北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

计算机科学与技术前沿

——图像处理

汇报人：董静涛 王梓臣 刘勇奇 王禹桥 黄梓颖

时　间：2024-11-18

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

1

基于事件相机的典型应用——眼动分析

事件相机

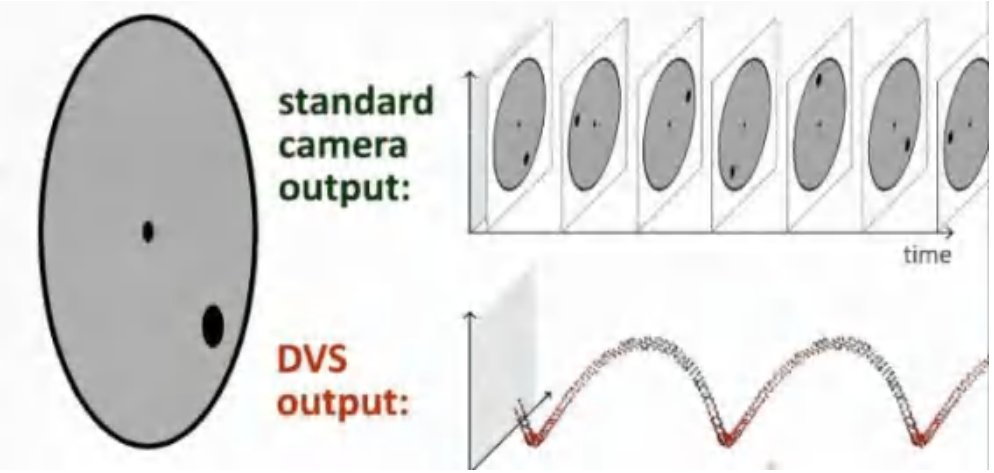
简介

- 我们把事件相机称为一种bio-inspired sensor，研究它的领域称为bio-inspired vision，即“仿生的”。制作事件相机的人的灵感就是来自于仿生。事件相机制作的目的就是：**敏感地捕捉到运动的物体。**
- 在传统的视觉领域，相机传回的信息是同步的，所谓同步，就是在某一时刻t，相机会进行曝光，把这一时刻所有的像素填在一个矩阵里回传，一张照片就诞生了。
- 一张照片上所有的像素都对应着同一时刻。至于视频，不过是很多帧的图片，相邻图片间的时间间隔可大可小，这便是我们常说的帧率（frame rate），也称为时延（time latency）。

事件相机

事件相机

- 逐像素响应亮度变化
- 异步运行
- 帧速率高（微秒级）
- 动态范围高

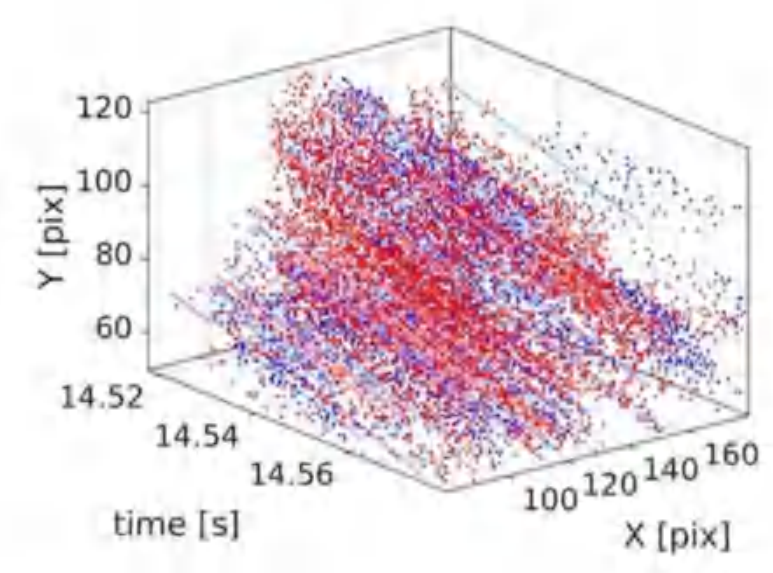


事件相机

每个事件数据（event）可以表示如下：

$$\mathbf{e} = \{t, x, y, p\}$$

其中包含事件发生位置的 x，y 坐标、时间 t 和亮度变化的 1 位极性 p（即亮度增加（“ON”，正事件）或亮度减少（“OFF”，负事件）。



介绍

- 人眼是一个动态的器官，能够执行快速和精确的运动，反映潜在的神经过程和认知状态。眼动分析一直是临床医生和各个学科的研究人员感兴趣的一个重要领域，包括心理学、神经病学和眼科学。
- 细粒度眼动分析可以启用或改进现实世界的应用程序，如虚拟现实、注意力估计、嗜睡检测、诊断等等。

对比

传统相机

标准RGB相机的高延迟导致事件发生和检测之间的延迟，在给定的时间范围内未能捕获足够的数据点，导致关键信息的丢失，特别是在涉及快速和复杂运动的情况下。因此，传统的相机往往不能提供准确和详细的眼球运动分析所必需的时间和空间分辨率。

