG 网络在特征提取时，因为不断的池化导致图像的缩小，进而导致预测精度的降低。

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

图 2-2 原始 VGG 网络模型

修改后的 DCN-VGG 网络结构如图 2-3 所示，其中红色标注框部分即为相对于原始 VGG-16 网络的修改部分，对于该网络，一共进行了四次最大池化，前三次池化的步长为 2，最后一次池化的步长为 1，因此网络的输出图像大小将会缩放为原来的$\frac{1}{8}$，同时最终输出的通道数为 512。

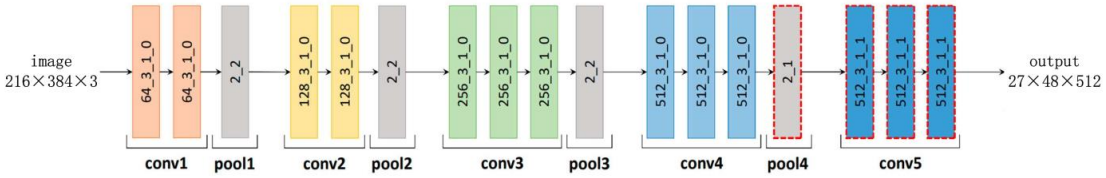

图 2-3 DCN-VGG 网络模型

2.3 ConvLSTM 网络

LSTM（Long Short-Term Memory）网络，即长短期记忆网络，原先用于处理时间序列信息，通过扩展将网络变为 ConvLSTM 网络，即卷积 LSTM 网络，利用卷积来代替原始 LSTM 的点积运算，以处理图像的空间特征；利用迭代来代替原始 LSTM 的时间依赖性的处理，从而在每一个时间戳内处理图像不同区域的显著性特征，最终获取 ConvLSTM 网络下图像的多通道特征输出。其网络结构如图 2-4 所示。

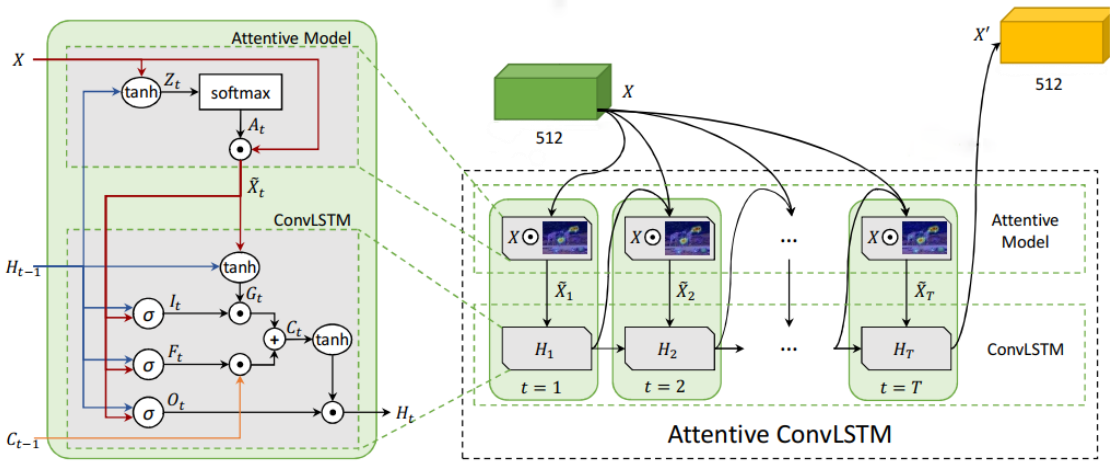

图 2-4 ConvLSTM 网络模型

在 LSTM 中，有两个状态向量C和H，其中C作为 LSTM 的内部状态向量，可以理解为 LSTM 的内存状态向量 Memory，而H表示 LSTM 的输出向量。相对于基础的 RNN 来说，LSTM 把内部 Memory 和输出分开为两个变量，同时利用三个门控：输入门(Input Gate)、遗忘门(Forget Gate)和输出门(Output Gate)

