基于stable diffusion的数据增强及数据标注

> *在计算机视觉领域，图像生成指根据模型或算法生成新的图像。将图像生成技术应用于实际产业能有效节省生产时间和人工成本，提高工作效率。目前该技术已被应用于艺术品生成、广告设计、游戏开发、机器人视觉和虚拟现实等领域，具有广阔的应用前景。*

1 扩散模型与Stable Diffusion

1.1 扩散模型

我们根据使用技术的不同，可将图像生成方法分为传统方法和基于深度学习的方法。传统方法包括基于纹理合成和基于马尔可夫随机场等方法。然而，传统方法的生成效果欠佳，且通常仅适用于特定的任务场景。基于深度学习的方法根据学习类型可分为生成对抗网络 、自回归模型 、变分自编码器、流模型、能量模型和 扩散模型等。与传统方法相比，基于深度学习的图像生成方法生成的图像质量更好，能满足实际应用需求。在实际应用中大多对生成内容有一定要求，条件引导的图像生成方法能增强对生成过程的控制，使生成内容向目标要求迈进；与其他生成模型相比，扩散模型在条件引导的图像生成中具有生成质量高和多样性强等优点，亦展现出巨大的发展潜力。

1.1.1 扩散模型的定义

扩散模型的目标可以总结为逆转数据逐渐退化的过程，包括符合马尔科夫链的正向过程和逆扩散过程，如图 1 所示。正向过程，在原始数据中逐步添加噪声，使其逐渐退化为几乎各向同性的高斯噪声，破坏原始数据；逆扩散过程，通过神经网络学习从高斯噪声中恢复原始数据。需要指出的是，逆扩散过程输入和输出的数据维度需保持不变，由于U-Net符合此要求且开销较小，被广泛用于构建扩散模型的去噪网络。

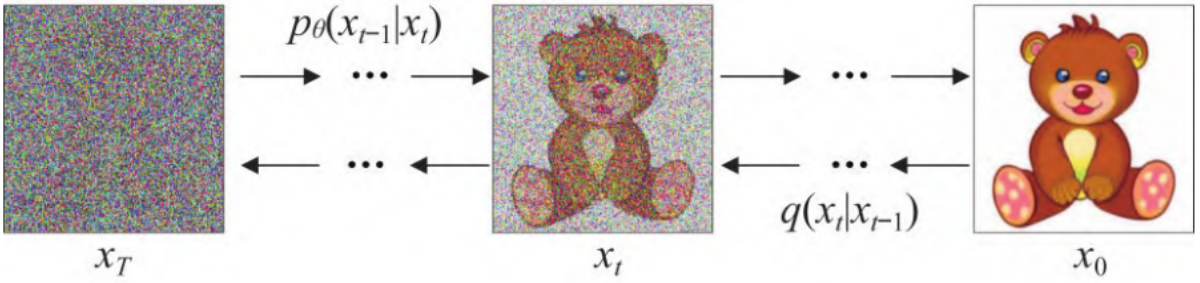

图 1　扩散模型的正向和逆扩散过程

根据定义方法，扩散模型可分为去噪扩散概率模型（denoising diffusion probabilistic models，

DDPMs）、分数生成模型（score-based generative models，SGMs）和 随机微分方程（stochastic differential equations，SDEs）3 类。

1.1.2 分类器引导

在基于扩散模型的图像生成中，显式分类器引导、隐式分类器引导和多模态引导均能提高图像生成的灵活性和质量。

1 显式分类器

虽然 DDPM和改进的 DDPM等方法均能生成逼真的图像，但在 FID 等指标上与基于生成对抗网络的主流方法仍存在一定差距。为了解决此问题，DHARIWAL 等引入了显式分类器引导（classifier guidance），通过在图像生成过程中注入类别标签信息增强对模型的约束，从而提高图像生成质量。基于该技术，由扩散模型生成的图像 FID 等指标超过了基于生成对抗网络的 BigGAN 。

