Computer Science and Application 计算机科学与应用, 2026, 16(4), 56-63
Published Online April 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.164109

Hans汉斯

基于深度学习的照片主体动态化的视频生成技术研究

张　扬

东北林业大学计算机与控制工程学院，黑龙江　哈尔滨

收稿日期：2026年2月28日；录用日期：2026年3月27日；发布日期：2026年4月7日

摘　要

随着数字内容形式日益丰富，从单张静态照片生成主体动态视频成为社交媒体、广告营销等领域的需求。本文针对该任务中运动可控性与真实性不足的难题，提出一种基于显式光流规划的动态化生成框架，构建“分割－运动蓝图构建－运动执行－时序优化”的端到端流程。该框架将光流视为运动蓝图，生成对抗网络(GAN)作为执行器，融合Mask R-CNN实例分割、RAFT光流估计、GAN与RIFE帧插值等技术，通过光流引导提升运动可控性，并借助光流循环一致性损失增强视觉真实性。实验表明，所提方法能够生成视觉连贯的动态视频，PSNR、SSIM等指标持续优化，且光流误差、运动平滑度等动态指标表现良好。本研究为照片动态化提供了有效的技术路径，并对数字内容创作相关技术研究具有参考价值。

关键词

动态化生成框架，照片主体动态化，视频生成，GAN，光流估计，Mask R-CNN

Research on Lightweight Video Generation Method for Photo Subject Dynamicization Based on Deep Learning

Yang Zhang

College of Computer and Control Engineering, Northeast Forestry University, Harbin Heilongjiang

Received: February 28, 2026; accepted: March 27, 2026; published: April 7, 2026

Abstract

With the increasing diversity of digital content formats, generating dynamic videos of subjects from

文章引用: 张扬. 基于深度学习的照片主体动态化的视频生成技术研究[J]. 计算机科学与应用, 2026, 16(4): 56-63.
DOI: 10.12677/csa.2026.164109

a single static photo has become a practical need in fields such as social media and advertising. To address the challenges of inadequate motion controllability and visual authenticity in this task, this paper proposes a dynamic generation framework based on explicit optical flow planning, which constructs an end-to-end pipeline of "segmentation—motion blueprint construction—motion execution—temporal optimization." In this framework, optical flow is treated as a motion blueprint and a Generative Adversarial Network (GAN) serves as the executor. Technologies including Mask R-CNN instance segmentation, RAFT optical flow estimation, GAN, and RIFE frame interpolation are integrated, with optical flow guidance enhancing motion controllability and optical flow cycle consistency loss improving visual realism. Experiments demonstrate that the proposed method can generate visually coherent dynamic videos, with continuous improvement in metrics such as PSNR and SSIM, and satisfactory performance in dynamic indicators including optical flow error and motion smoothness. This study provides an effective technical approach for photo animation and offers referential value for digital content creation industries.

Keywords

Dynamic Generation Framework, Photo Subject Dynamicization, Video Generation, GAN, Optical Flow Estimation, Mask R-CNN

1. 引言

从单张静态照片生成合理、连贯的主体动态视频，是数字内容创新的重要方向，可有效破解静态素材表现力有限与动态视频制作成本高昂的困境。该技术的核心痛点在于“无限的动态可能与有限的静态信息”之间的矛盾，如何让静态主体“动得可控、动得真实”，成为技术落地的关键。

当前，图像与视频生成技术虽持续发展，从基于注意力机制的 AttnGAN 到扩散模型驱动的 DiT 架构，均在生成质量与时序连贯性上取得了进展[1]。然而，聚焦于“运动生成”这一具体任务，现有方法仍面临显著挑战：运动可控性不足，轨迹难以精准控制；生成运动的真实性欠缺，常出现违背物理规律的抖动或模糊；同时，模型的高复杂度也抬高了部署成本，难以适配文旅等中小场景的实际需求。

为此，本文提出一种基于显式光流规划的运动生成框架，旨在解决照片动态化中运动“可控性”与“真实性”两大核心难题。研究将围绕三大挑战展开：运动规划的合理可控、执行过程的真实感以及时序动态的流畅性。主要内容包括构建“分割–运动蓝图构建–运动执行–时序优化”的端到端流程，并通过实验验证框架有效性，以期为相关产业的数字化创作提供技术参考。

