MAGICLICK：A Video Cropping System

1.Introduction

随着智能移动设备的广泛普及，手机录像已成为我们捕捉生活精彩瞬间的重要手段。但美好时刻转瞬即逝，用户常常来不及深思熟虑地构图，加之大多数人缺少后期处理的技能和意识，导致许多经过处理的视频作品缺乏基本的美学价值，最终成果往往令人失望。鉴于此，一种简单快捷的交互平台能够迅速且精确地满足用户需求。MAGICROP（Wang等，2023年）在图片裁剪方面已做出显著贡献。然而，目前尚无适用于视频输入的有效裁剪算法。若仅将视频帧作为独立图片处理，可能会导致含糊不清的语义表达。在本项目中，我们提出了MAGICLICK这一新模型，它允许用户选择突出的主题，并通过适当的分割模型，提供用户选择裁剪比例的选项，或由模型自动决定。我们采用了一种创新的基于动态规划的裁剪方法，结合了不同裁剪结果的视频帧重要性采样，并通过Neural Image Assessment（NIMA）模型（Talebi等，2018年）进行美学评估，让用户轻松选择心仪的裁剪视频。

我们的主要工作包括：

1. 以视频为输入，在MAGICROP框架基础上，结合用户的即时提示（通常为一次Click），使用Track Anything Model（Yang等，2023年）作为分割模型，实现用户所选视频主体的突出显示和实时跟踪。
2. 采用基于动态规划的裁剪方法，其评估函数基于视频主体的变化幅度。我们的算法全面考虑视频主体的突出效果和画面变化，实现在指定裁剪比例和中心定位下的最佳裁剪结果。
3. 提出了一种基于视频帧变化幅度的重要性采样算法，利用NIMA模型对加权视频帧进行美学评估，为用户的选择提供分数指导。

2.Related work

1. MAGRICROP

MAGICROP是一个图像裁剪系统，专为新手用户设计，以帮助他们制作艺术化的人像照片。MAGICROP结合深度学习和摄影构图规则，提供了一个直观的用户界面，使用户能够轻松地进行裁剪操作。系统包括两个主要模块：灵感模块和裁剪模块。灵感模块帮助用户发现图片的裁剪潜力，而裁剪模块则允许用户调整自动裁剪过程的参数，生成个性化的裁剪结果。

研究者进行了形成性研究，包括对新手用户、商业摄影师和摄影专家的深入访谈，以理解裁剪过程中的困难和需求。基于这些洞见，开发了MAGICROP系统。该系统通过其独特的裁剪方法，结合显著性对象检测、裁剪候选项生成和美学质量评估，提供了个性化的裁剪结果，特别适用于人像摄影。

此外，研究者还进行了用户研究，包括问卷调查和专注小组访谈，以验证MAGICROP的有效性和易用性。结果表明，该系统对于缺乏图像编辑知识的新手用户尤其有用，提高了他们对裁剪结果的满意度，并且用户界面简单直观，易于学习和使用。

2. TAM

TAM（Tracking Anything Model）是一种视频分析和对象跟踪的模型，设计用于在动态环境中准确跟踪各种对象，包括罕见或新颖的对象。它适用于安全监控、运动分析、自动驾驶车辆感知等多种应用。TAM的关键优势包括其能够识别和跟踪多样的对象、高效的实时处理能力、适应动态环境变化的能力以及对挑战性条件（如遮挡、光照变化）的鲁棒性。技术上，它结合了ViT等深度学习技术，并可能使用先进算法提高跟踪的准确性和稳定性。TAM在处理复杂、动态环境中的实时跟踪任务方面具有重要价值。