通过去噪网络构建$p(x_{t-1}|x_t)$是扩散模型的关键，利用贝叶斯公式显式分类器将预训练无条件图像生成模型中的$p(x_{t-1}|x_t)$转化为$p(x_{t-1}|x_t,y)$：

$$\$p(x_{t-1}|x_t,y)=\frac{p(x_{t-1}|x_t)p(y|x_{t-1},x_t)}{p(y|x_t)}\$$$

其中 y 表示给定的条件。由于该过程仍基于预训练无条件图像生成模型的生成过程$p(x_{t-1}|x_t)$，所以不需要训练新模型。

经计算，$p(x_{t-1}|x_t,y)$近似于$\mathcal{N}(x_{t-1};\mu(x_t)+\sigma_t^2\nabla_{x_t}\lg p(y|x_t),\sigma_t^2I)$，其中中间项就是无条件生成过程的新增项。

虽然显式分类器的成本较低，但存在以下缺陷：（1）需额外训练一个分类器，增加了模型的复杂性；（2）分类器在一定程度上决定了模型的上限，优化该分类器也具有一定的难度；（3）显式分类器引导会破坏生成结果的多样性等。

2　隐式分类器

受显式分类器的启发，HO 等提出了隐式分类器引导（classifier-free guidance）。与显式分类器不同，隐式分类器可直接将转换函数$p(x_{t-1}|x_t,y)$定义为，$\mathcal{N}(x_{t-1};\mu(x,y),\sigma_t^2I)$这是两者的本质差别。

隐式分类器无需训练额外的分类器，而是训练条件生成和无条件生成 2类模型。在训练模型时，引入额外的条件 y，在不需要条件时用空值代替条件y，从而降低训练难度，但大大增加了训练成本。目前的图像生成大模型 Glide、Stable Diffusion、DALLE2 和 Imagen 等均采用隐式分类器，生成效果惊人。

3　多模态引导

基于 CLIP 的多模态引导被应用于条件引导的图像生成、文本引导的图像编辑和文本引导的图像转换等任务。预训练的 CLIP模型能很好地衡量图像与文本之间的相似程度特性，有助于模型构建

$$\mathcal{L}\left(x_{\text{gen}}, y_{\text{tar}}; x_{\text{ref}}, y_{\text{ref}}\right) = 1 - \frac{\langle \Delta I, \Delta T \rangle}{\|\Delta I\| \|\Delta T\|}$$

其中$\Delta T = E_T(\mathrm{y}_{\text{tar}}) - E_T(\mathrm{y}_{\text{ref}})$为文本之间的向量差值,$\Delta I = E_I(\mathrm{y}_{\text{tar}}) - E_I(\mathrm{y}_{\text{ref}})$ 为图像之间的向量差

