天津大学

计算机视觉期末作业

基于 StableDiffusion 和 ControlNet 的视频生成与转绘

学　　院　智能与计算学部

年　　级　2020 级

成　　员　王家鹏 3020210111

刘锦帆 3020202184

罗超凡 3020209258

程子姝 3019244160

李雨霄 3020244279

指导教师　万亮　林迪

摘　要

AI 生成视频是一种常见的计算机视觉应用，然而，当连续生成具有复杂结构或高频细节的视频时，其性能会降低，并有可能出现频闪或变形的问题。因此，本文提出了一种基于 ControlNet 的对于 StableDiffusion 的改进方法，用于增强 StableDiffusion 模型在视频生成任务中的处理能力。具体而言，我们使用 ControlNet 引入了一个能随着基于图像内容生成的控制字符与前后两帧之间的光流数据变化向 StableDiffusion 产生影响的控制器，以增强 AI 生成视频模型在细节不断变化时保持平稳流畅视频输出的能力。

此外，我们提出在进行 ControlNet 改进过程前进行基于光流 wrap 和运动掩码的前处理，以进一步增强其视频生成效率。光流 wrap 和运动掩码的引入有助于模型缩小处理范围，在提升生成效率的同时降低可能对原图固定部分的过度处理。预测生成图像中的动态变化。

我们的实验结果表明，引入控制文本和光流数据的 ControlNet 改进方法在生成的图像质量和稳定性方面都产生了高质量的结果。

关键词：StableDiffusion；ControlNet；AI 视频生成；控制文本；光流数据；运动掩码

ABSTRACT

AI-generated videos are a common computer vision application. However, the performance of generating videos with complex structures or high-frequency details can decrease, and flickering or distortion may occur. This article proposes an improvement method for StableDiffusion based on ControlNet to enhance the processing capabilities of StableDiffusion models in video generation tasks. Specifically, we use ControlNet to introduce a controller that can affect StableDiffusion with changes in control characters based on image content and optical flow data between the previous and current frames, enhancing the ability of AI-generated video models to maintain smooth video output even as details change.

In addition, we propose pre-processing based on optical flow wrap and motion masks before the ControlNet improvement process to further enhance its video generation efficiency. The introduction of optical flow wrap and motion masks helps to reduce the processing range of the model, improving generation efficiency while reducing excessive processing of fixed parts in the original image, predicting dynamic changes in the generated image.

Our experimental results show that the ControlNet improvement method with control text and optical flow data produces high-quality results in both image quality and stability of the generated videos.

Keywords: StableDiffusion, ControlNet, AI video generation, control text, optical flow data, motion masks

目　录

第一章　研究问题与相关背景

1.1　待解决的问题

AI 生成视频闪烁、前后帧不一致。

1.1.1　问题介绍

近来，随着人工智能模型的迅速发展，各类多模态数字内容生成应用也持续出现，生成图像、视频的效率质量也在不断提高。然而其中诸如 AI 生成视频等应用却依旧存在着生成的视频频繁闪烁、前后帧不一致等问题，导致最终呈现给用户的体验不佳，产生的视频与真实世界的视频存在较大差异，降低了视频生成技术的实用性，并无法达到拟真的效果。

1.1.2　应用意义

解决 AI 生成视频闪烁和生成前后帧不一致的问题是当前视频生成技术研究的热点问题，对于提高生成视频的质量和稳定性具有很强的应用意义。若将本报告中的方法与现有的 AI 生成视频技术相结合，便能以相对较小的代价获得用户观感更加流畅连贯、更加接近真实视频的输出结果；上述问题的解决将在电影和电视制作、游戏开发、虚拟现实、安防监控等应用场景中具有重要的应用意义。

1.2　背景知识

1.2.1　ControlNet

ControlNet 是一种可以将给定的文本控制信息作为条件来生成图像的技术，通过在图像生成器中添加条件控制器，可以更加精确地控制生成的图像。

传统的文本到图像生成模型（如 GAN 和 VAE）通常是在图像生成器中引入条件输入，例如类别标签或文本描述。但是，这些模型往往难以在生成过程中精确控制所需的特定视觉特征；然而，ControlNet 可以通过将给定的文本控制信息映射到条件向量中，然后将该向量与生成器中的噪声向量进行合并，从而生成具有所需视觉特征的图像。同时，过程中添加的条件控制器还可以学习到不同的控制策略，例如对图像进行微调或引导生成器生成更具创意的图像。