3. NIMA 模型

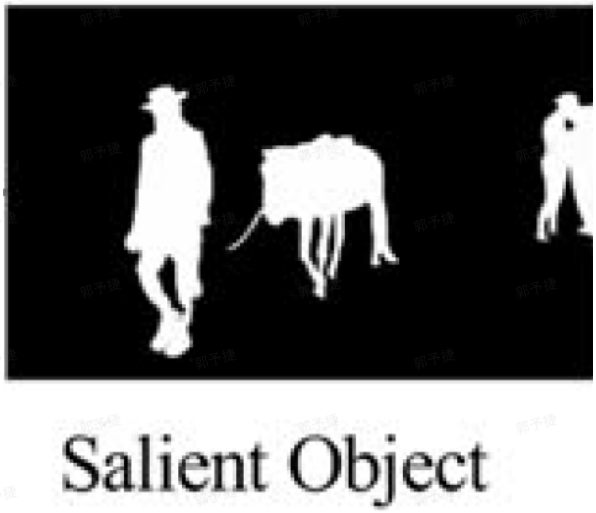
这是一种新颖的利用卷积神经网络（CNN）进行图像质量和审美评估的方法。与传统的平均意见得分（MOS）预测方法不同，NIMA评估了人类对图像质量感知的分布，从而提供了更全面的理解。该模型采用了最先进的深度目标识别网络作为其架构，这些网络被重新训练以适应图像质量评估任务。与更复杂的方法相比，这种架构更为简单但同样有效。NIMA能够在不需要参考图像的情况下评估单幅图像，这是一种无参考质量评估方法。它不仅能够评估图像的技术质量，如噪声、模糊和压缩伪影，还能评估图像的审美质量，这包括更主观的方面，如艺术性和情感吸引力。此模型在AVA（美学视觉分析）和TID2013等大规模数据集上进行了训练，这些数据集包含了带有人类评分的图像。训练过程包括对这些带注释的数据集进行微调，以完成感知质量评估任务。NIMA模型的应用不仅限于评估图像质量，它还有助于调整和优化照片编辑和增强算法，为图像处理流程的各种应用做出贡献。总体来看，NIMA在自动图像质量评估方面实现了重大进步，结合了深度学习的能力和对人类感知图像质量的更细腻理解。

3.Motivation

1. 我们注意到，在已有的MAGICROP中，用户无法自行指定需要突出的主体。例如输入图如下时：

如果我们希望的拍摄主体为画面中心的黑马，MAGICROP算法不能给出指定的裁切结果，而是选择将人与马共同作为画面主体而进行决策。而使用成熟的分割模型，例如SegAnything Model, 通过用户手动给出prompt(一次简单的click), 这样则可以很好解决该问题。