值，E_T 和E_I分别为 CLIP 模型的文本编码器和图像编码器。

由于 CLIP 模型在生成过程中计算每步损失均极其耗时，随着隐式分类器的提出，基于 CLIP 的多模态引导逐渐被淘汰。

综上，条件引导图像生成的条件和对应方法如图 2 所示。

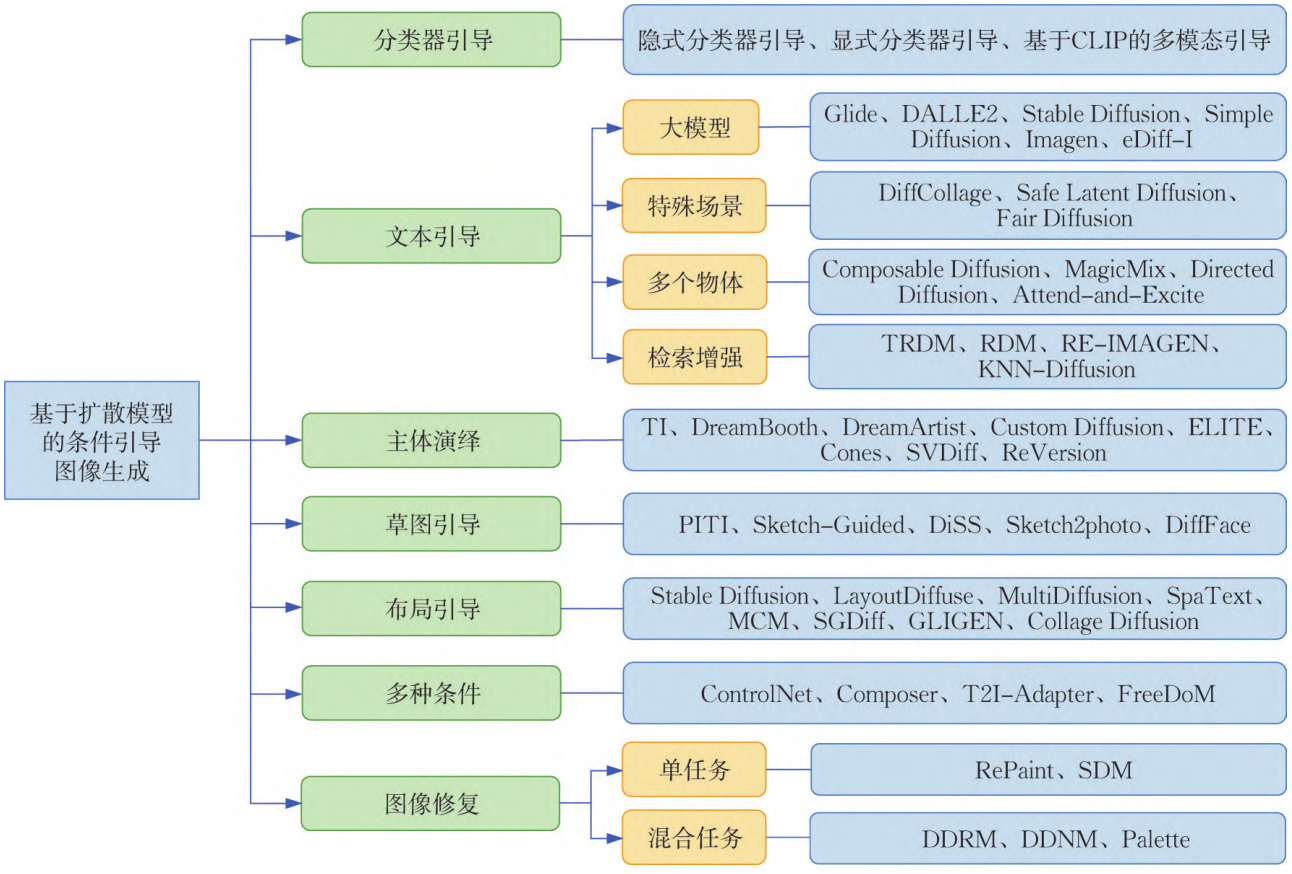

图 2　条件引导图像生成的条件和对应方法

1.1.2 条件引导图像生成的应用

随着扩散模型的发展，近几年出现了许多条件引导的图像生成模型，其中与文本相关的条件引导

生成模型占多数，这与文本数据易获取、数量多有关，且利用文本控制图像生成的过程相对简单。此外，还有针对特定任务的生成模型，如基于草图、布局的生成模型等。

1 基于文本引导的图像生成模型

计算机视觉和自然语言处理是深度学习的两个热门方向，为此有大量的图像数据和文本数据被保存，其中有多个为开源的大规模图像文本数据集。与无条件图像生成方法相比，基于文本引导的图像生成方法能显著提高图像生成的可控制性，不仅自动化程度高、灵活性强，还能在节省时间成本的同时应对复杂的场景。下面主要介绍大型文本生成图像模型、特定场景下的文本生成图像模型、生成包含多个物体图像的文本生成图像模型和基于检索增强的文本生成图像模型。

目前实际应用最多的文本生成图像模型主要有Stable Diffusion、 Midjourney 和 DALLE2。BORJI A.在真实场景下对上述模型生成人脸图像的能力进行了对比。结果表明，Midjourney 生成的图像具有明显的动漫和艺术风格，真实性最差，Stable Diffusion 生成的图像真实性优于 DALLE2。FID 指标，Stable Diffusion 最好，其次是 DALLE2，最后是Midjourney。

2 对图像的主体内容进行演绎的图像生成模型

尽管现有的大型文本生成图像模型生成的图像具有良好的真实性和多样性，但无法在给定参考图像的情况下对参考图像中的主体进行相关的演绎。Textual Inversion在不改变图像主体基本属性的情况下，根据文本对图像中的主体进行了创造性演绎，首先通过隐向量空间的文本编码器学习新的概念，然后根据文本所包含的特定概念实现对图像的精细控制。

3 以草图为条件的图像生成模型

基于扩散模型的条件图像生成方法能够产生具有显著多样性和真实感的图像，然而，大多方法只允许对标签或文本提示进行调节，限制了对最终结果的控制强度。因此，出现了一些以草图为条件的图像生成方法。