来控制内部信息的流动。将原始 LSTM 的点乘运算替换为卷积运算后得到 ConvLSTM 的内部信息流动方程为

$$I_t = \sigma\left(W_i * \tilde{X}_t + U_i * H_{t-1} + b_i\right) \tag{1.1}$$

$$F_t = \sigma\left(W_f * \tilde{X}_t + U_f * H_{t-1} + b_f\right) \tag{1.2}$$

$$O_t = \sigma\left(W_o * \tilde{X}_t + U_o * H_{t-1} + b_o\right) \tag{1.3}$$

$$G_t = \tanh\left(W_c * \tilde{X}_t + U_c * H_{t-1} + b_c\right) \tag{1.4}$$

$$C_t = F_t \odot C_{t-1} + I_t \odot G_t \tag{1.5}$$

$$H_t = O_t \odot \tanh\left(C_t\right) \tag{1.6}$$

其中 I_t、F_t、O_t 分别表示输入门，遗忘门，输出门的控制变量，G_t 表示输入向量，C_t 为本时间戳新的状态向量，H_t 为本时间戳的 ConvLSTM 输出，其中 $\odot$ 表示元素点乘运算，所有的 ConvLSTM 门输出和状态向量均为三维矩阵。其中 $\tilde{X}_t$ 通过以下方程求解

$$Z_t = V_a * \tanh\left(W_a * X + U_a * H_{t-1} + b_a\right) \tag{1.7}$$

$$A_t^{ij} = p\left(att_{ij} \mid X, H_{t-1}\right) = \frac{\exp\left(Z_t^{ij}\right)}{\sum_i \sum_j \exp\left(Z_t^{ij}\right)} \tag{1.8}$$

$$\tilde{X}_t = A_t \odot X \tag{1.9}$$

首先对输入 X 和上一时刻状态 H_{t-1} 进行卷积和加权求和，经过 tanh 函数后再与单通道的卷积核 V_a 进行卷积获取 Z_t，对 Z_t 进行 softmax 归一化后再与输入 X 进行点乘获取 $\tilde{X}_t$，其中每一时刻的输入 X 为 DCN-VGG 网络的输出，初始状态 C_0 和 H_0 全部初始化为 0。

2.4 Gaussian Prior 网络

Gaussian Prior，即高斯先验。由于观察者看图像时，他们的目光会偏向中心。这种现象主要是由于摄影师倾向于将感兴趣的物体定位在图像的中心。此外，当

人们反复观看具有显著信息的图像时，他们希望在其中心周围找到图像中信息量最大的内容。即使当图像没有高度显着的区域时，人类也倾向于看图像的中心。

因此，中心先验是图像显著性预测的关键工作，通过预先定义的先验值，让网络自己学习先验知识。实际算法如图 2-5 所示，将每个先验约束为二维高斯函数，均值和方差可以自由学习，通过不断的训练自动从数据中学习先验知识。

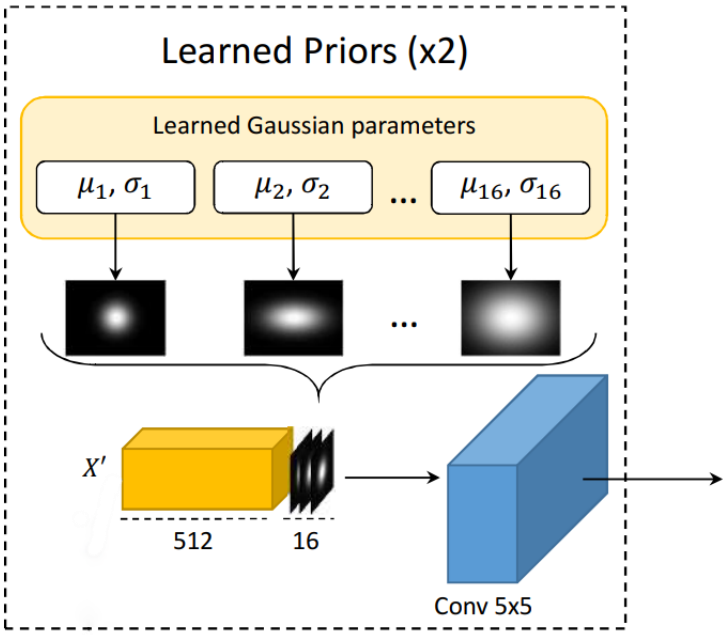

图 2-5 Gaussian Prior 网络模型

其中二维高斯函数为

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\left(\frac{(x-\mu_x)^2}{2\sigma_x^2}+\frac{(y-\mu_y)^2}{2\sigma_y^2}\right)\right) \tag{1.10}$$