通过点击中心（蓝色点）选择马，马所在的区域将会被分割

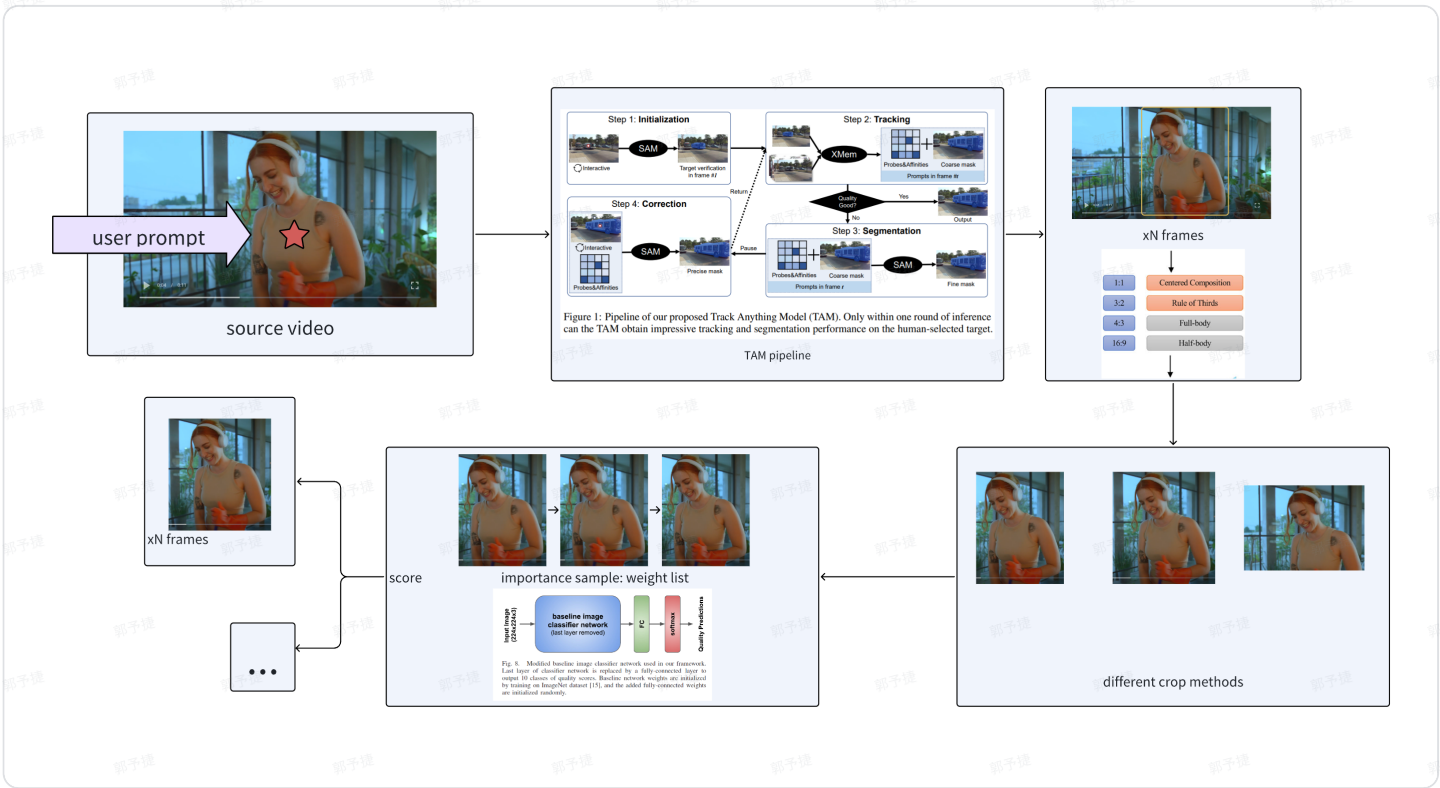


MAGICROP采用的salient object detection将选择所有主体

2. 同时，我们希望输入不只是静态的图片，而可以接受动态视频的输入。当拍摄主体始终在画幅内时，镜头应跟踪并裁切它。同时，应该兼顾效率与稳定性，即逐帧的视频裁切，不应该出现巨大的主体晃动，应该尽可能平滑稳定，模拟专业摄影师的平稳运镜过程。
3. 在使用现有的成熟美学评估算法，例如NIMA模型时，我们考虑到视频主体的不同变化幅度对结果的评估应该有不同的影响。举例来说，例如一个主体长时间静止不动，在考虑这一主体对视频美学贡献时，它所占的权重应该相对较小，这样可以更好的平衡运动物体与静止物体间的相对关系。

4.Design Overview

如下是MAGICLICK系统的设计流程



首先，接收用户的click prompt, 使用Track Anything Model 的pipeline进行主体的追踪与分割。用户可以选择裁切比例，分为 `1:1, 3:2, 4:3, 16:9` 四种裁切方式，居中方式有黄金分割（三分法则）或中心法则。在确定了裁切方式后，MAGICLICK根据用户选择的主体，以动态规划算法[5.1节]完成对应的裁切。将不同的裁切方式分别生成的帧序列，经过重要性采样后通过评估模型进行打分，最后返回最佳的top K裁切结果。

5. Design detail

下面，我们将展开介绍模型的实现细节，帮助读者更好的理解MAGICROP系统。

5.1 Dynamic-programming-based cropping

通过TAM分割模型在pipeline中的应用，我们能够获取每个视频帧中用户所选主体的边界框（bounding box）。针对每个视频帧及其对应的边界框，我们采用用户提供的比例进行扩散算法计算，从而得到一系列可选的裁切区域，表示为[x1, y1, x2, y2]。基于这些信息，我们为每个视频帧的不同裁切方案创建节点，每个节点包括裁切区域的属性（propertity）、边界框信息（bounding box），以及计算得到的比例变化（alpha）和锚点位移（theta）。节点还包含一个列表（befores），用于存储与其相关的前向节点及其距离。