1.2 Stable diffusion

扩散模型最大的问题是它的时间成本和经济成本都极其“昂贵”。Stable Diffusion的出现就是为了解决上述问题。如果我们想要生成一张 1024 × 1024 1024 \times 10241024×1024 尺寸的图像，U-Net 会使用 1024 × 1024 1024 \times 10241024×1024 尺寸的噪声，然后从中生成图像。这里做一步扩散的计算量就很大，更别说要循环迭代多次直到100%。一个解决

方法是将大图片拆分为若干小分辨率的图片进行训练，然后再使用一个额外的神经网络来产生更大分辨率的图像（超分辨率扩散）。

2021年发布的Latent Diffusion模型给出了不一样的方法。 Latent Diffusion模型不直接在操作图像，而是在潜在空间中进行操作。通过将原始数据编码到更小的空间中，让U-Net可以在低维表示上添加和删除噪声。

1.2.1 潜在空间(Lantent Space)

潜在空间简单的说是对压缩数据的表示。所谓压缩指的是用比原始表示更小的数位来编码信息的过程。比如我们用一个颜色通道（黑白灰）来表示原来由RGB三原色构成的图片，此时每个像素点的颜色向量由3维变成了1维度。维度降低会丢失一部分信息，然而在某些情况下，降维不是件坏事。通过降维我们可以过滤掉一些不太重要的信息你，只保留最重要的信息。假设我们像通过全连接的卷积神经网络训练一个图像分类模型。当我们说模型在学习时，我们的意思是它在学习神经网络每一层的特定属性，比如边缘、角度、形状等.....每当模型使用数据（已经存在的图像）学习时，都会将图像的尺寸先减小再恢复到原始尺寸。最后，模型使用解码器从压缩数据中重建图像，同时学习之前的所有相关信息。因此，空间变小，以便提取和保留最重要的属性。这就是潜在空间适用于扩散模型的原因。

1.2.2 Latent Diffusion

“潜在扩散模型”（Latent Diffusion Model）将GAN的感知能力、扩散模型的细节保存能力和Transformer的语义能力三者结合，创造出比上述所有模型更稳健和高效的生成模型。与其他方法相比，Latent Diffusion不仅节省了内存，而且生成的图像保持了多样性和高细节度，同时图像还保留了数据的语义结构。

语义压缩
在学习的第二阶段，图像生成方法必须能够捕获数据中存在的语义结构。 这种概念和语义结构提供了图像中各种对象的上下文和相互关系的保存。 Transformer擅长捕捉文本和图像中的语义结构。 Transformer的泛化能力和扩散模型的细节保存能力相结合，提供了两全其美的方法，并提供了一种生成细粒度的高度细节图像的方法，同时保留图像中的语义结构。

1.2.3 感知损失

潜在扩散模型中的自动编码器通过将数据投影到潜在空间来捕获数据的感知结构。论文作者使用一种特殊的损失函数来训练这种称为“感知损失”的自动编码器。该损失函数确保重建限制在图像流形内，并减少使用像素空间损失（例如 L1/L2 损失）时出现的模糊。

1.2.4扩散损失

扩散模型通过从正态分布变量中逐步去除噪声来学习数据分布。换句话说，扩散模型使用长度为 T TT 的反向马尔可夫链。这也意味着扩散模型可以建模为时间步长为 t = 1 , … , T t =1,\dots,Tt=1,...,T 的一系列“T”去噪自动编码器。由下方公式中的 ϵ_θ表示：

$$L_{DM} = E_{x,\epsilon\sim N(0,1),t}[|| \epsilon - \epsilon_\theta(x_t, t) ||_2^2]$$

公式(1)给出了扩散模型的损失函数。在潜在扩散模型中，损失函数取决于潜在向量而不是像素空间。我们将像素空间元素x替换成潜在向量$\varepsilon(x)$，将t时间的状态x_t替换为去噪U-Net在时间t的潜在状态zz_t，即可得到潜在扩散模型的损失函数，见公式(2)：

$$L_{LDM} := E_{\varepsilon(x),\epsilon\sim N(0,1),t}[|| \epsilon - \epsilon_\theta(z_t, t), \tau_\theta(y) ||_2^2]$$

将公式(2)写成条件损失函数，得到公式(3)：

$$L_{LDM} := E_{\varepsilon(x),y,\epsilon\sim N(0,1),t}[|| \epsilon - \epsilon_\theta(z_t, t), \tau_\theta(y) ||_2^2]$$