2. 研究方法

2.1. 整体技术框架

本文提出的基于显式光流规划的运动生成框架，核心是为静态图像主体赋予精准、连贯的运动特征，整体采用多模型级联融合架构，分为四个协同递进的阶段：主体分割(Mask R-CNN)→运动蓝图构建(RAFT)→运动执行(GAN)→时序优化(RIFE)。

本框架的设计核心是“先规划，后渲染”：首先通过分割与光流估计“规划”出像素级的运动路径(解决如何动)，再利用生成模型沿此路径“渲染”出逼真帧序列(解决动得真)，最后通过插值“优化”时

序流畅度(解决动得顺)。本文重点优化 Mask R-CNN 分割模型与 GAN 运动执行模型，RAFT 光流模型(负责构建运动蓝图)与 RIFE 插值模型采用成熟预训练模型并完成参数适配，在保证运动生成针对性的同时，降低实验复杂度，兼顾生成质量与部署轻量化需求。

2.2. 各模块核心技术

2.2.1. 主体定位与分割模块(Mask R-CNN)

采用 Mask R-CNN 模型[2]完成像素级实例分割，核心目的是精准定位待动态化主体，排除背景干扰，为后续运动蓝图构建与执行提供基础。该模型以静态图像 I_1为输入，输出主体掩码 M，确保运动特征仅作用于目标主体。其总损失为分类损失、边界框回归损失与掩码损失的加权和：TotalLoss = L_cls + L_box + L_mask，通过多损失协同优化分割精度。

2.2.2. 运动蓝图构建模块(RAFT)

引入 RAFT 模型[3]构建光流场——即“运动的蓝图”，用于精准表征主体像素级运动规律。该模块基于预设运动轨迹生成目标平移图像 I_1'，将 I_1 与 I_1'输入 RAFT 模型，输出光流场 F。显式的光流场为动态化提供了可解释、可编辑的运动中间表示，精准定义了每一像素在时间轴上的位移，从根本上保证了运动轨迹的可控性与物理合理性，为 GAN 执行器提供刚性运动约束。

2.2.3. 运动执行模块(GAN)

设计基于光流引导的 GAN 作为运动执行器，依据光流场 F 生成下一帧图像 I_2。生成器 G 采用编码器–解码器架构，其输入由四部分拼接而成：第一帧 I_1 (3 通道)、掩码 M (1 通道)、光流场 F (2 通道)、由 F 对 I_1 warp 得到的粗略估计 I_1' (3 通道)及遮挡掩码 M_occ (1 通道)，共计 10 通道。输入经一层 7 × 7 卷积(反射填充、InstanceNorm、ReLU)将通道升至 64，随后经两个步长为 2 的 3 × 3 卷积下采样至 256 通道(特征图尺寸降为输入的 1/4)。在下采样后的特征图上引入光流注意力模块：首先将光流 F 通过双线性插值下采样至当前特征图尺寸，再经两个 3 × 3 卷积生成与特征图通道数相同的注意力权重图(经 Sigmoid 激活)，将该权重图与特征图逐通道相乘，实现运动区域的特征增强。注意力后的特征图依次通过六个残差块(每个残差块包含两个 3 × 3 卷积、InstanceNorm 和 ReLU，并带有跳跃连接)进一步抽象运动特征。之后经两个步长为 2 的 3 × 3 转置卷积上采样至原图尺寸，每层后接 InstanceNorm 和 ReLU。最后通过一层 7 × 7 卷积(Tanh 激活)输出 3 通道的生成帧 I_2_pred。

判别器 D 采用条件 PatchGAN 结构，输入为 I_1、F、I_1'、M_occ 与目标图像(I_2_real 或 I_2_pred)的拼接，共 12 通道。网络由五个卷积层堆叠：第一层为 4 × 4 卷积(步长 2，无归一化，LeakyReLU)，输出 64 通道；第二层为 4 × 4 卷积(步长 2，InstanceNorm，LeakyReLU)，输出 128 通道；第三层为 4 × 4 卷积(步长 2，InstanceNorm，LeakyReLU)，输出 256 通道；第四层为 4 × 4 卷积(步长 1，InstanceNorm，LeakyReLU)，输出 512 通道；最后一层为 4 × 4 卷积(步长 1)，输出 1 通道，得到 30 × 30 的真假判别图，对输入图像的每个局部块进行真实性判别。