对于本研究而言，该技术可以应用于 AI 生成视频的过程中，以解决生成视频时出现的闪烁和前后帧不一致等问题。具体而言，Control Net 可以将前面帧的信息作为条件来控制后续帧的生成，从而使得生成的视频更加连续和流畅。

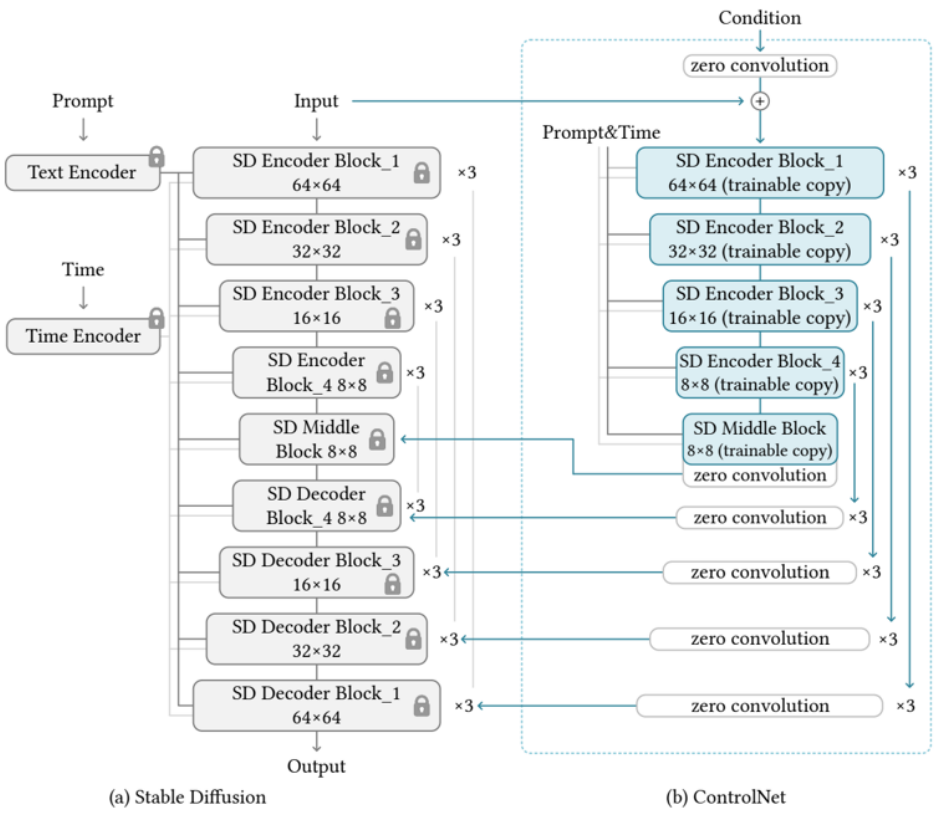


图 1-1　　ControlNet 的结构

1.2.2　StableDiffusion 与 ControlNet

StableDiffusion 是一个基于数十亿张图像进行训练的大型文本图像扩散模型。该模型本质上是一个 U-net，包括编码器、中间块和跳跃连接解码器，可用于生成高质量的图像和视频。U-Net 是一种用于图像分割任务的深度卷积神经网络。它的名字来源于其网络结构的形状，其具有对称的 U 字形，并已经被广泛应用于各种计算机视觉任务中。

StableDiffusion 以通过引入随机噪声，再经过多步扩散的过程逐渐去除噪声，从而逐步生成出高质量的图像或视频。该模型的核心思想是将噪声视为一种不确定性，在随着时间的推移而逐渐减少。具体来说，整个生成过程包括噪声的添加、扩散过程、条件图像或视频帧的生成，以及最终图像或视频的重建。该模型的优点是能够生成高质量、清晰度高、细节丰富的图像和视频，并且在生成过程中能够很好地控制生成的结果。

我们可以采用 ControlNet 对 StableDiffusion 中 U-Net 的每个级别进行控制。这种方法可以加速训练过程并节省 GPU 内存，由于 Stable Diffusion 采用了典型的 U-Net 结构，因此这种 ControlNet 架构很可能也可用于其他扩散模型。

第二章　实验方法

2.1　ControlNet 引入光流信息控制

在进行 AI 视频生成的过程中，保持生成前后帧的连贯性是决定所生成视频是否流畅真实的关键。我们将可以通过给出的文本控制信息和光流信息来影响所生成图片的 ControlNet 技术与表现真实视频中前后相邻帧的像素位置关系的光流图像数据进行结合，通过学习真实视频中相邻帧的前导帧与光流数据，并利用以对前导帧中内容的文字描述和获取到的光流信息为控制信息的 ControlNet 尝试生成正确后继帧，并在与数据集中对应后继帧的比较中不断逼近最优结果。