在实际算法中，选取 N 个高斯函数，通过将 ConvLSTM 的输出 X' 和 N 个高斯函数拼接后利用 5×5 的卷积核进行空洞卷积，以扩大感受野，经过两次 Gaussian Prior 的学习获取高斯先验网络的输出。

最后再将高斯先验网络的输出和单通道的卷积核进行卷积并进行上采样恢复至原图像大小，归一化后获取预测的显著性图。

2.5 评估指标

对于 SAM 模型，评估指标不再单一的选取 KL 散度或者皮尔逊系数 CC，而是通过 KLD 与 CC 的线性组合构成 SAM 模型的损失函数。

CC（Correlation Coefficient Loss），即皮尔逊系数，通过协方差与方差来评价预测显著性图和实际显著性图之间的相关性，其具体公式如式(2-11)所示

$$L_1\left(\tilde{y}, y^{\text{den}}\right)=\frac{\sigma\left(\tilde{y}, y^{den}\right)}{\sigma(\tilde{y})\cdot\sigma\left(y^{den}\right)} \tag{1.11}$$

其中 $\tilde{y}$ 表示预测的概率分布，y^{den} 表示实际的概率分布，当 CC 为 1 时，表示两图像完全相关，即预测显著性图等于实际显著性图。

KLD（Kullback-Leibler divergence），即 KL 散度，主要用于评估分布 $\tilde{y}$ 和分布 y^{den} 在概率上的近似程度，，其具体公式如式(2-12)所示

$$L_2\left(\tilde{y}, y^{\text{den}}\right)=\sum_i y_i^{\text{den}}\ \log\left(\frac{y_i^{\text{den}}}{\tilde{y}_i+\varepsilon}+\varepsilon\right) \tag{1.12}$$

ε 为正则化系数，表示一个计算机的存储的无穷小数，防止分母为 0 造成数值震荡，i 表示第 i^{th} 个像素，由式可以看出，KLD 越低，两分布越接近，预测显著性图越接近真实显著性图。

在原论文中，还采用了 NSS 指标，但针对本数据集，由于缺少人眼的二进制注视图，故不考虑 NSS 指标，对 KLD 和 CC 指标进行线性组合，最终的评价指标如式(2-13)所示

$$L\left(\tilde{y}, y^{\text{den}}\right)=\beta L_1\left(\tilde{y}, y^{\text{den}}\right)+\gamma L_2\left(\tilde{y}, y^{\text{den}}\right) \tag{1.13}$$

其中 β 和 γ 表示 KLD 和 CC 指标的加权系数。

三、实验过程

3.1 数据集处理

对于本数据集，训练集共 1600 副图片，测试集共 400 副图片，训练数据共 20 类，每类数据 80 张，测试数据也 20 类，每类数据 20 张，原始图像大小为 1080×1920，实际实验时考虑到计算资源有限，将图像缩放为 216×384。

将 1600 组训练数据划分为训练集和验证集，训练集和验证集比例为 9:1，验证集用于防止网络训练过拟合。对训练集和验证集随机打散并对像素进行归一化，测试数据集不打散，只做归一化，方便后面统计不同类别数据的预测性能。

3.2 网络搭建

将 DCN-VGG、ConvLSTM、Gaussian Prior 网络组合起来，上一级网络输出作为下一级网络的输入，并在最后对 Gaussian Prior 网络输出进行单通道卷积、上采样和归一化获取最终的预测图，构成完整的 SAM 模型。

在实际实验中，根据论文完整的复现了 SAM 网络进行数据集测试，但实验中发现网络震荡不收敛，故对 DCN-VGG 网络除空洞卷积外卷积部分添加 batch normalization 操作（以下简称 BN）。在 DCN-VGG 除空洞卷积外每一次卷积后添加 BN 操作，再经过 Relu 激活函数。

网络参数设置与初始化方面，DCN-VGG 所有卷积的激活函数全部采用 Relu 函数，卷积核大小为 3，空洞卷积扩张率为 2；ConvLSTM 内部权重卷积核和二维显著性图输出卷积核的通道数均为 512，所有权重矩阵和U_a矩阵初始化值从均值为 0，标准差为 0.05 的正态分布中采样，U_i,U_f,U_o,U_c初始值为随机正交矩阵，V_a和所有偏置参数初始化为 0，卷积核大小全部为 3，初始状态C_0和H_0全部初始化为 0，时间步长为 4；Gaussian Prior 的高斯函数个数N设置为 16，卷积核大小为 5，空洞卷积扩张率为 4，激活函数为 Relu 函数，并对每一个高斯函数的均值和方差进行了范围限制。