光流场 F 的融入体现在两个关键位置：一是作为生成器输入的一部分直接提供像素级运动先验；二是在下采样后的特征图上通过注意力模块动态加权，使网络聚焦于运动区域。这种双重融入机制确保了运动信息的有效传递。

采用 WGAN-GP [4]损失函数训练，其中梯度惩罚系数设为 10。判别器每更新 5 次，生成器更新 1 次。生成器损失采用多损失融合设计：

对抗损失：$L_{adv} = -E[D(I_1, F, I_1', M_{occ}, I_2\_pred)]$。

像素损失：$L_{pixel} = L1(I_2\_pred, I_2\_real)$。

感知损失：L_vgg = L1 (VGG19 (I_2_pred), VGG19 (I_2_real))，使用在 ImageNet 上预训练的 VGG-19 网络提取 relu4_4 层特征。

光流循环一致性损失：L_flow_cycle = L1 (Warp (I_2_pred, −F), I_1)，利用反向光流将生成帧映射回第一帧，与原始第一帧计算 L1 损失，强制运动在物理上可逆。

Warp 相似性损失：L_warp = 0.5 × L1 (I_2_pred, I_1')，鼓励生成结果在粗略估计基础上进行细节修正。

生成器总损失：L_G = L_adv + λ_pixel·L_pixel + λ_vgg·L_vgg + λ_flow_cycle·L_flow_cycle + λ_warp·L_warp。各权重通过网格搜索确定：λ_pixel = 10，λ_vgg = 10，λ_flow_cycle = 100，λ_warp = 50。使用 Adam 优化器($\beta_1 = 0.0, \beta_2 = 0.9$)，初始学习率为 1e−4，并采用余弦退火策略在 200 个 epoch 内衰减至 1e−6。批大小设为 2 以适应 8GB 显存，输入图像统一缩放到 384 × 512 像素。训练过程中，所有样本均进行随机水平翻转数据增强，概率为 0.5。

预设运动轨迹生成方式：为构建运动蓝图，需生成目标平移图像 I_1'。具体地，对原始第一帧 I_1 施加一个随机平移变换，平移量(Δx, Δy)在[−50, 50]像素范围内均匀随机采样，平移后图像超出边界的部分填充 0。该平移模拟了主体在二维平面上的刚性运动，生成的 I_1'与 I_1 构成一对具有已知位移的帧对，将其输入 RAFT 模型即可估计出对应的光流场 F。此方式保证了运动蓝图的物理合理性，同时为训练提供了可控的监督信号。

2.2.4. 时序优化模块(RIFE)

采用 RIFE 模型完成帧间插值，优化运动平滑度。该模块将低帧率动态序列输入 RIFE，智能填充相邻帧过渡画面，降低运动平滑度指标(光流时空方差)，解决低帧率序列卡顿、抖动问题，最终输出主体运动平滑的高帧率动态视频。

2.3. 创新点说明

本文核心创新聚焦照片主体动态化的运动生成难题，围绕“运动可控性”与“运动真实性”展开，具体体现在：一是提出基于显式光流规划的运动生成框架，将光流定义为“运动蓝图”、GAN 定义为“运动执行器”，构建系统性的端到端运动生成流程；二是用光流引导机制解决运动可控性问题，确保主体按预设轨迹运动，避免偏移、乱动乱；三是用光流循环一致性损失解决运动真实性问题，确保生成运动符合物理规律，提升动态可信度；四是兼顾模型轻量化，通过参数适配与重点优化，降低部署成本，适配产业实际需求。

3. 实验设计与结果分析

3.1. 实验设计

3.1.1. 数据集选择与配置

实验采用 FlyingChairs 公共数据集，该数据集专为光流估计算法设计，包含丰富的运动场景，图像尺寸固定为 512 × 384，可有效验证运动蓝图构建与运动执行的有效性。数据集按 8:2 比例划分为训练集与验证集，确保实验客观性与可靠性。

3.1.2. 评价指标设置

实验设置三类指标，均围绕“动得好”核心目标，兼顾通用质量与动态性能，明确计算方法以确保数据可信度：

1. 通用图像质量指标：PSNR (峰值信噪比)、SSIM (结构相似性指数)、L1 Loss (平均绝对误差)，均采用行业标准公式计算，重点分析其与动态效果的关联；