2.1.1　光流数据获取

为了有效建立已知图片（前导帧）与生成图片（后继帧）之间的关系，我们采取获取样本数据图片对之间的光流数据，并进一步生成光流图像的方式来完成这一目标。RAFT（Recurrent All-Pairs Field Transforms）是一种基于神经网络的光流估计算法，可以用来估计图像序列相邻帧之间的光流。RAFT 使用一个具有六层的卷积神经网络来估计光流。该网络分别接受视频或动图中的前导帧和后继帧作为输入，随后输出一个光流场，表示两个输入帧之间的运动。

在随后的过程中，我们通过获得的光流场来输出光流图。具体来说，光流图可以计算出每个像素在两个帧之间的运动向量，该向量的大小和方向表示了该像素在两个帧之间移动的距离和方向。输出的光流图使用颜色编码来可视化，通过颜色的不同显示每个像素的运动方向，并通过颜色的饱和度来体现每个像素的速度大小。

2.1.2　文字描述生成

在形成光流图的同时，我们还需要与图片内容相关的文字描述，作为 ControlNet 的控制信息。CLIP（Contrastive Language-Image Pre-Training）是一种基于图像和文本的联合训练模型，可以将图片和相关的文本描述联系起来。将源数据中的前导帧图像进行预处理后加载预训练的 CLIP 模型，输入图像并获取图像向量，指定描述图像的文本，将文本描述转换为词向量，并计算图像和文本向量之间的相似度，便可以方便地获得相关文字描述，用以指导 ControlNet 的生成。文字描述的生成对于输出结果的质量具有决定性作用，因为其为预测原图像在未来可能产生变化的重要依据。