1.2.5 条件扩散

扩散模型是依赖于先验的条件模型。在图像生成任务中，先验通常是文本、图像或语义图。为了获得先验的潜在表示，需要使用转换器（例如 CLIP）将文本/图像嵌入到潜在向量τ \tauτ中。因此，最终的损失函数不仅取决于原始图像的潜在空间，还取决于条件的潜在嵌入。

1.2.6 注意力机制**

潜在扩散模型的主干是具有稀疏连接的 U-Net 自动编码器，提供交叉注意力机制2。Transformer 网络将条件文本/图像编码为潜在嵌入，后者又通过交叉注意力层映射到 U-Net 的中间层。这个交叉注意力层实现了注意力 $(Q, K, V) = softmax(QKT/\sqrt{d})V$其中 Q、K 和 V 是可学习的投影矩阵

1.2.7 文本-图像合成

在 Python 实现中，我们可以使用使用 LDM v4 的最新官方实现来生成图像。 在文本到图像的合成中，潜在扩散模型使用预训练的 CLIP 模型3，该模型为文本和图像等多种模态提供基于Transformer的通用嵌入。 然后将Transformer模型的输出输入到称为“diffusers”的潜在扩散模型Python API，同时还可以设置一些参数（例如，扩散步数、随机数种子、图像大小等）。

1.2.8 图像-图像合成

相同的方法同样适用于图像到图像的合成，不同的是需要输入样本图像作为参考图像。生成的图像在语义和视觉上与作为参考给出的图像相似。这个过程在概念上类似于基于样式的 GAN 模型，但它在保留图像的语义结构方面做得更好。

1.2.9 整体架构

上面介绍了潜在扩散模型的各个主要技术部分，下面我们将它们合成一个整理，看一下潜在扩散模型的完整工作流程。

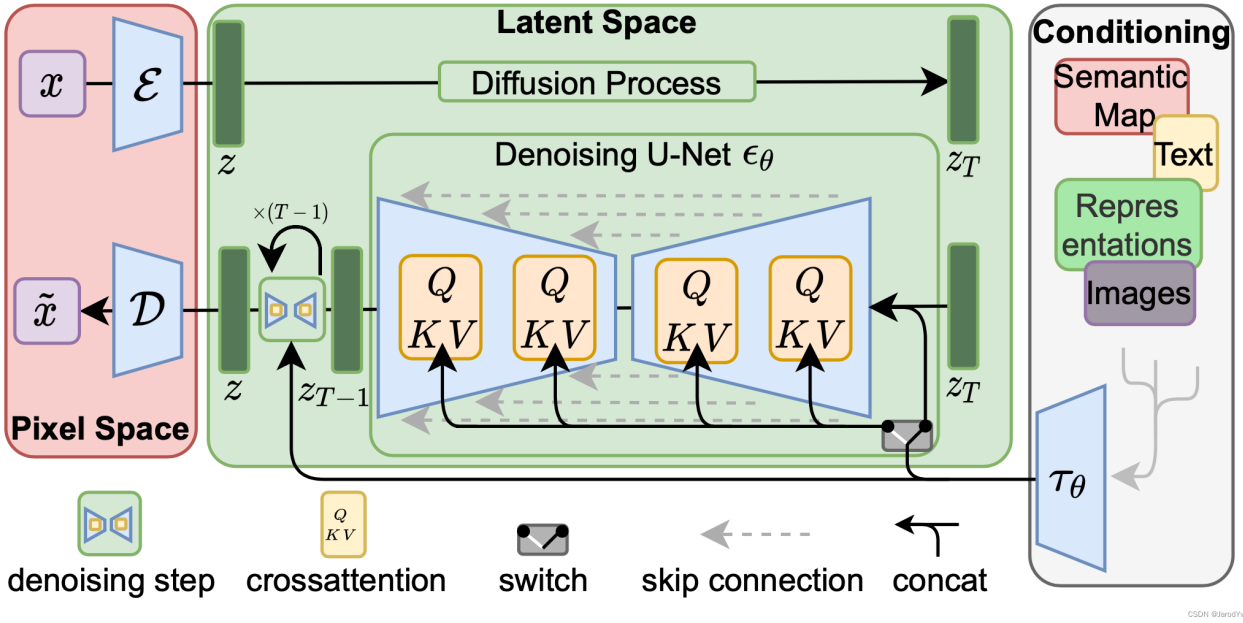

图 3　Stable Diffusion整体架构

上图中 x 表示输入图像，$\tilde{x}$ 表示生成的图像；ε是编码器，$\mathcal{D}$是解码器，二者共同构成了感知压缩；z 是潜在向量；z_T是增加噪声后的潜在向量；τ_θ是文本/图像的编码器（比如Transformer或CLIP），实现了语义压缩。