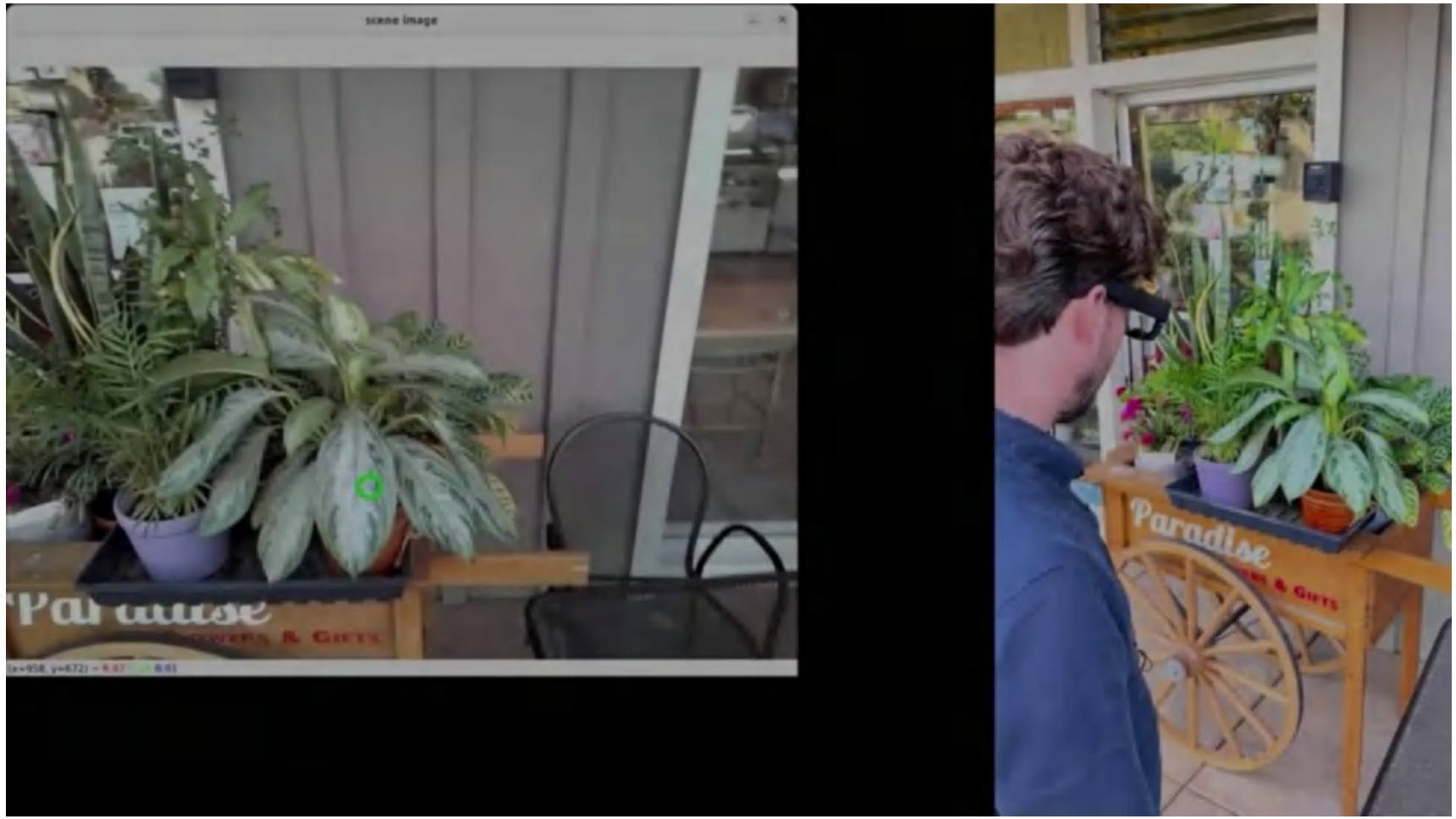
事件相机

与以固定间隔捕捉帧的传统相机不同，它通过检测每个像素的场景变化来工作，从而产生连续的事件流，而不是离散的帧。能够以高度的时间精度检测眼睛位置的微小变化。这种独特的操作机制使事件相机能够实现更低的延迟，使它们能够实时捕捉快速的眼球运动。

当前应用


Glint illuminator assemblies
World-facing camera
Microphone location
Eye-tracking cameras
Speaker location

当前应用

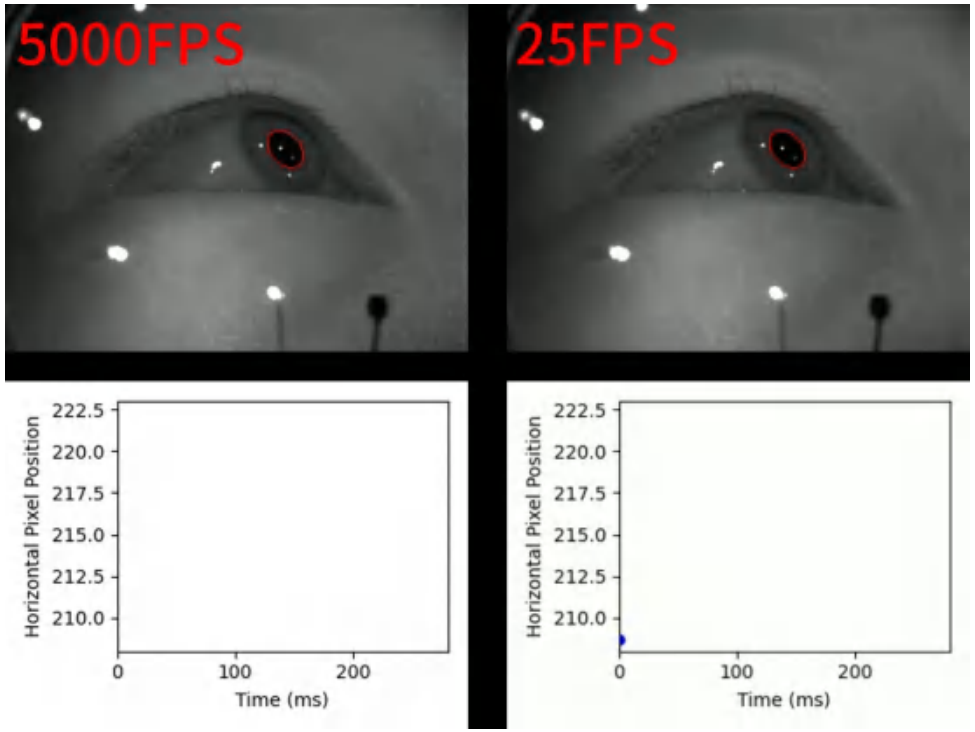


介绍

- Swift-Eye一文提出了一种基于事件相机的**高精度、高频**眼部运动追踪框架，旨在支持虚拟现实（VR）等领域中的高精度眼动分析。该框架通过融合事件相机的高时间分辨率优势和深度学习算法，实现了对眼部细微运动的精确捕捉，特别是在处理眨眼遮挡问题时表现出色。

创新点

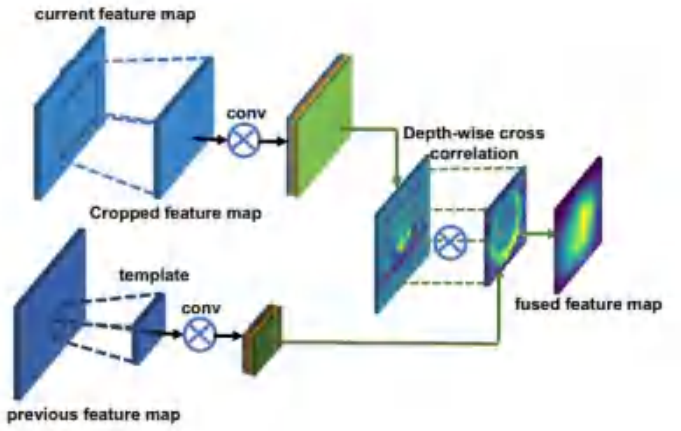
- **采用了一种基于事件驱动的视频插值方法（Timelens），将低帧率视频和高时间分辨率的事件流转换为高帧率视频，从而支持高频眼动的可视化和分析。**
- 特别设计了组件来处理由于不自主眨眼引起的瞳孔区域遮挡问题，提供了连续且平滑的瞳孔追踪轨迹。
- 凭借精确高频瞳孔轨迹，为心理健康诊断等高频眼动分析的潜在应用提供了支持。