3.3 评估指标

采用 KLD 和 CC 指标线性组合的形式，由于网络的目标为最大化 CC 和最小化 KLD，故权重因子β设置为-2，γ设置为 10。

3.4 训练模型

设置最大的 epoch 为 50，batch size 大小为 20，方便统计不同类型数据的测试结果；学习率为2×10^{-4}，优化器选择 Adam 优化器；正则化项的正则化系数2×10^{-6}，采用 L2 正则化，在网络训练时，保存模型进入过拟合前的最优参数。

3.5 测试模型

加载验证集上最优模型进行测试，由于 batch size 设置为 20，正好对应测试集一个类别的图片数，最终再统计所有评估指标的均值。

四、实验结果

4.1 训练结果

网络训练时，最终训练集和测试集的误差曲线如图 4-1 和 4-2 所示

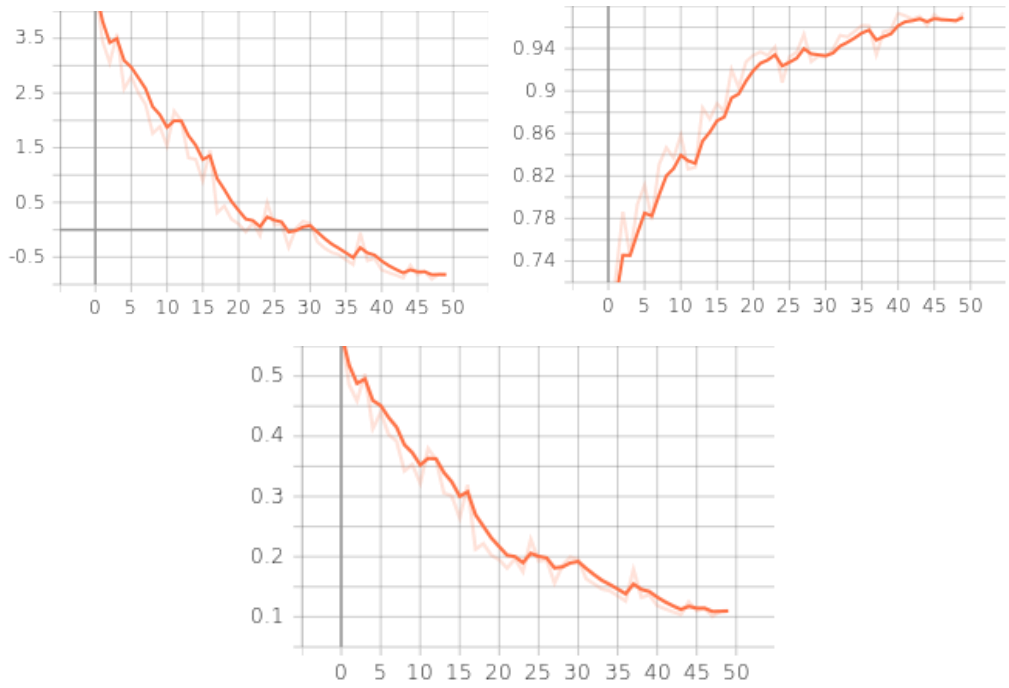

图 4-1 训练集 loss、CC、KLD 曲线

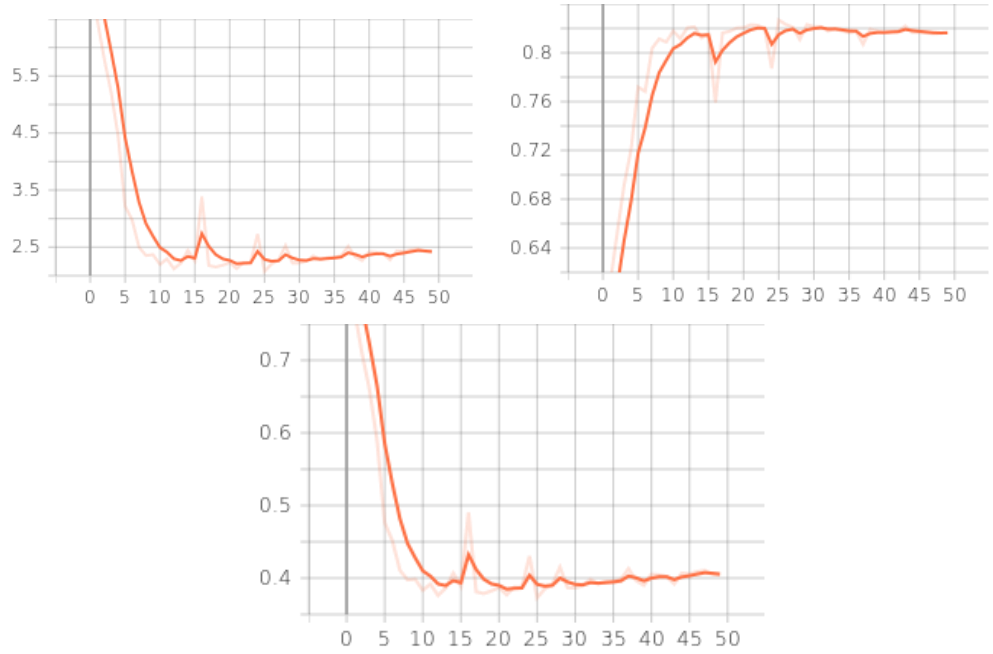

图 4-2 验证集 loss、CC、KLD 曲线

通过观察误差曲线可以发现，随着迭代次数增加，训练集的 loss 和 KLD 不断下降，CC 不断上升，在训练结束阶段，更是达到了 0.9776 的相关性，但验证

集在 epoch 为 26 时获得了误差最小值，随着训练的进行，模型开始出现些许过拟合现象趋势，验证集效果开始变差。其输出图像如图 4-3~4-5 所示，其中每行依次表示原始 RGB 图像、真实显著性图，单通道卷积归一化 DCN-VGG 输出，单通道卷积归一化 ConvLSTM 输出，单通道卷积归一化 Gaussian Prior 输出与最终预测输出；每一列表示一组图像数据。