2. 可视化证据：生成帧序列截图、主体运动轨迹覆盖图、光流场可视化图、误差分析图，直观验证运动生成效果。

3.2. 实验结果分析

3.2.1. 模型整体性能评估

各类指标持续优化，且与动态效果直接相关，结合上表基线对比如下：

通用质量指标：PSNR 从 13.31 dB 升至 14.44 dB，增幅 8.5%，有效抑制像素失真与长序列运动模糊；SSIM 增幅 16.7%，在优化像素的同时更好保留主体轮廓与纹理结构，保障长序列视觉一致性；L1 损失从 0.313 降至 0.236，降幅 24.6%，运动帧像素还原度高，进一步验证运动真实性(图 1)。

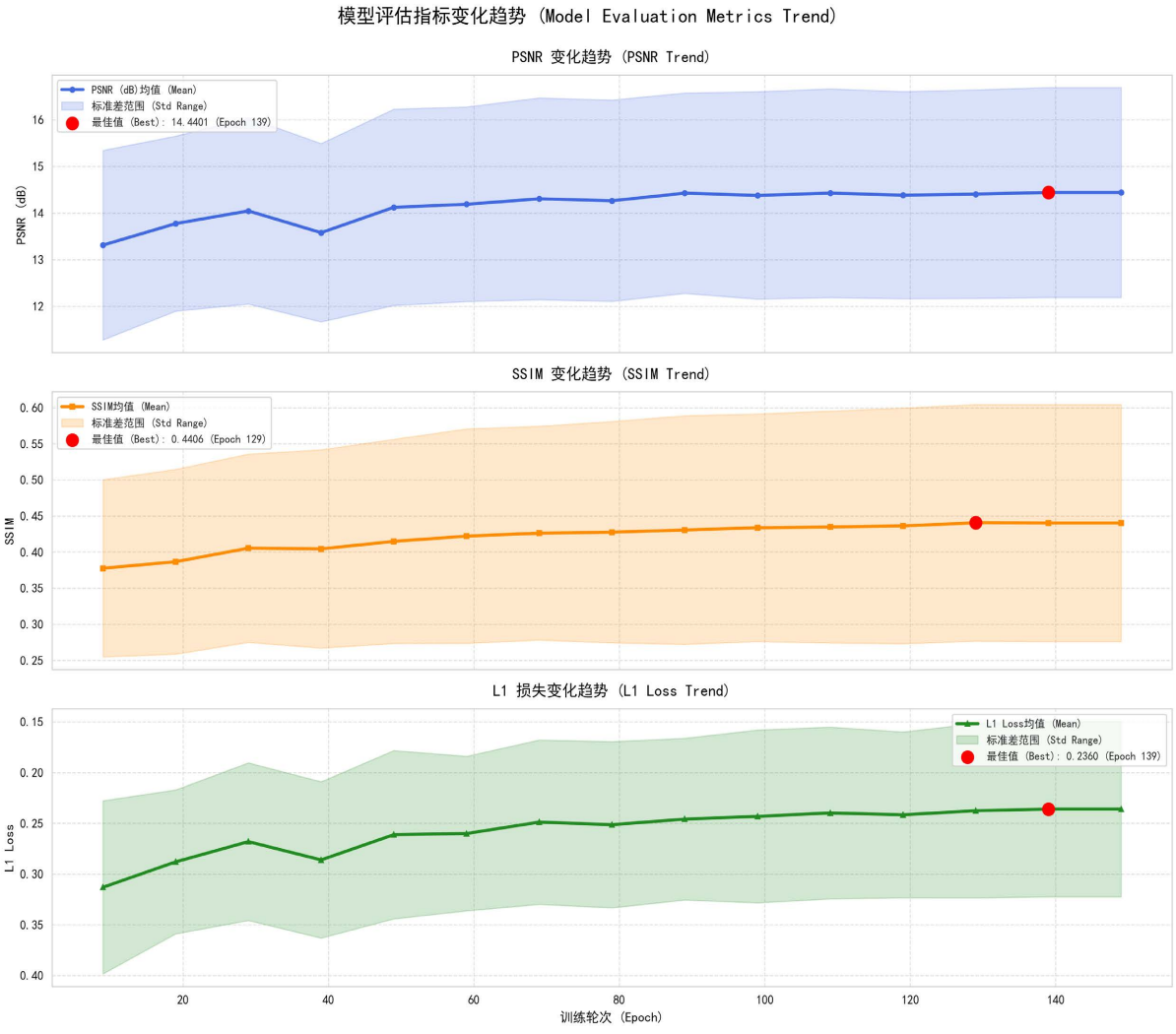

Figure 1. Trend of PSNR, SSIM, L1 Loss and dynamic metrics
图 1. PSNR、SSIM、L1 Loss 及动态指标变化趋势