为了最小化相邻帧之间的画面变化幅度，我们采用动态规划算法，将每个视频帧的不同裁切方案视为一个点集，并按照时序顺序连接各个点集中的点以找到最短路径。两点之间的距离由以下公式计算：

$$\mathrm{dist}(\mathrm{a},\mathrm{b}) = \left|\frac{\mathrm{a}.\alpha_0}{\mathrm{b}.\alpha_0}\right| \cdot |\mathrm{a}.\theta_0 - \mathrm{b}.\theta_0| + \left|\frac{\mathrm{a}.\alpha_1}{\mathrm{b}.\alpha_1}\right| \cdot |\mathrm{a}.\theta_1 - \mathrm{b}.\theta_1|$$

其中alpha_0和alpha_1表示x轴和y轴的比例变化，theta_0和theta_1表示x轴和y轴的锚点位移。

通过连接相邻帧的点集，我们形成了一个有向图，并依据上述公式计算边的长度。随后，采用动态规划的思想计算从起始帧到末尾帧的最短路径。状态转移公式如下：

$$\mathrm{dp}[i][j] = \min(\mathrm{dp}[i-1][k] + dist((i,j),(i-1,k)), dp[i][j]), \forall k \in points(i-1)$$

其中，dp[i][j] 表示从起始帧到第 i 帧的第 j 个裁切所需的最短距离。通过此方法，我们可以获得在画面变化幅度最小化条件下的多个裁切方案。

5.2 Video Assessment

将一个裁切后的视频进行评分，首先要将其分离成逐帧的形式。我们需要考虑各个帧的重要性。视频主体在大尺度变动的视频帧，对整体视频美感的重要性贡献是更大的，因为长时间静止不动的主体，并不影响整体视频的美感，而主体在大尺度运动时，作为观察者，我们更容易注意并捕捉这些变化帧。

我们将bounding box的变化幅度做一次差分，构建出变化的差分数组。将这一数组归一化处理得到p_i. 有:

$$\sum_{i=1}^{N} p_i = 1$$

现在，我们已经可以得到各个帧的重要性权重，可以根据这一权重对视频帧序列进行采样。不选择将所有视频帧逐个进行美学评估的重要原因是，审美模型的运行需要一定的时间开销，这样输入一个长序列视频帧，用户将等待相当长的一段时间，不利于用户体验。

采样过程为：

$$\int f(x)\,dx \approx \frac{1}{N}\sum_{i=1}^{N}\frac{c_i * N(f_i)}{diff(f_i)}$$

其中， f_i 为第i个视频帧。c_i取值为0或1,有：

$$c_i = \begin{cases} 1 & \text{if } p_i < rand, \\ 0 & \text{otherwise.} \end{cases}$$

diff(f_i)为差分数组的第i项。N(f_i)为视频帧f_i经由NIMA模型得到的评估得分。

6. Conclusion

在本文中，我们详细介绍了MAGICLICK，一种创新的视频裁剪系统，它在现有的MAGICROP图像裁剪系统基础上进行了显著的扩展和改进。MAGICLICK的主要创新之处在于其对视频内容的支持，使用户能够通过简单的点击交互来选择视频中的主要主题，系统则自动完成裁剪过程。通过结合动态规划裁剪方法、Track Anything Model以及基于视频帧重要性采样的美学评估，MAGICLICK不仅提高了视频裁剪的效率和质量，还极大地简化了非专业用户的视频编辑过程。尽管当前系统已经表现出色，但未来仍有进一步优化算法性能和扩展功能的空间，以支持更广泛的视频处理应用。MAGICLICK代表了视频裁剪技术的一大步进，为视频编辑领域提供了一种新的、高效的解决方案。