论文速读：SwinOCSR - Optical Chemical Structure Recognition with Swin Transformer and Transformer Decoder

参考文献：https://link.springer.com/content/pdf/10.1186/s13321-022-00624-5.pdf

📌 一、研究背景与任务定义

📍任务定义：

OCSR（Optical Chemical Structure Recognition）：将科学出版物中的化学结构图像（如 JPEG/PNG）转换为机器可读的化学表示，如 SMILES 或 DeepSMILES。

🧱 OCSR 类似于“图像字幕生成任务”：

- 图像 → 文本序列（分子描述语言）
- 本质是 **图像到序列的翻译任务**

⚠️ 挑战：

1. 化学结构图形复杂
2. SMILES 表达式较长，容易积累错误
3. Token（元素字符）分布极度不均衡（如：C/H/O 频繁，Br/Cl 稀有）

💡 技术路线创新点（贡献）：

1. 使用 **Swin Transformer** 替代传统 CNN，提取更全面图像特征
2. 采用 **Transformer 解码器** 生成 DeepSMILES（结构合法性更高）
3. 引入 **多标签 Focal Loss** 处理 token 不均衡问题
4. 构建了 500 万分子图合成数据集（4类结构）

🔧 架构概览（SwinOCSR）

模块分为三层：

模块名称	功能描述
Swin Transformer	局部+全局注意力提取图像特征
Transformer Encoder	融合上下文语义
Transformer Decoder	自回归地生成结构语言 DeepSMILES 表达式

模型架构与技术实现

🧩 1. 图像编码器：Swin Transformer（作为 Backbone）

📌 原理：

Swin Transformer 是一种分层的 Vision Transformer，用**局部窗口 + 平移窗口**提取局部与全局特征，解决 CNN 只能感知局部的局限。

🔨 (1) Patch 分割与嵌入：

输入图像 $I \in \mathbb{R}^{H \times W \times 3}$，划分为 $P \times P$ 大小的非重叠 patch，形成 token 序列。

每个 patch 映射为向量 $x_i \in \mathbb{R}^{P^2 \cdot 3}$，通过线性变换 $W \in \mathbb{R}^{(P^2 \cdot 3) \times d}$ 得到嵌入向量：

$$z_i = W x_i + b$$

输出为 token 序列 $Z = \{z_1, z_2, ..., z_N\}$

🧠 (2) Swin Block（见 Fig.3 图示）

每个 Swin Block 包括两个注意力阶段：

👉 **第一阶段：Window-based Multi-head Self-Attention (W-MSA)**

- 局部注意力，仅在小窗口中进行
- 对每个窗口中的 tokens：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

- 多头版本为：

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O$$

👉 **第二阶段：Shifted Window Multi-head Self-Attention (SW-MSA)**

- 对窗口做平移，捕捉跨窗口的长距离依赖

两阶段 attention 的输出通过残差连接和前馈网络：

$$x' = \text{W-MSA}(x), \quad x'' = \text{MLP}(\text{LayerNorm}(x + x'))$$

📚 2. Transformer Encoder（见 Fig.4）

输入为 Swin 提取的 patch 特征序列，加上位置编码：

$$S_e = \text{Encoder}(S_b + \text{PosEmbedding})$$

每层 Transformer Encoder 包含：

1. **多头自注意力层**：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V$$

2. **前馈网络（MLP）**：

$$\text{MLP}(x) = W_2 \cdot \text{GELU}(W_1 x + b_1) + b_2$$

3. **LayerNorm + 残差连接**：

$$x' = \text{LayerNorm}(x + \text{Attention}(x)), \quad x'' = \text{LayerNorm}(x' + \text{MLP}(x'))$$

✏️ 3. Transformer Decoder（见 Fig.5）

作用是逐步生成 DeepSMILES 字符串。结构与 Encoder 类似，但包含：

- **Masked Multi-Head Attention**：防止当前 token 看到未来信息
- **Cross-Attention**：与 Encoder 输出交互（融合图像上下文）