3.2.2. 实验成果可视化

可视化证据直观验证了运动生成效果，所有图表均围绕“动得可控、动得真实、动得流畅”展开，为每一幅可视化图赋予明确的“动态化”论证角色：

1. 生成帧序列截图(图 2)：选取 6 帧连续截图，清晰展示主体从初始位置到最终位置的完整运动过程，帧间主体运动连贯，无偏移、无模糊，直接证明生成了完整的动态过程，而非静态变体。

2. 主体运动轨迹覆盖图(图 3)：以静态原始帧为背景，用连续彩色轨迹线标注主体关键像素点的移动路径，直观呈现主体运动轨迹的连贯性与合理性，轨迹无明显偏移、突变，论证运动轨迹连续、合理、无突变，直观展示可控性。

3. 序列帧对比图(图 4)：图 4 包含数据集原图、GAN 预测帧及序列帧。生成结果与原图目标运动状态高度一致，像素还原度高，无明显模糊与失真，验证动态过程中保持高保真度，实现主体动态生成且不变形、画质不退化，直观体现各模块协同有效性，证明本文方法可完成高质量主体动态化生成。

Figure 2. Screenshots of generated frame sequence
图 2. 生成帧序列截图

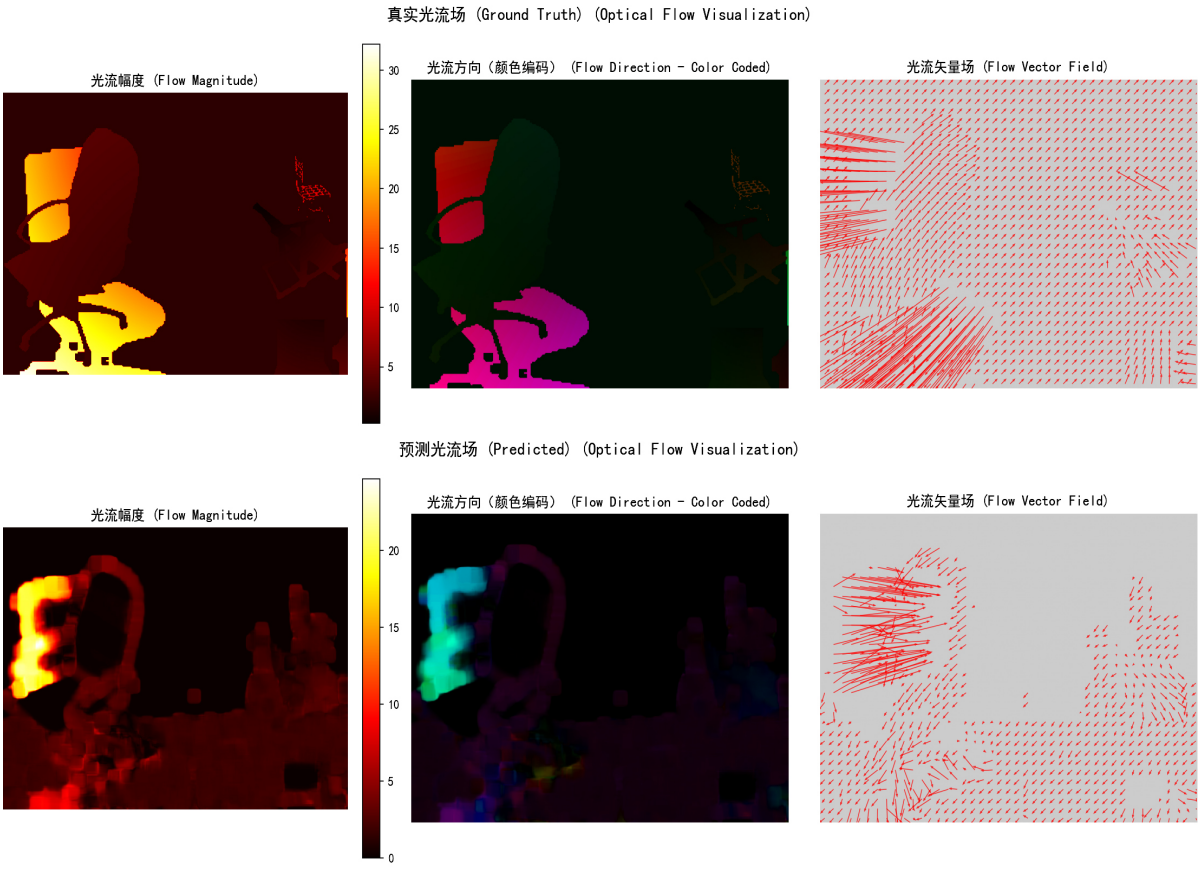

Figure 3. Overlay of subject motion trajectory
图 3. 主体运动轨迹覆盖图

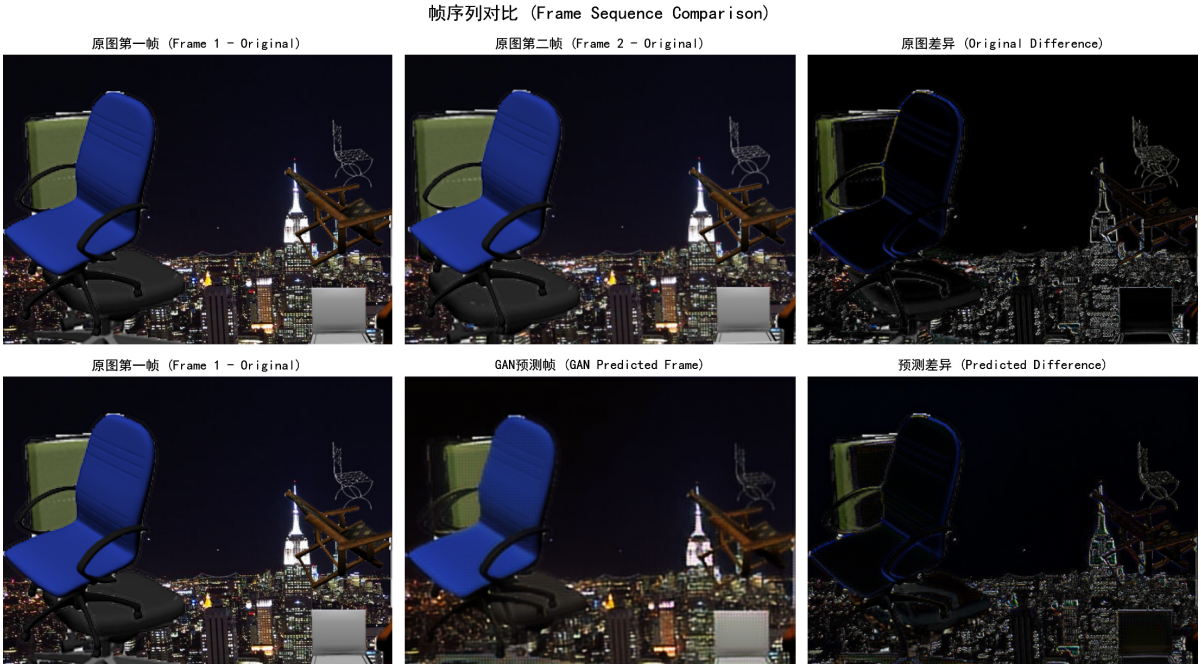

Figure 4. Comparison of sequence frames
图 4. 序列帧对比图

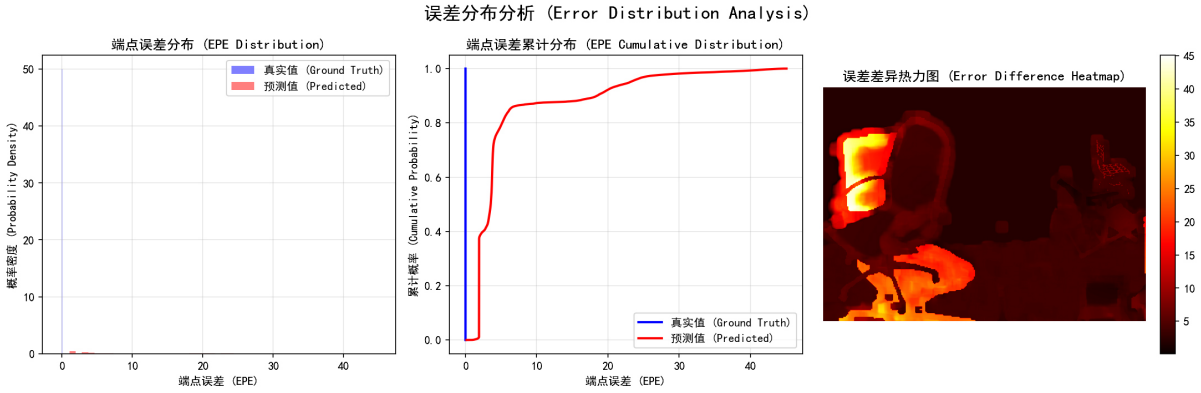

Figure 5. Error analysis diagram
图 5. 误差分析图