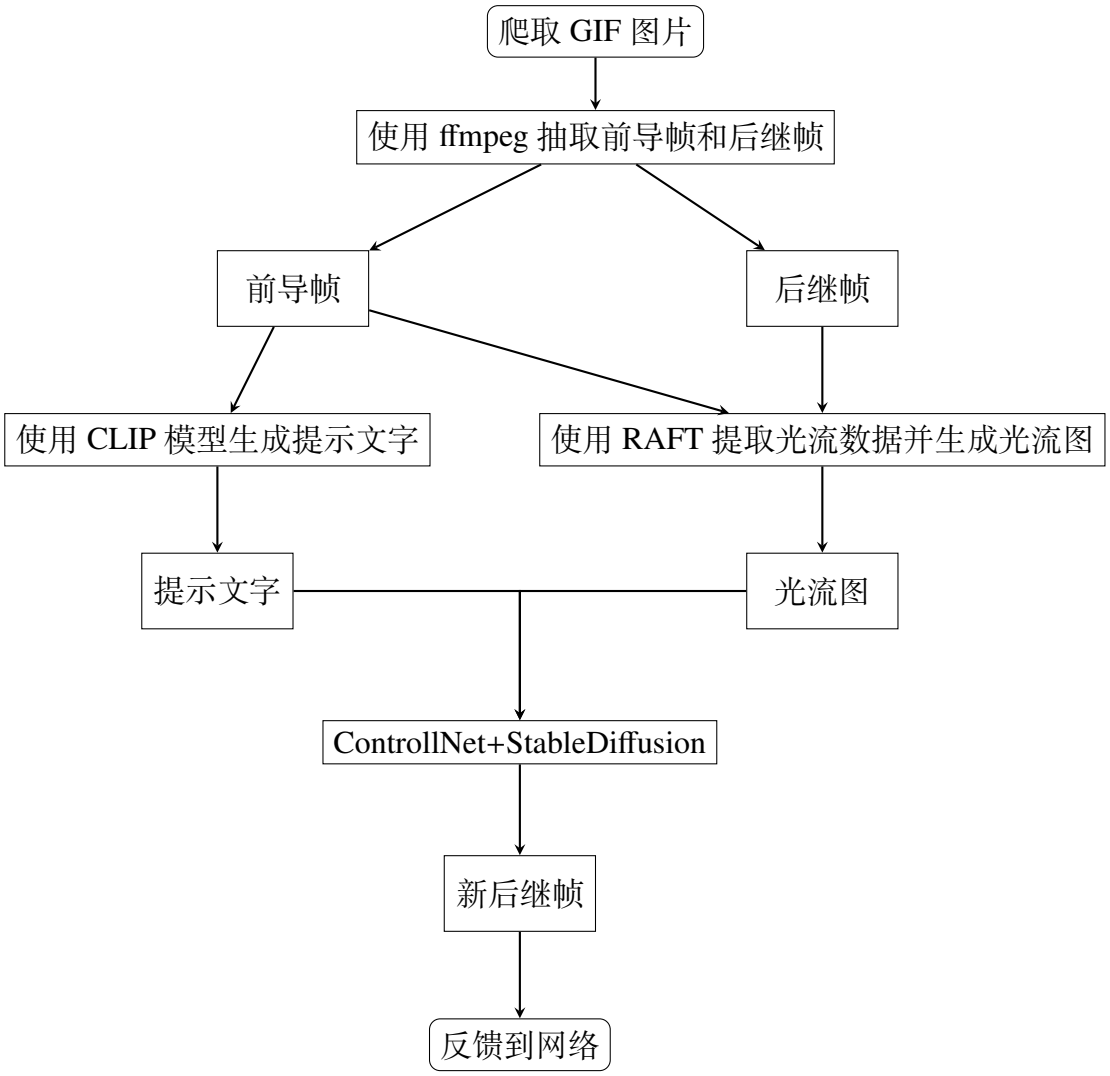


图 2-1　　总体流程图

至此，用于进一步训练的数据集制作完成。具体的数据集格式与示例将在第三章中专门介绍。

2.1.3　StableDiffusion 的 ControlNet 改进

StableDiffusion 是一种强大的图像处理和计算机视觉方法，模拟了像素之间的稳定扩散过程。然而，当处理具有复杂结构或高频细节的图像时，其性能会降低，导致输出结果失真。我们提出了一种 ControlNet 改进方法，用于增强 StableDiffusion 模型在视频生成任务中的处理能力，以进一步增强其视频生成能力。

具体而言，我们将文字描述生成产生的控制文本作为输入，并在模型中添加相应的模块，以利用输入图像中的语义信息来生成更准确的图像。我们还将光流数据与输入图像一起提供给模型，可以帮助模型更好地理解视频中的运动信息，

从而更好地预测后一帧图像中的动态变化。具体而言，光流数据可以告诉模型前一帧图像中的物体如何移动，并估计这些移动在后一帧图像中的位置。控制器可以利用这些信息来自适应地调整模型中的扩散速率和扩散时间等参数，最终使得生成的图片与提供的参数信息相符合。

2.2 基于光流数据的局部图像生成

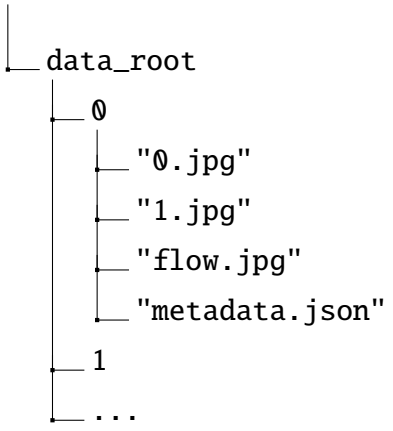
我们利用先前获得的光流信息，可以进一步提升 ControlNet 引入光流信息控制预测后继帧生成的效果。可以注意到，现实中的视频总是高度连续的，即前后两帧之间变动的范围并不大。因此，对前后两帧间不甚运动的背景像素进行反复的重新训练则会消耗运算资源，并存在被错误重绘的风险。例如，在仅仅使用 ControlNet 指导 StableDiffusion 生成图片的过程中，ControlNet 对于生成的图片仅仅只能施加部分影响，这将有可能导致控制效果不明显，进而丢失控制文本和光流图中所描述的部分细节。

介于光流图能够描述像素运动的特性，我们将光流图中不参与前后帧移动的像素排除在图像生成的范围之外，藉此在一定程度上增强 StableDiffusion 的输出效果。

第三章　实验数据

3.1　数据集概述

为了满足我们的客制化需求，我们需要制作新的数据集。为了确保 Control Net 能学习到控制文本和光流信息对 Stable Diffusion Model 进行指导，我们考虑将数据集的一个单元设计为 **“前导帧”**，**“后继帧”**，**“两者之间的光流”** 以及一个 **描述性 的 Json 文件**，且不划分训练和测试集。结构如下：



其中，"metadata.json" 格式如下：

```
{
  "source"       #源片段hashid
  "min"          #光流最小值（用于归一化）
  "max"          #光流最大值
  "prompt"       #与图片内容对应的文字提示
}
```