👇 解码器中的三重注意力层：

1. **Masked Self Attention**：

$$\text{Mask}(i, j) = \begin{cases} 0, & j \leq i \\ -\infty, & j > i \end{cases}$$

注意力权重被遮蔽掉未来位置。

2. **Encoder-Decoder Attention**：

- Query 来自 decoder token，Key 和 Value 来自 encoder 输出。

3. **MLP 和输出层**：

最终输出通过：

$$y_t = \text{softmax}(W_o h_t + b_o)$$

每一步预测一个字符 token，直到结束符为止。

🧾 4. DeepSMILES 表达与 Tokenization

相比传统 SMILES，**DeepSMILES 避免使用括号和数字标记环结构**，能减小语法出错率。

示例：

- SMILES: `C1=CC=CC=C1`
- DeepSMILES: `C=CC=CC=`

整个预测序列是字符级生成的，使用了 76 个唯一字符，每个字符作为一个 token。

🔍 1. 问题背景： Token 不均衡

DeepSMILES 表达中出现的字符（token）分布极度不均衡（典型的长尾分布）：

- 高频 token 如：`)`, `C`, `c`, `=`
- 低频 token 如： `[Li]`, `#`, `Cl`, `Br`

📊 总 token 数：**234,706,822**

这种情况下使用普通 **交叉熵（Cross-Entropy, CE）** 会导致：

- 高频 token 的损失主导整个目标函数
- 低频 token 预测错误对总损失贡献小 → **模型“懒得学”低频 token**

❌ 原始损失函数： Cross Entropy Loss

传统 CE 定义如下：

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{n} y_i \log(p_i)$$

- y_i：真实标签（one-hot）
- p_i：模型预测的 softmax 概率
- n：token 的总类别数（共 76 个）

📉 问题：**所有 token 权重一致**，无法处理频率不均问题。

✅ 2. 解决方案： 多标签 Focal Loss（Multi-label Focal Loss, MFL）

Focal Loss 是为目标检测中解决类不平衡设计的，SwinOCSR 将其改造为适用于字符级多分类的**多标签版本**。

🔁 多标签转换机制

作者将 token 分类任务视为 **多标签二分类**，即每个类别都输出一个 logit，做独立判断。

1. 对每个类别 i，输出一个 logit o_i
2. 经 sigmoid 激活得概率：

$$p_i = \sigma(o_i) = \frac{1}{1 + e^{-o_i}}$$

3. 构建 Focal Loss：

$$\mathcal{L}_{\text{MFL}} = \frac{1}{n} \sum_{i=1}^{n} -\alpha_i (1 - p_i)^{\gamma} \log(p_i)$$

✅ 参数定义：

符号	含义
α_i	类别 i 的权重（低频 token 给更高权重）
γ	聚焦参数（常用 2）抑制容易分类样本的影响
p_i	模型预测的概率
y_i	真实标签（0 或 1）
n	类别数（本任务中为 76）

数据构建与训练实验设计

🧬 1. 数据来源与预处理流程

🔍 来源：

- 下载自 **PubChem**（前 850 万条 SMILES）
- 处理后得到 **698 万个唯一 SMILES**
- 最终构建了 **500 万条**用于训练的数据（每类 125 万）

📐 步骤：

1. **SMILES → DeepSMILES**
   - 使用 **RDKit** 进行 Aromatic 转换与结构标准化
2. **图像渲染**：
   - 使用 **CDK (Chemistry Development Kit)** 生成分子结构图
   - 采用字体调整、角距、下标控制等，贴近真实文献风格
3. **图像转为灰度 + 二值图**：
   - 二值图复制 3 通道 → 模拟 RGB 输入格式
   - 图片尺寸统一为 **224×224**