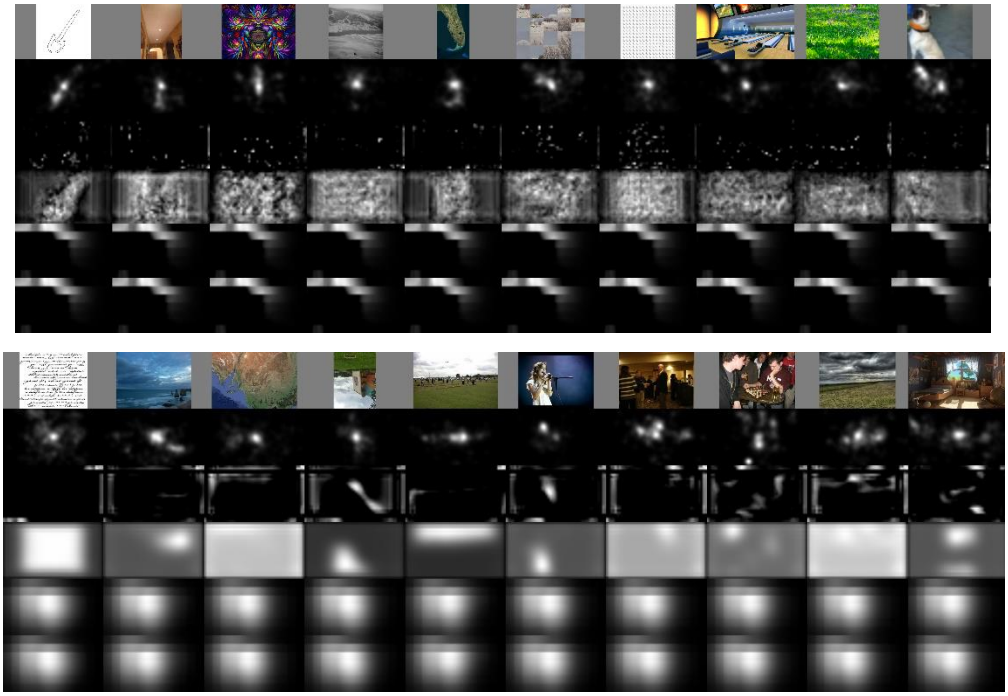

图 4-3 第 1 个 epoch 训练集与验证集图

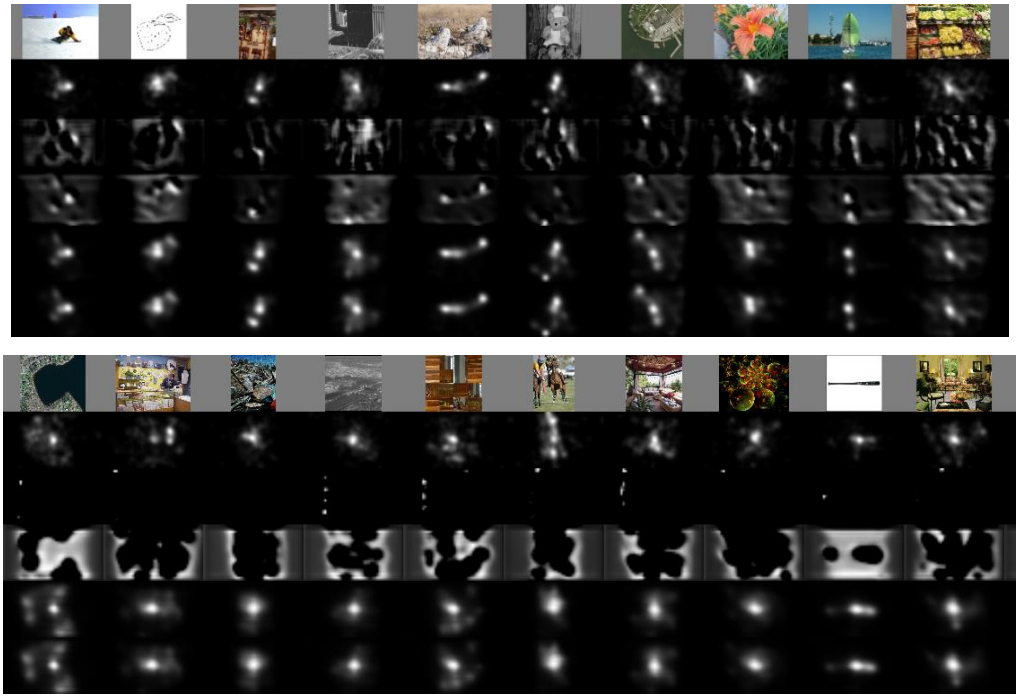

图 4-4 第 26 个 epoch 训练集与验证集图

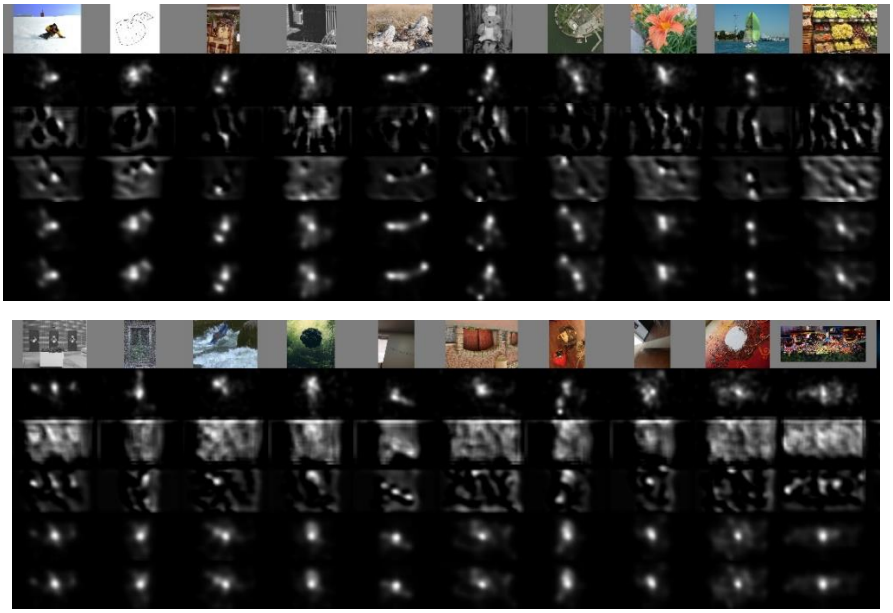

图 4-5 第 50 个 epoch 训练集与验证集图

对比图 4-3~4-5 第二行和最后一行可以发现，在第 1 个 epoch，即初始状态，参数完全随机，基本没有预测效果；在第 26 个 epoch，训练集和验证集均取得不错的效果；在第 50 个 epoch，训练集预测效果更好，但验证集预测效果有些许变差，没有产生非常严重的过拟合现象，但无论是预测点的分布还是密度，相对于实际显著性图均有些许差别。