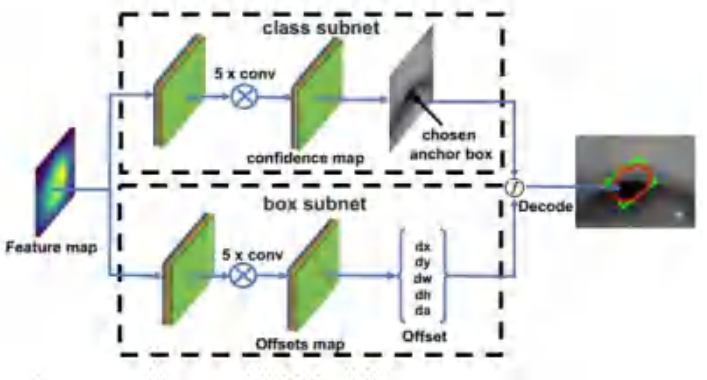


Swift-Eye: 面向抗眨眼瞳孔追踪的精确且鲁棒的高频近眼运动分析的事件相机应用

创新点

- 采用了一种基于事件驱动的视频插值方法（Timelens），将低帧率视频和高时间分辨率的事件流转换为高帧率视频，从而支持高频眼动的可视化和分析。
- **特别设计了组件来处理由于不自主眨眼引起的瞳孔区域遮挡问题，提供了连续且平滑的瞳孔追踪轨迹。**
- 凭借精确高频瞳孔轨迹，为心理健康诊断等高频眼动分析的潜在应用提供了支持。

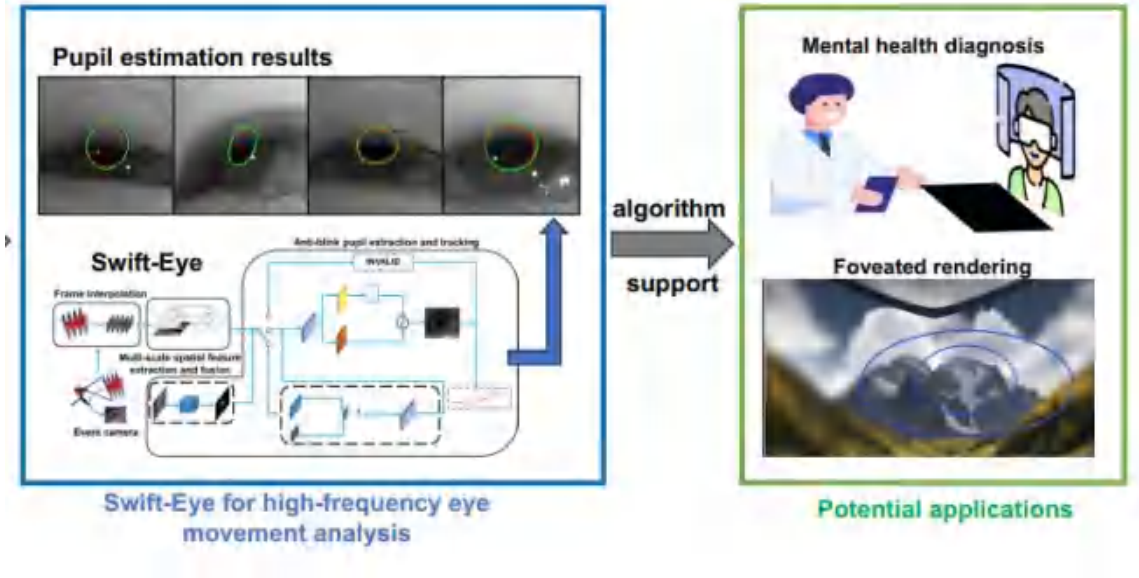




Swift-Eye: 面向抗眨眼瞳孔追踪的精确且鲁棒的高频近眼运动分析的事件相机应用

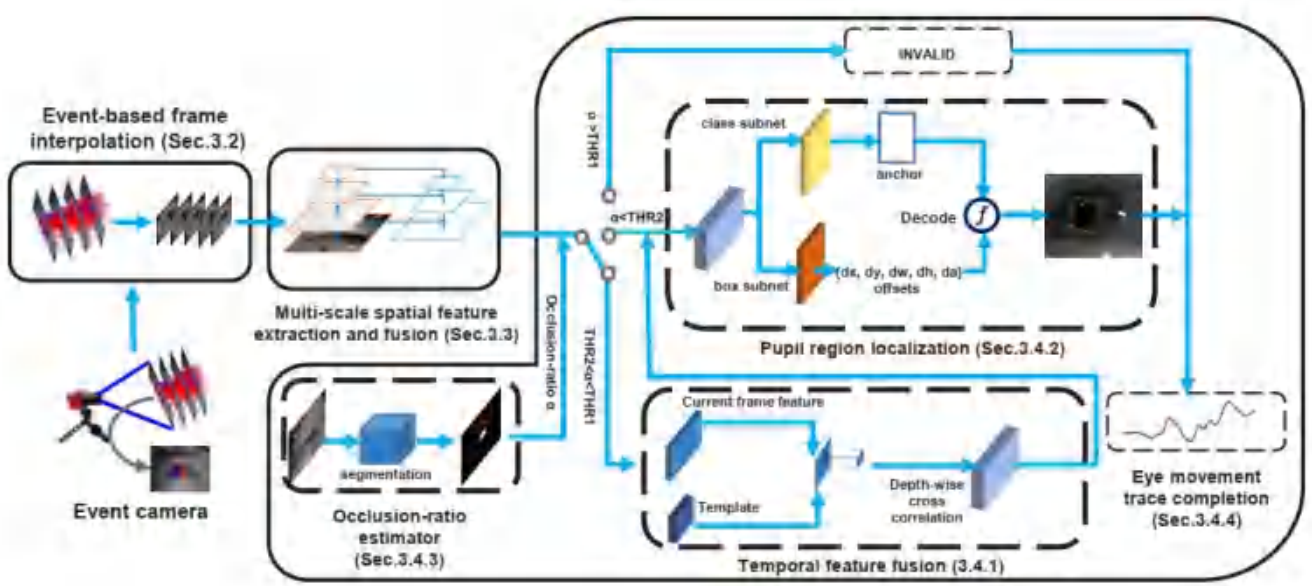
创新点

- 采用了一种基于事件驱动的视频插值方法（Timelens），将低帧率视频和高时间分辨率的事件流转换为高帧率视频，从而支持高频眼动的可视化和分析。
- 特别设计了组件来处理由于不自主眨眼引起的瞳孔区域遮挡问题，提供了连续且平滑的瞳孔追踪轨迹。
- **凭借精确高频瞳孔轨迹，为心理健康诊断等高频眼动分析的潜在应用提供了支持。**