🧩 2. 分子结构的四种类别

构建数据集时，分子被划分为以下四类（每类 125 万）：

类别	环结构类型	是否包含取代基
1	Kekule	✕ 无
2	Aromatic	✕ 无
3	Kekule	✅ 有
4	Aromatic	✅ 有

- **Kekule**：使用单/双键显示环结构，形状不明显
- **Aromatic**：使用圆形环表示，视觉上更清晰
- **取代基**：附加的官能团、原子符号如 `[Si]`、`[NH]` 等，出现在专利数据中较多

📌 目标：**覆盖真实出版物中常见的化学结构图表达方式**

🧪 3. 训练设置

项目	设置
图像输入	224×224，3 通道（复制二值图）
Batch Size	256
Optimizer	**Adam**
初始学习率	5×10^{-4}
Token嵌入维度	256
学习率调度器	Swin 使用 Cosine Decay，Transformer 使用 Step Decay
Epoch	30
GPU	NVIDIA Tesla V100

🧱 4. 数据集划分

比例为 **18:1:1**：

- 训练集：4,500,000
- 验证集：250,000

- 测试集：250,000

📊 5. 多维度实验设计

✅ (A) 分子结构类型影响（Table 5）

类别	准确率表现	分析原因
3, 4	更高	✅ 取代基在图中显著
2, 4	高于 1, 3	✅ Aromatic 圆环形状更清晰
总体差异	不大	✅ 模型鲁棒性强，对不同结构适应性好

✅ (B) DeepSMILES 长度影响（Fig. 12）

字符长度范围	准确率表现
[1–75]	准确率稳定 ✅
[76–100]	轻微下降（94.76%） 🔻

📌 说明：即使序列很长，模型也能稳定预测，体现 Transformer 解码器的 **长序列建模能力**。

✅ (C) 真实文献数据测试（Real-world Test）

- 构建 100 张真实分子图（来源于出版物），手动标注 SMILES
- 模型在真实数据上准确率：**25%**

📉 说明：

- 模型主要训练于合成图像，缺少文献图的字体变形、扫描噪声等特征
- 暗示需要后续引入 **领域自适应、数据增强或微调机制**

Backbone	Accuracy	Tanimoto	BLEU	ROUGE
SwinOCSR	**97.36%**	**99.65%**	**99.46%**	**99.64%**
ResNet-50	89.17%	-	-	-
EfficientNet-B3	86.70%	-	-	-

📌 SwinOCSR 在 **所有指标** 上均优于使用 CNN 的 Image2SMILES、DECIMER 等方法：

- Tanimoto 相似度几乎完美（>99.6%）
- BLEU 和 ROUGE 分别显示文本一致性和覆盖率极高
- 准确率提升明显（比 ResNet-50 高 8.19%，比 EfficientNet-B3 高 10.66%）

🧠 3. 性能关键影响因素回顾

因素	正向影响说明
Swin Transformer	具备更强图像语义建模能力，层次结构支持大分子解析
Transformer 解码器	优于 GRU，长序列建模更稳健
Focal Loss	解决 token 长尾分布，提升稀有字符识别
数据增强策略	多结构覆盖（Kekule/Aromatic + 有/无取代基）提升泛化性

🧾 4. 限制与改进方向

📉 真实文献图像准确率仅 25%

说明 SwinOCSR 尚未完全适应真实世界的图像分布差异。

📌 潜在改进方向：

- 引入 **领域自适应**（Domain Adaptation）
- 合成图与真实图混合训练
- 使用 OCR 噪声模拟、图像风格迁移等方法扩展鲁棒性

✅ 最终结论汇总

论文核心贡献	内容简述
技术创新	首次使用 Swin Transformer + Transformer 解码器用于 OCSR
表达优化	使用 DeepSMILES 替代 SMILES，提高预测合法性
损失优化	多标签 Focal Loss 解决 token 不均衡问题
数据集构建	构建 500 万分子图像合成数据集，含四种结构类型
性能提升	所有评估指标均优于现有方法，准确率达 97.36%，BLEU > 99.4%