4.2 测试结果

根据训练结果，加载 epoch 为 26 时保存的最优模型进行测试集测试，最终各个类别的测试结果如表 4-1 所示

表 4-1 SAM 模型测试集评估指标值

Metric	Action	Affective	Art	BlackWhite	Cartoon
KLD	0.4021	0.4289	0.3998	0.3936	0.3363
CC	0.7959	0.8061	0.7977	0.8157	0.8303
Metric	Fractal	Indoor	Inverted	Jumbled	LineDrawing
KLD	0.3975	0.3729	0.3704	0.3544	0.3168
CC	0.8218	0.8328	0.8361	0.8207	0.8679
Metric	LowResolution	Noisy	Object	OutdoorManMade	OutdoorNatural
KLD	0.3131	0.3869	0.3231	0.3979	0.3809
CC	0.8891	0.8509	0.8571	0.8083	0.8197

Metric	Pattern	Random	Satelite	Sketch	Social
KLD	0.3323	0.3656	0.3654	0.2591	0.4343
CC	0.8698	0.8414	0.8500	0.8916	0.7853

观察表 4-1 可以发现，在 20 个类别图像的测试中，在 LowResolution 和 Sketch 类别的图像上均取得了不错的效果，而在 Action 和 Social 类别图像上测试效果相对较差，即采用 SAM 模型方法对模糊图像和素描画的显著性图预测能取得较好的效果，而对人类行为和社交照片预测效果相对较差。其输出图像如图 4-6~4-9 所示。