系统架构

- **事件基础帧插值模块**：使用Timelens事件基础视频插值方法，将低帧率视频（例如25fps）和异步事件流转换为高帧率视频（例如5000fps）。
- **多尺度空间特征提取和融合模块：**使用Swin Transformer作为从不同尺度提取特征，再通过特征金字塔网络融合多尺度空间特征图，结合了底层的详细特征和高层的语义信息。
- **抗眨眼瞳孔估计和跟踪模块**：包括时间特征融合组件、旋转瞳孔区域检测器、遮挡比率估计器。通过遮挡比率估计器，根据遮挡比率自适应地切换不同的检测策略，以实现最佳准确性。



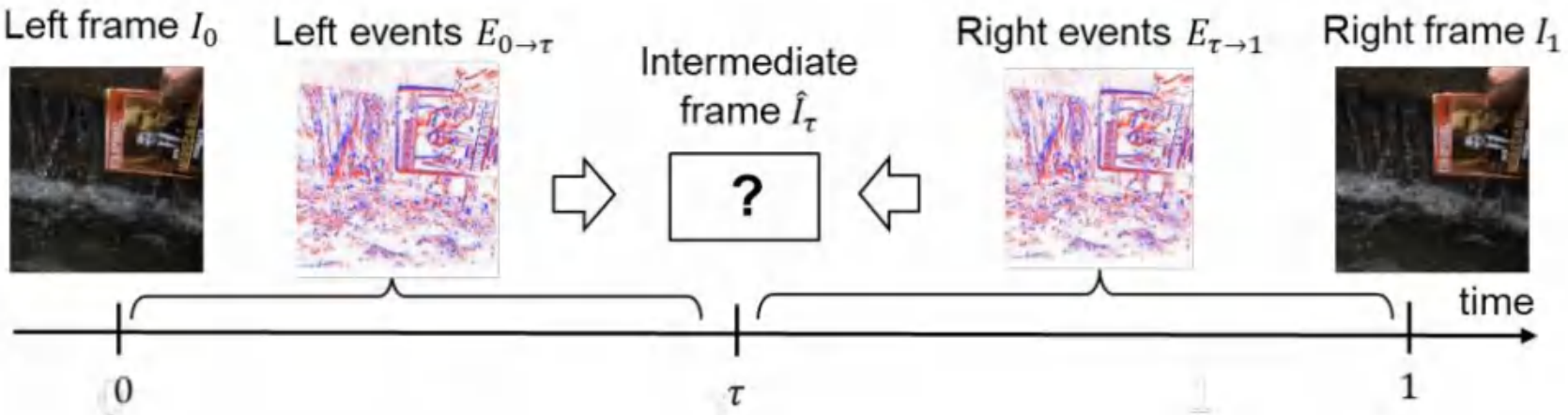
Timelens

Swift-Eye: 面向抗眨眼瞳孔追踪的精确且鲁棒的高频近眼运动分析的事件相机应用

Tulyakov et al., TimeLens, (2021)

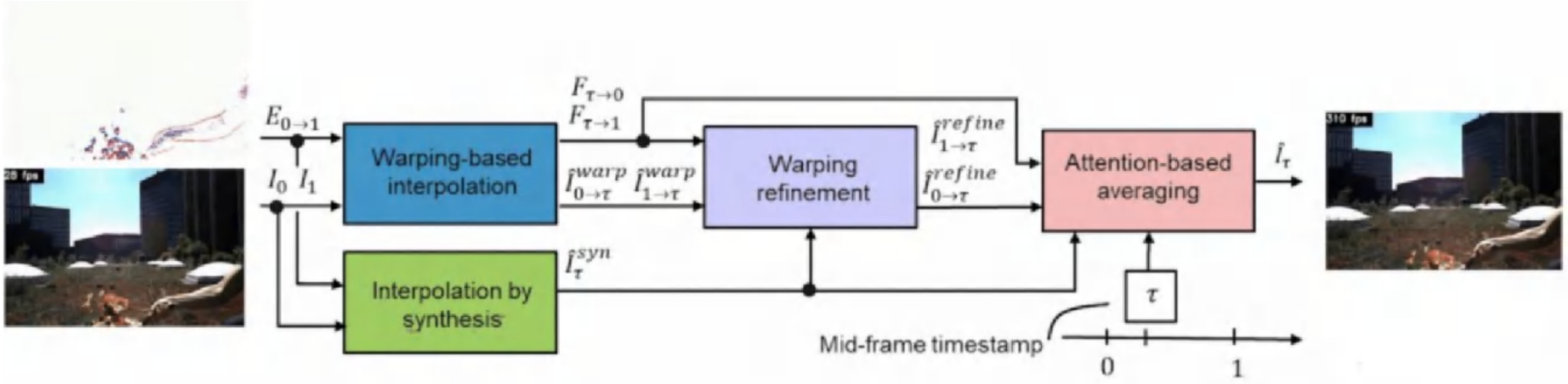
将左侧I_0和右侧I_1 RGB关键帧，以及左侧$E_{0\to\tau}$和右侧$E_{\tau\to1}$事件序列作为输入，目标是在关键帧之间以随机时间步长τ插入（一个或多个）新帧 $\hat{I}_\tau$。

事件序列（$E_{0\to\tau}$, $E_{\tau\to1}$）包含从各自（左I_0或右I_1）关键RGB帧同步采样的时刻触发的所有异步事件，直到我们想要插入新帧的时间步长τ。