3.2　数据采集与处理

对于互联网上的数据，我们经过如下处理将其转换为可用的数据集：

1. 从 giphy 网站多线程下载动图；
2. 使用 ffmpeg 抽取相邻帧，每张动图进行一次抽取；
3. 使用 RAFT 算法抽取光流，形成 “前导帧-光流-后继帧” 的初始数据集：RAFT 算法通过对前导帧和后继帧进行特征提取，可以确定图像中的特征点，并通过这些提取到的特征点计算出每个特征点在两帧之间的运动向量。这些向量可以被组合成一个完整的光流场，最终输出中间部分的光

流数据；

4. 使用 CLIP 对图片内容生成文字提示：

   CLIP (Contrastive Language-Image Pre-training)，即对比文本-图像预训练模型，使用大规模的图像和文本数据来学习将图像和文本联系起来的方式。其可以对给定的图像和文本进行语义匹配，从而将图像内容映射到与之相关的单词或短语。在 CLIP 模型中，图像和文本都表示为向量，因此在将图像输入模型之后，将获得一个与图像相关的向量表示。在搜索与图像相关的单词或短语时，可以使用 Python 的 NumPy 库等工具计算图像向量与词表中每个单词的余弦相似度。余弦相似度越大，表示图像向量与单词向量之间的关联性越强，因此可以将余弦相似度作为一个度量图像和单词之间关联的度量。在计算余弦相似度之后，可以选择余弦相似度最高的单词或短语作为图像的描述。最终，我们将这些描述输出到片段元数据的“prompt”字段中；

5. 数据完整性校验
6. 重组并编号，形成最终数据集

3.3 格式要求

1. 尺寸要求：

   对于每张提取的图片，尺寸统一处理为 256x256，对于不符合规格的图片，我们先按照比例对图片进行缩放，使得其最短边与目标尺寸匹配，随后对多余长边部分进行剪裁，最终得到符合规范的图片。

2. 存储格式要求：

   所有图片均保存为 jpg 格式，计算出的光流数据也同时保存为 jpg 格式。

3.4 数据集示例

我们在此提供一些实际参与实验过程的数据示例。

示例 1：hash_id = 1gQtwPbKiQIWxfDGpt

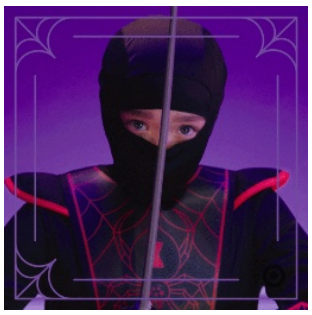


图 3-1 前导帧 0.jpg

图 3-2 光流 flow.jpg

图 3-3 后继帧 1.jpg

```
// metadata.json
{
  "source": "1gQtwPbKiQIWxfDGpt",
  "min": "-54.65134",
  "max": "27.403206",
  "prompt": "a person in a costume with a sword . "
}
```

示例 2：hash_id = 1eguaknQlhpg81aCU4

图 3-4 前导帧 0.jpg

图 3-5 光流 flow.jpg

图 3-6 后继帧 1.jpg

```
// metadata.json
{
  "source": "1eguaknQlhpg81aCU4",
  "min": "-37.065155",
  "max": "22.992125",
  "prompt": "a man holding a box with a birthday message on it . "
}
```

第四章　实验内容

4.1　训练 ControlNet 网络

我们使用 1 块 NVIDIA RTX3090 GPU 对自建数据集中的数据训练了 33 epoch，batch_size 大小设置为 4，学习率为 1e-5。对于训练输入，我们设想了以下三种方法。

1. 将光流图以马赛克的形式 patch 到第一帧图片上。这种方法可以提供稀疏的运动信息，且仍然保持输入图片的三通道数目不变，更利于训练。
2. 将光流的可视化 HSV 图片与原图进行 concat，得到 6 通道 tensor，再输入网络进行训练。
3. 将光流的可视化 HSV 图片逆向转换为两通道光流数据，去除没有意义的明度通道，再与原图 concat，得到 5 通道输入数据。

综合考虑提供信息的丰富度与效率，我们选择最后一种方法实现 data_loader。

4.2　基于光流 warp 和运动掩码的前处理

仅使用 ControlNet 并不足以维持前后帧的高度一致，原因主要在于：训练过程中，原本的 StableDiffusion 参数被锁定，仅更新 ControlNet 的参数。因此原本的 StableDiffusion 能够完整地保留其图像生成能力。为了确保视频前后帧一致性，需要对原 StableDiffusion 施加较强的约束。而 ControlNet 参数量只有原 StableDiffusion 的一半，因此只能一定程度上引导 StableDiffusion 进行图像内容的生成，并不能对生成的图像细节进行严格的约束。