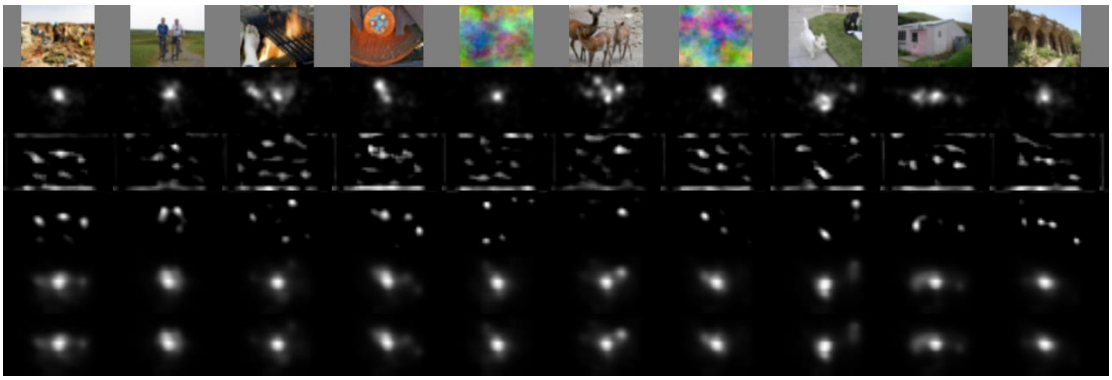

图 4-6 LowResolution 图集测试输出

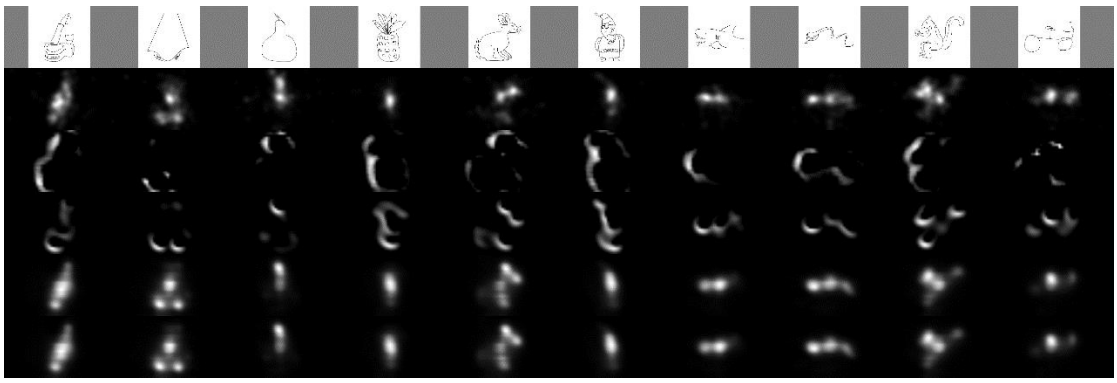

图 4-7 Sketch 图集测试输出

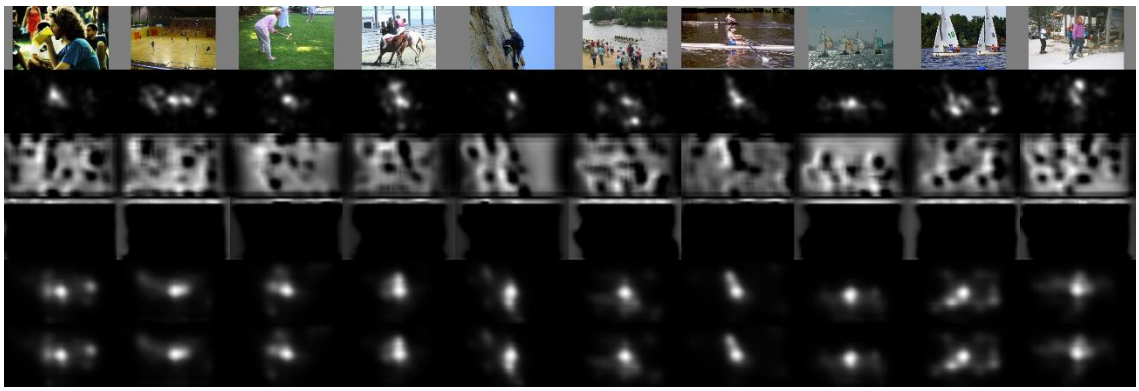

图 4-8 Action 图集测试输出

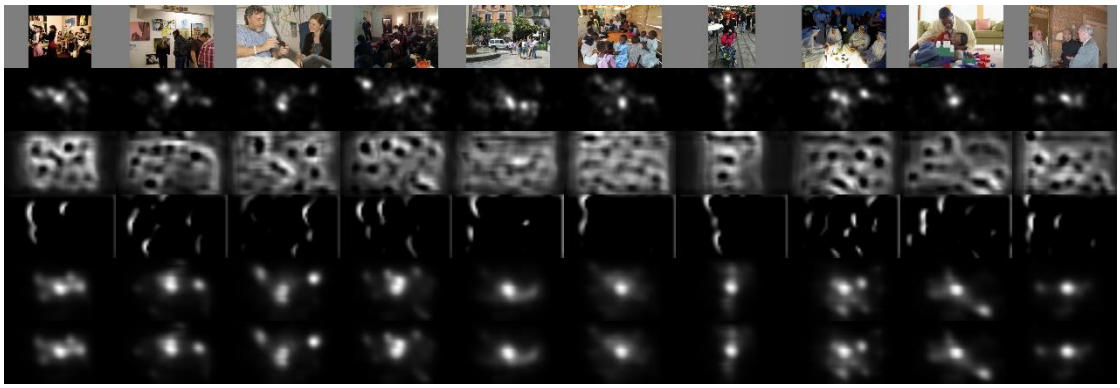

图 4-9 Social 图集测试输出

观察图 4-5~4-9 可以发现，在模糊图像和素描画上其预测显著性图相对于真实显著性图非常接近，而对人类行为和社交照片的预测，其亮点密度相对较大。初步估计，可能由于对于模糊图和素描画，其纹理较少或图像特征较少，人眼观察区域单一，对于深度网络相对好预测；而对于人类行为和社交照片，其内容非常丰富，人眼在观察时，注意点相对较多，不仅局限于中心区域，对于深度网络，难以完整的提取所有特征并转换为显著性图；其次，该网络是对所有类别图像建立的一个模型，其参数可能较难兼顾所有类型图像进行显著性预测，故其对于部分复杂图片，其预测效果不会非常显著。

最终总体评估的 KLD 均值与 CC 均值分别为 0.3666 和 0.8344，总体来说，其测试结果相对较好，与原论文在其他数据集上测试结果相接近，故针对此数据集，SAM 模型达到基本要求。

参考文献

[1]. M. Cornia, L. Baraldi, G. Serra and R. Cucchiara, "Predicting Human Eye Fixations via an LSTM-Based Saliency Attentive Model," in IEEE Transactions on Image Processing, vol. 27, no. 10, pp. 5142-5154, Oct. 2018, doi: 10.1109/TIP.2018.2851672.