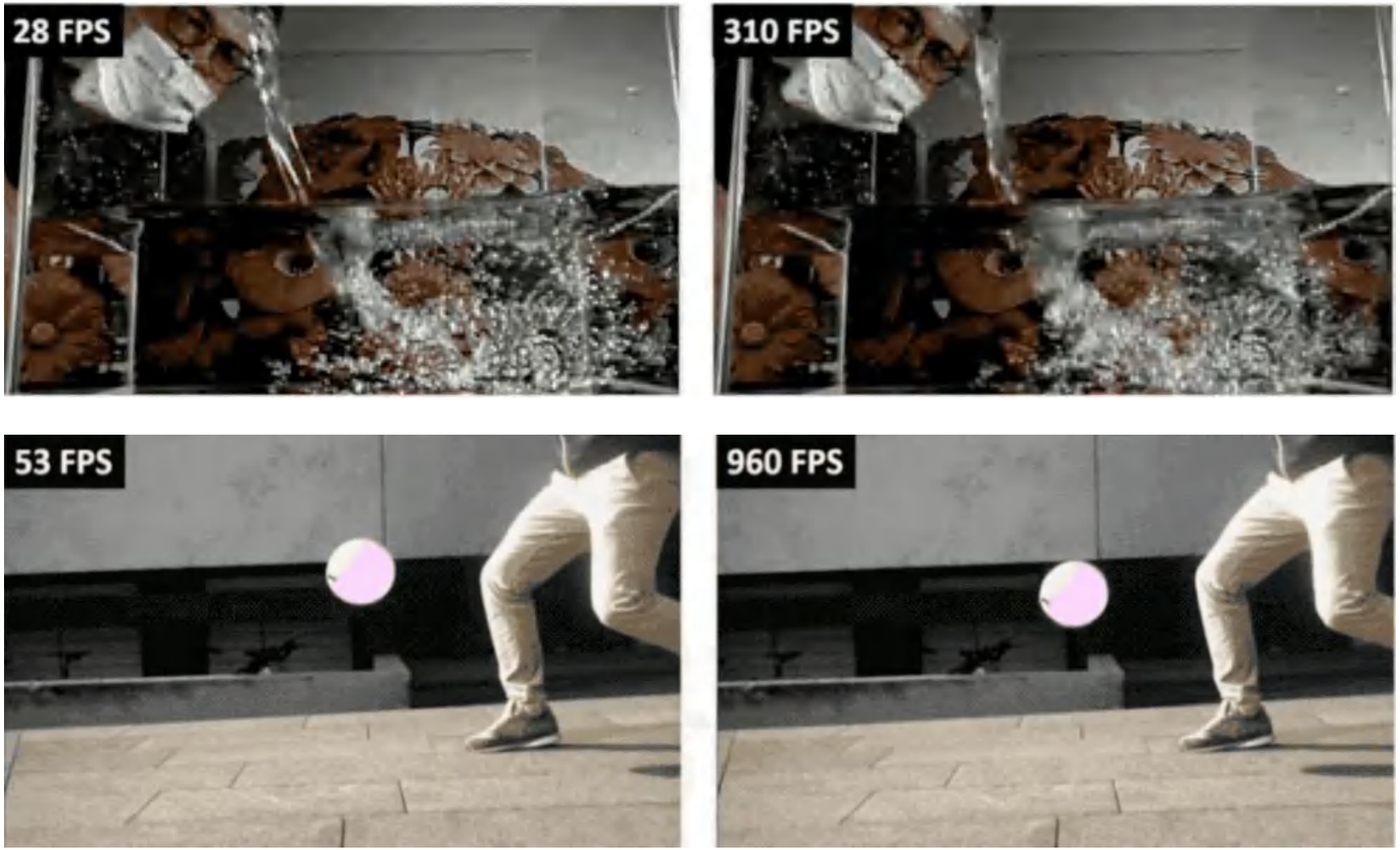
Timelens

Swift-Eye: 面向抗眨眼瞳孔追踪的精确且鲁棒的高频近眼运动分析的事件相机应用

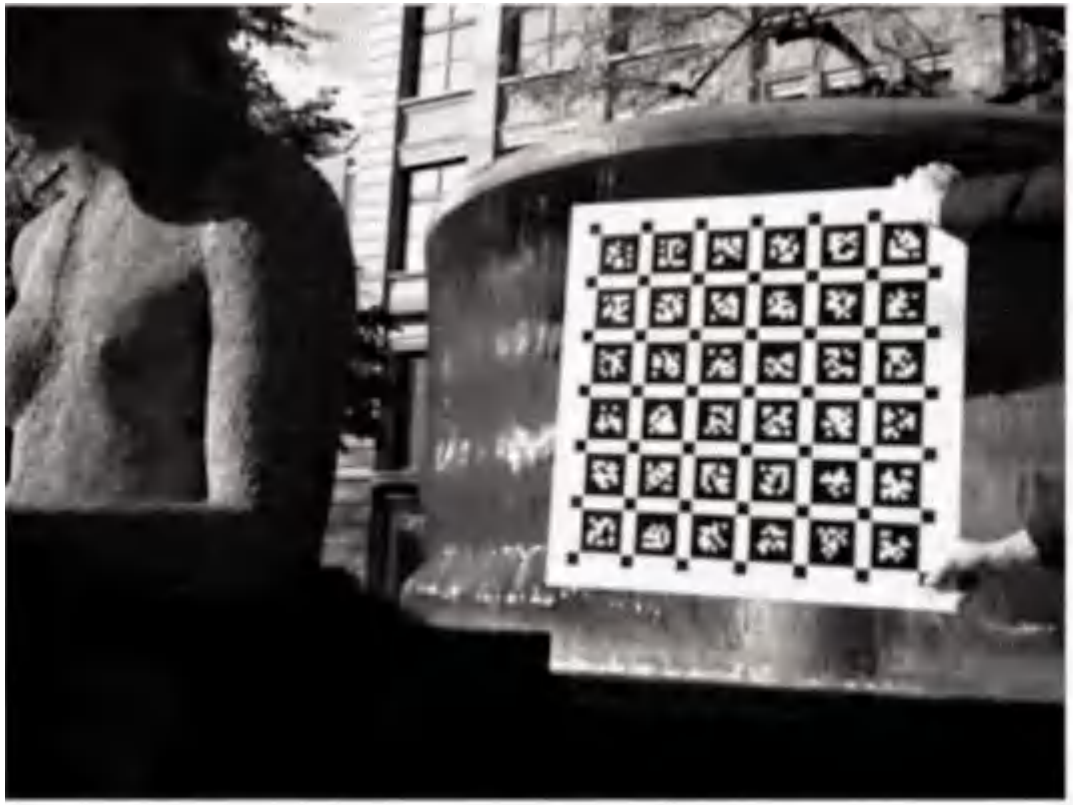


- Warping-based interpolation模块利用U形网络将运动转换为光流表示，然后将事件转换成真实的帧。
- Interpolation by synthesis模块也是利用U形网络将事件置于两个帧之间，并直接为每个事件生成一个新的可能帧。
- Warping refinement模块提取同一事件的两个生成帧中最有价值的信息，进行变形优化。