基于以上考虑，我们设计了基于光流 warp 和运动掩码的前处理步骤，和 ControlNet 互为补充。首先通过 warp 操作得到一张粗略的下一帧图片，然后使用运动掩码标记 inpainting 区域，最后使用 StableDiffusion 和 ControlNet 进行细化，得到最终结果。

4.2.1　光流 warp

光流描述了两帧图片之间像素的相对位移信息。通过给定前一帧 I_A 以及 I_A 到后一帧 I_B 的光流 $F_{A\to B}$，可以进行前向 warp 操作得到图片 $I_{B'}$。当运动不大、亮度变化不明显时，$I_{B'}$ 可看作 I_B 的粗略近似。

需要注意的是，前向 warp 会造成像素缺失的空洞现象，一般能够使用最近邻插值等对其进行填补。我们选择用噪声填充空洞区域，然后使用 StableDiffusion 进行去噪修复。

图 4-1 前导帧

图 4-2 光流

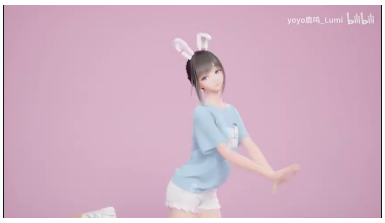


图 4-3 后继帧

4.2.2 运动掩码

由于我们希望视频连续帧中不运动的物体保持一致，因此我们使用运动掩码进行前景和背景分割，运动物体作为前景，标记为 inpainting 区域，背景和上一帧保持一致。

mask 的作用是分割运动物体，运动物体在图像中的掩码应当具有一定的连续性，不应出现太多孔洞。因此我们对光流二值化产生的掩码进行形态学闭操作去除 mask 的孔洞区域。

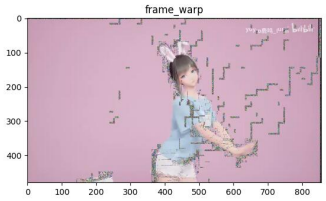


图 4-4 warp

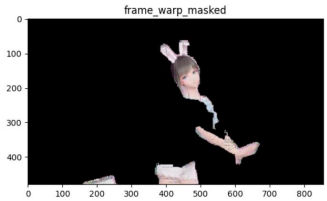


图 4-5 光流掩码

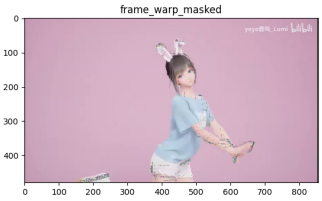


图 4-6 warp&masked

第五章　实验结果分析与讨论

5.1　实验平台

我们基于 Gradio 搭建了 WebUI, 便于进行试验操作与展示。

观察实验结果，可以看出示例 1 图（3）的手部和图（4）的头部有较大幅度运动，生成的图像也都具有对应位置的运动模糊。示例 2 图（2）的杯子经对比具有相对位移且很好维持了手部形态。这些说明了 ControlNet 学习到了运动信息并发挥了比较重要的作用。

