工作

- HumanML3D：文本（词性标注），动作序列（263个特征值 * 帧数：脚是否在地上，关节的velocity），来自于AMASS动作捕捉数据集
- 100Style，我们收集的数据跟他对齐，动作序列，文本（词性标注），显式的style标签。

next step

建立在场景标签是正确的。

- （1）场景无关风格提取器：输出场景无关的style
- （2）场景提取器： 这两个接受style motion，一个输出scene（提取出scene的特征（512维），然后做100分类（100维向量）），一个输出场景无关的style，输出style的特征（512维）（vector）；这两个特征向量彼此正交。 这里可以有一个正交的loss。网络结构用原来的就可以。

- （3）场景和style融合的网络：MLP直接做融合。
- （4）GAN的判别器：判别style确实不包含场景信息。

要求：

- scene和style要尽量正交，scene当中要包含场景信息（假定是100类），
- 场景无关的style，不包含场景信息（考虑一个discriminator）

建议：

- MCM-LCM的风格提取器输出当作是融合网络的ground truth。去掉diffusion，train 上面的（1）（2）（3）（4）.相当于只是对于style vector（f_s）做了重新解释，解耦开场景和风格。
- train的时候有场景标签和场景下的motion，数据集是真的场景风格数据集。可以参考