Timelens



Timelens



Time Lens (ours)

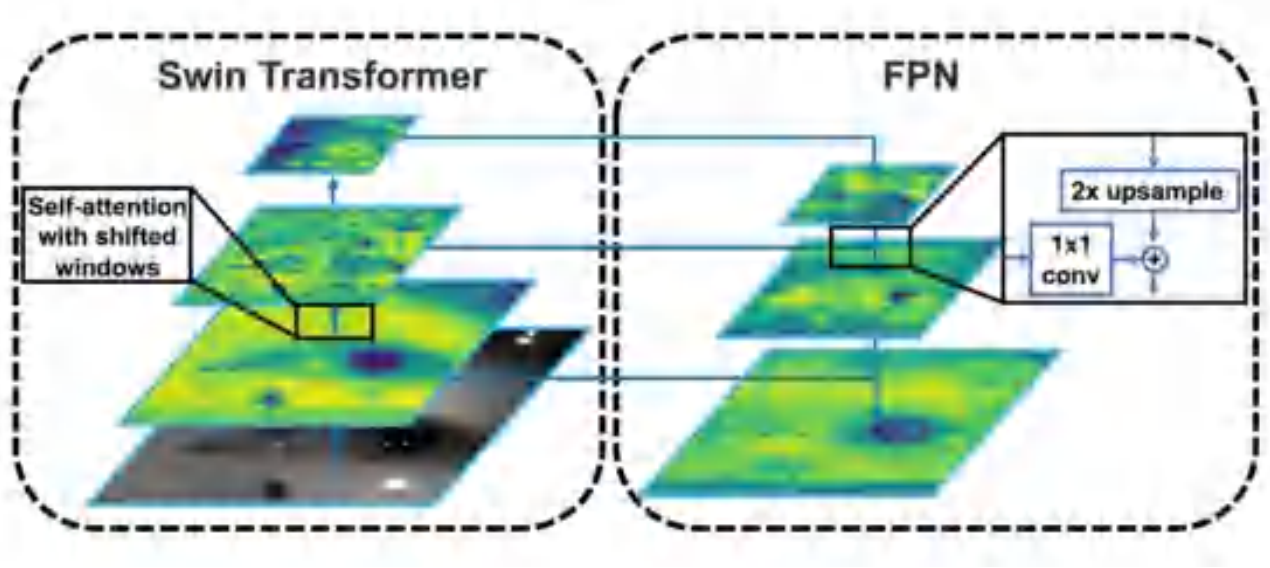
DAIN [Bao CVPR'19]

Multi-scale spatial feature extraction

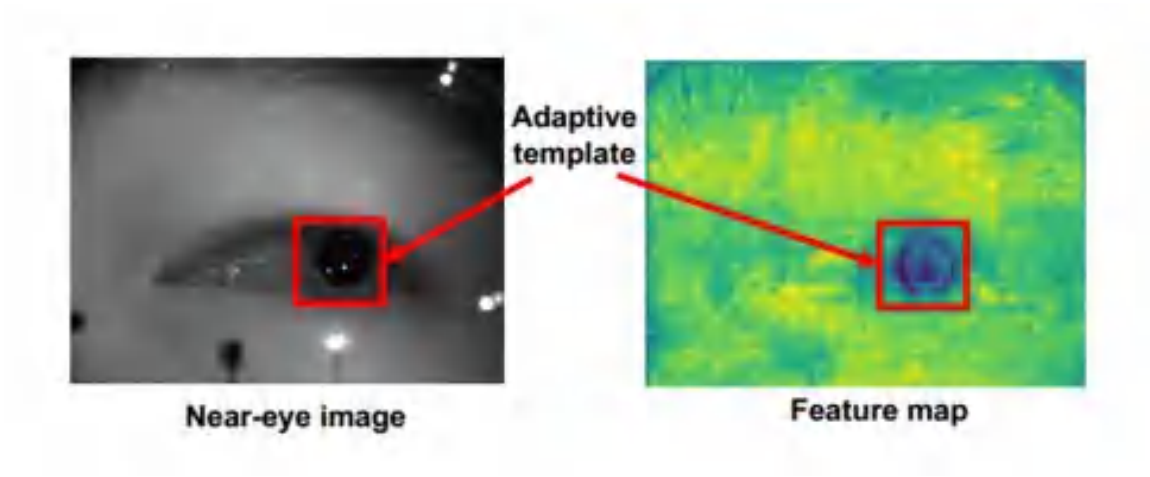
提取具有代表性和鲁棒性的特征来编码近眼图像的局部空间信息是瞳孔区域估计成功的关键因素之一。

利用移动窗口计算一个窗口内的自注意力，在降低计算复杂度的同时提取局部空间特征。

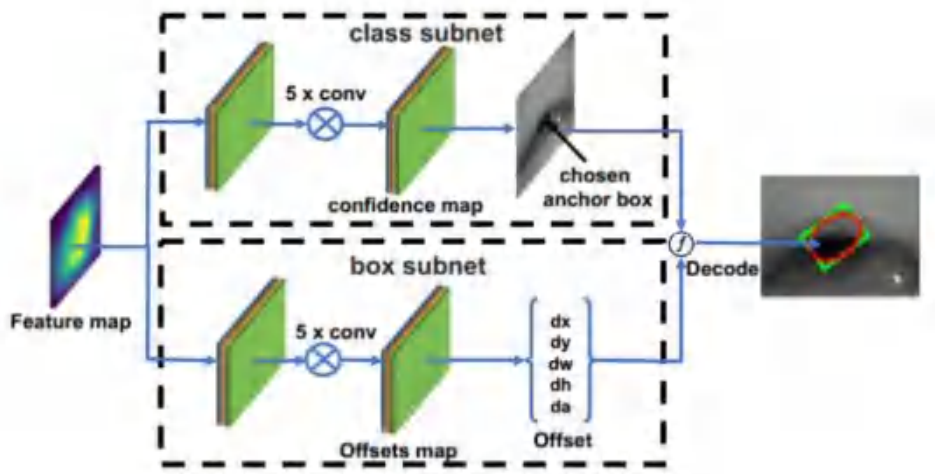
对于FPN，从上到下，上层特征图上采样两次，下层特征图进行1×1卷积。将来自相邻两层的处理后的特征图加在一起，形成融合的特征图。**该特征图既包含来自较低层的丰富细节，也包含来自较高层的高级语义信息。**