5.2　ControlNet 网络测试生成结果

见下页实验结果展示。

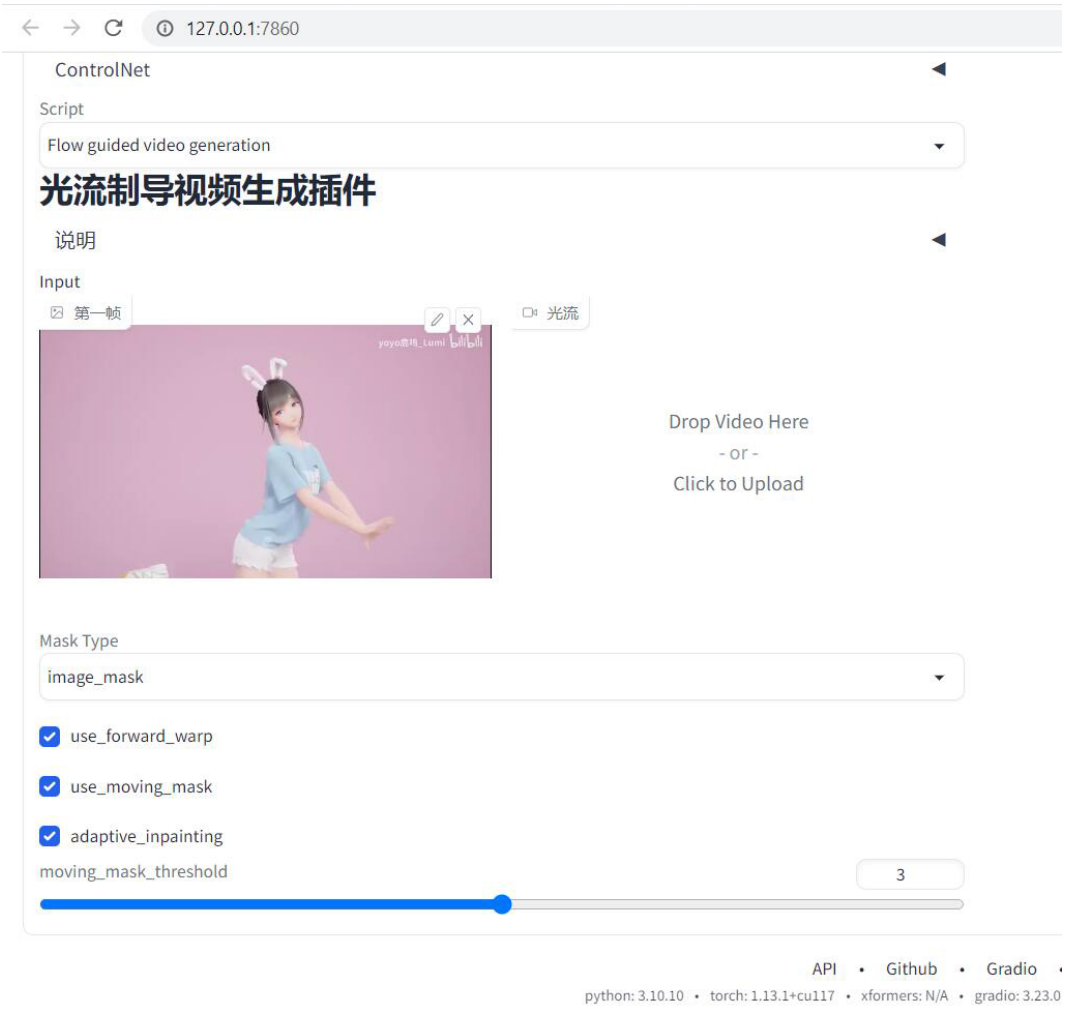


图 5-1　WebUI

图 5-2　　第一帧（示例 1）

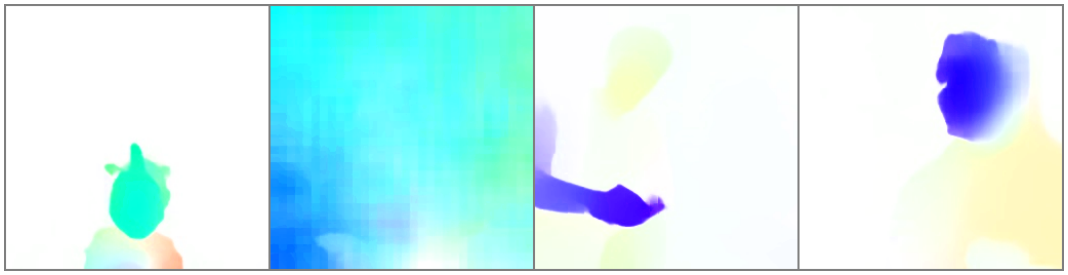


图 5-3　　光流（示例 1）

图 5-4　　生成图像（示例 1）

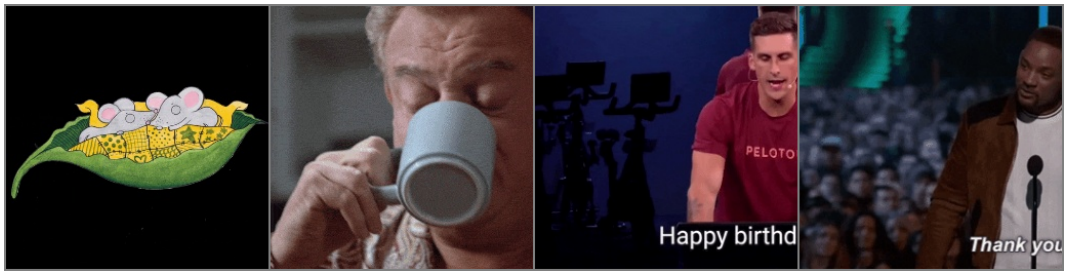


图 5-5　　第一帧（示例 2）

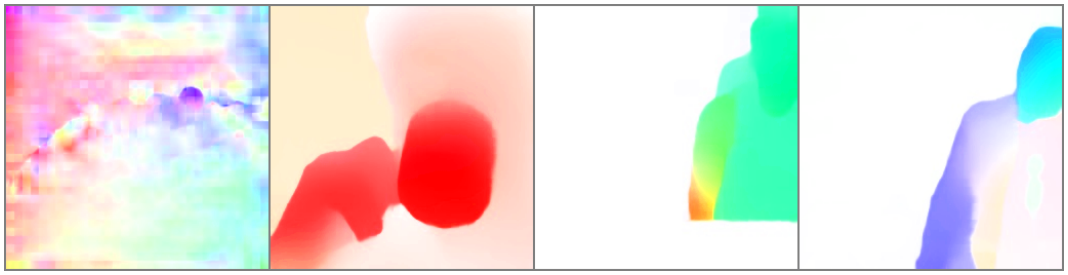


图 5-6　　光流（示例 2）

图 5-7　　生成图像（示例 2）

第六章　讨论与结论

6.1　讨论

局限性及分析

可以发现单独使用 ControlNet，根据上一帧和光流直接生成的下一帧图片效果并不理想。尽管 ControlNet 已经能够使生成的新视频帧具备合理的运动幅度和运动模糊效果，但新视频帧和原视频帧在色彩、纹理方面仍存在一定差距。

我们推测造成这个现象的根本原因是 ControlNet 对负责生成图像的 LDM 约束力欠缺。具体而言可能由以下几个因素导致：

1. 数据集规模太小。
2. 训练步数太少。
3. ControlNet 本身参数量是 LDM 的一半，而 LDM 在训练时参数锁定，因此单纯依靠 ControlNet 可能无法有效控制 LDM。（LDM 由于参数锁定，因此具备完整的图像生成能力。因此可能需要 ControlNet 具备接近 LDM 的参数量才能对 LDM 进行有力地控制，使生成图像的前后帧高度一致。）

我们进而希望在 image to image 模式下，通过适当降低去噪强度来确保前后帧的一致性。

调低去噪强度会使模型生成能力欠缺，过高的去噪强度会加剧前后帧的不一致。我们希望通过 ControlNet 对其进行平衡，同时引入前向 warp 和运动物体掩码进一步改善质量。

未来工作

由于 stable diffusion webui 的 ControlNet 插件没有实现光流模态的控制信息输入，因此我们分别实现了光流模态的 ControlNet 和视频生成的 webui 插件。

由于时间限制，我们没能将已经实现的光流模态 ControlNet 完整移植到 webui 上，插件部分仅实现了前向 warp 和运动掩码部分。未来我们将完成 ControlNet 的移植并探索不同参数对生成质量的影响，寻找合适的去噪强度参数。