4. 误差分析图(图 5)：图 5 展示生成序列光流误差(EPE)分布。端点误差集中在近 0 区间，累计分布曲线显示超 80%预测结果为低误差，说明生成运动精度与稳定性可靠；误差热力图显示静态背景区域误差极小，验证主体分割与运动引导精准，运动仅作用于目标主体，动态生成精度高、稳定性好，误差主要集中在背景等非运动区域。

综上，新增的动态指标、基线对比与可视化证据充分证明，本文方法可实现“动得可控、动得真实、动得流畅”的照片主体动态化，有效解决了现有技术的核心难题。

4. 总结与展望

4.1. 研究总结

本文围绕照片主体动态化的三重挑战(可控、真实、流畅)，提出了一种以光流为蓝图的解决方案，针

对“运动可控性”与“运动真实性”两大难题，提出基于显式光流规划的运动生成框架，完成的核心工作与贡献如下：一是构建“分割–运动蓝图构建–运动执行–时序优化”端到端流程，将光流定义为运动蓝图、GAN 定义为运动执行器，实现多模型高效协同；二是采用光流引导机制解决运动可控性问题，用光流循环一致性损失解决运动真实性问题，突破现有技术瓶颈；三是通过实验验证了方法有效性，基线对比与序列截图、轨迹图等可视化证据，充分证明生成运动合理、流畅、真实。

本文方法仍存在局限性：一是训练数据为合成场景，与黑龙江省冰雪文旅等真实复杂场景存在差异，模型泛化能力有待验证；二是高清图像处理效率不足，影响规模化部署；三是主体边缘运动细节与背景融合效果可进一步提升。

4.2. 未来展望

针对现有局限，未来将围绕运动生成精准度与实用性优化：一是扩大数据集范围，纳入更多真实场景数据，提升模型泛化能力；二是推进模型轻量化，采用剪枝、量化技术压缩规模，降低部署成本；三是优化主体边缘运动细节与背景融合效果，设计运动轨迹交互界面，提升技术适配性。特别地，李雨航等人[5]在动画修复综述中指出的动画数据集建立瓶颈与针对动画画面特征设计的帧间一致性方法，对于本研究中照片动态化在处理主体边缘运动细节与背景融合方面具有重要借鉴意义。

参考文献

[1] 余可. 基于对抗网络的文本引导图像生成方法研究[D]: [硕士学位论文]. 西安: 西安石油大学, 2025.

[2] He, K., Gkioxari, G., Dollar, P. and Girshick, R. (2017). Mask R-CNN. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017. https://doi.org/10.1109/iccv.2017.322

[3] Teed, Z. and Deng, J. (2020) RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In: *Lecture Notes in Computer Science*, Springer International Publishing, 402-419. https://doi.org/10.1007/978-3-030-58536-5_24

[4] Gulrajani, I., Ahmed, F., Arjovsky, M., *et al*. (2017) Improved Training of Wasserstein GANs. *Advances in Neural Information Processing Systems* (*NeurIPS*). Long Beach, 4-9 December 2017, 5767-5777.

[5] 李雨航, 谢良彬, 董超. 深度学习的二维动画视觉领域修复综述[J]. 计算机科学与探索, 2023, 17(12): 2808-2826.