Pupil region estimation



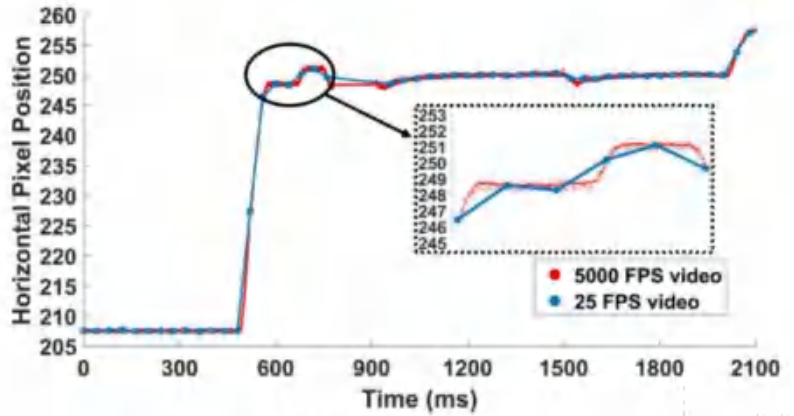
瞳孔区域定位成功后，选取一个以瞳孔区域中间为中心的小正方形区域。由于低层次的局部空间信息对于区分瞳孔和虹膜区域的边缘至关重要，因此文章将特征金字塔最低层次的对应区域作为下一步进行时间特征融合的模板。



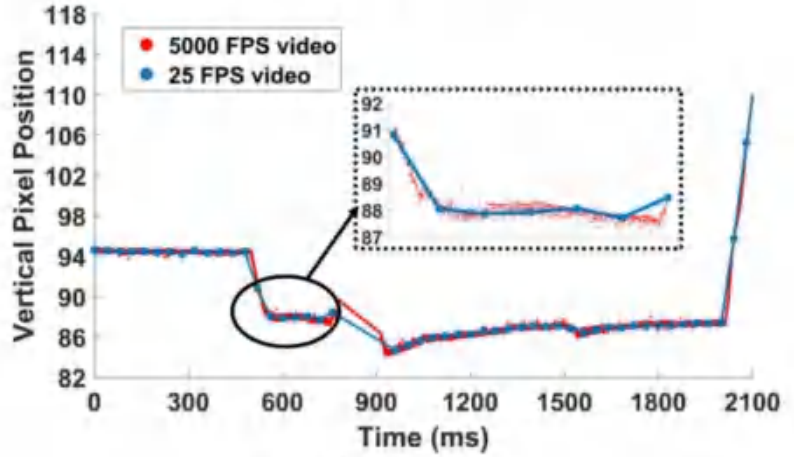
事先生成多个不同大小的锚盒，类子网确定锚框包含瞳孔区域的置信度，盒子网计算对应锚框到瞳孔区域边界框的偏移量，确定置信度最高的锚盒和相应的偏移量。

Pupil region estimation

“从右图的放大视图中，我们可以从密集的红点中清楚地看到高速视频中详细的眼球运动，这与简单地插值低帧率视频中获得的离散蓝点不同。这些差异通常发生在眼睛快速移动时，因此需要高频跟踪技术来揭示眼球运动背后难以察觉的知识。”