同时我们将从模型结构出发，设计更加合理的 ControlNet 结构使其具备更强的约束力以适配视频生成与转绘任务。

6.2　结论

针对视频生成中前后帧闪烁、一致性差的问题，我们提出了一种基于 ControlNet 使用光流作为输入的视频生成或转绘模型。我们构建了包含 **7925 组**视频

帧-光流-文本标注的多模态数据集，并训练了对应的 ControlNet 模型。为了进一步提升生成质量，我们使用了通过光流进行前向 warp 的预处理步骤，通过运动掩码分割前景背景，实现 inpainting 区域的精确标记。最后我们基于目前最常用的 stable diffusion 平台 automatic1111 stable diffusion webui 编写了光流视频生成插件，将我们的工作集成到了 DDPM 图像生成 pipeline 中。

综上所述，我们的贡献主要可以分为三点：

1. 提出了基于 ControlNet 结合光流信息进行 AI 视频转绘的模型
2. 构建了包含 **7925 组**视频帧-光流-文本标注的多模态数据集。
3. 编写了对应的 webui 的插件，将上述工作整合进了最流行的 DDPM 图像生成流水线中。

参考文献

[1] Radford A, Kim J W, Hallacy C, *et al*. Learning transferable visual models from natural language supervision [C]. In International conference on machine learning, 2021: 8748–8763.

[2] Zhang L, Agrawala M. Adding conditional control to text-to-image diffusion models [J]. arXiv preprint arXiv:2302.05543, 2023.

[3] Rombach R, Blattmann A, Lorenz D, *et al*. High-resolution image synthesis with latent diffusion models [C]. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10684–10695.