2 实验过程

2.1 Stable Diffusion的本地部署

本地部署Stable Diffusion最简单的方法是使用Stable Diffusion Web Ui。

Stable Diffusion Web Ui是一套无代码、可视化的Stable Diffusion集成运行环境。它将Stable Diffusion的安装部署集成打包，提供一键安装脚本，并提供Web界面操作界面，极大简化了Stable Diffusion的操作和使用，让没有不懂代码的小白也能轻松上手使用Stable Diffusion模型。

我们采用整合包直接安装Stable Diffusion Web Ui（在这里感谢秋葉aaaki)），安装后只需下载启动依赖即可。

2.2 使用Stable Diffusion进行数据增强

2.2.1 提示词

prompt: 主要是对于图像进行描述。prompt对Stable Diffusion图像生成质量至关重要，因此如果想生成高质量图片，一定要在提示设计上下功夫。一个好的提示需要详细和具体。

Negative prompt：主要是告诉模型我不想要什么样的风格或元素；

prompt语法
为了产生具有特定风格的图像，必须以特定格式提供文本提示。 这通常需要添加提示修饰符或添加更多关键字或关键短语来实现。下面为大家介绍一下Stable Diffusion的prompt语法规则。

Stable Diffusion提示文本中的关键字或关键短语通过半角逗号分割，一般越靠前权重越高。我们可以通过提示修饰符来认为修改权重。

- (tag)：增加权重5%
- [tag]：降低权重5%
- (tag: weight)：设置具体权重值

括号可以嵌套使用，例如：(tag)的权重为$1 \times 1.05 = 1.05$((tag))的权重为$1 \times 1.05 \times 1.05 = 1.10255$。同理[tag]的权重为$\frac{1}{1.05} = 0.952$[[tag]]的权重为$\frac{1}{1.05^2} = 0.907$

- [tag1 | tag2]：将tag1和tag2混合；
- {tag1 | tag2 | tag3}：从标签集合中随机选择一个标签；
- [tag1 : tag2 : 0.5]：表示先用tag1生成，当生成进程到50%时，改用tab2生成；如果输入整数的话表示步长，比如10，意思是生成10步后改用tag2；
- <lora:filename:multiplier>：LoRA模型引用语法

2.2.2 Stable Diffusion 模型

与DALL·E和Midjourney相比，Stable Diffusion最大的优势是开源，这就意味着Stable Diffusion的潜力巨大、发展飞快。Stable Diffusion已经跟很多工具和平台进行了集成，且可用预训练模型数量众多（参见Stable Diffusion资源列表）。正是由于社区的活跃，使得Stable Diffusion在各种风格的图像生成上都有着出色的表现。

我们在civitai.com上选择合适的stable diffusion模型下载并导入，由于本次需要生成画风写实的图片，我们挑选了PicX_real作为Stabl Diffusion模型。

2.3.3 调参

由于stable参数量众多，碍于篇幅，这里仅对调参过程中用到的参数进行介绍

迭代步数

模型生成图片的迭代步数，每多一次迭代都会给 AI 更多的机会去对比 prompt 和 当前结果，从而进一步调整图片。更高的步数需要花费更多的计算时间，但却不一定意味着会有更好的结果。当然迭代步数不足肯定会降低输出的图像质量。

采样方法

扩散去噪算法的采样模式，不同采样模式会带来不一样的效果，具体需要在实际使用中测试。

重绘程度

决定算法对图像内容的保留程度。0什么都不会改变，1会得到一个完全不同的图像；

Controlnet

调节完成以上内容后，只需点击批量生成即可开始生成图片

2.3 数据标注

使用基于百度飞桨的EasyDL平台对新扩充的数据进行标注。标注的类别主要分三类：分别是未满溢的垃圾桶、满溢的垃圾桶和垃圾（和论文中相同）。