(a) Horizontal movement



(b) Vertical movement

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

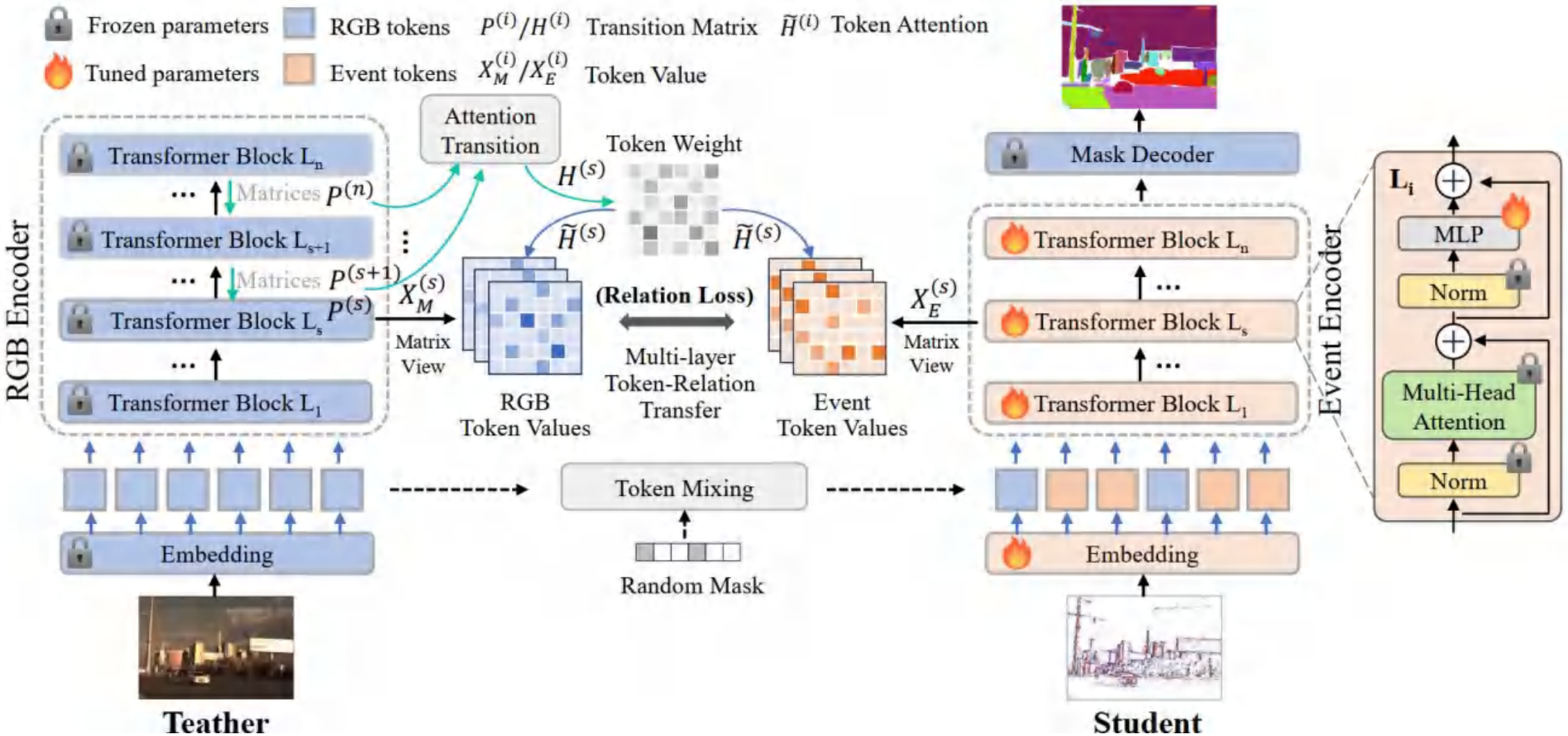
2

Segment Any Event Streams via Weighted Adaptation of Pivotal Tokens

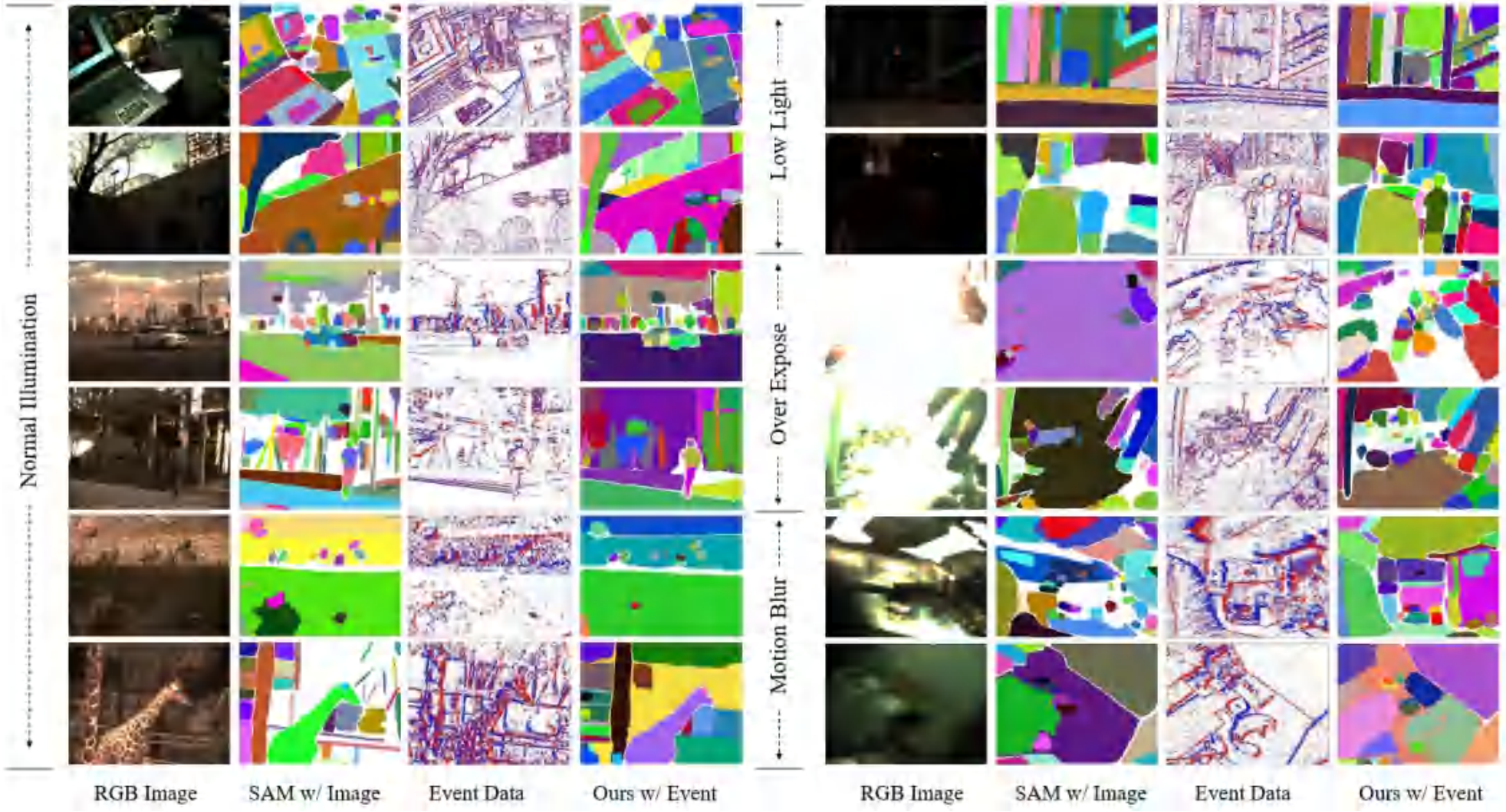
基于事件相机的典型应用（二）

这篇文章提出了深入探讨了将“Segment Anything Models”（**SAMs**）与事件数据结合的挑战，旨在实现**事件领域内**稳健和通用的**对象分割**。为此，其引入了一种多尺度特征蒸馏方法，该方法严格优化源自事件数据的标记嵌入与其RGB图像对应嵌入的对齐。针对不同数据集的广泛实验表明所提出的蒸馏方法的有效性。

2 框架


Frozen parameters
Tuned parameters
RGB tokens
Event tokens
$P^{(i)}/H^{(i)}$ Transition Matrix
$\tilde{H}^{(i)}$ Token Attention
$X_M^{(i)}/X_E^{(i)}$ Token Value
RGB Encoder
Transformer Block L_n
Matrices $P^{(n)}$
Transformer Block L_{s+1}
Matrices $P^{(s+1)}$
Transformer Block L_s $P^{(s)}$
Transformer Block L_1
Embedding
Teather
Attention Transition
Token Weight
$H^{(s)}$
$\tilde{H}^{(s)}$
$\tilde{H}^{(s)}$
$X_M^{(s)}$
Matrix View
RGB Token Values
(Relation Loss)
Multi-layer Token-Relation Transfer
Event Token Values
$X_E^{(s)}$
Matrix View
Token Mixing
Random Mask
Mask Decoder
Transformer Block L_n
Transformer Block L_s
Transformer Block L_1
Embedding
Student
Event Encoder
L_i
MLP
Norm
Multi-Head Attention
Norm

2 结果展示


Normal Illumination
Low Light
Over Expose
Motion Blur
RGB Image
SAM w/ Image
Event Data
Ours w/ Event
RGB Image
SAM w/ Image
Event Data
Ours w/ Event

2 结果展示

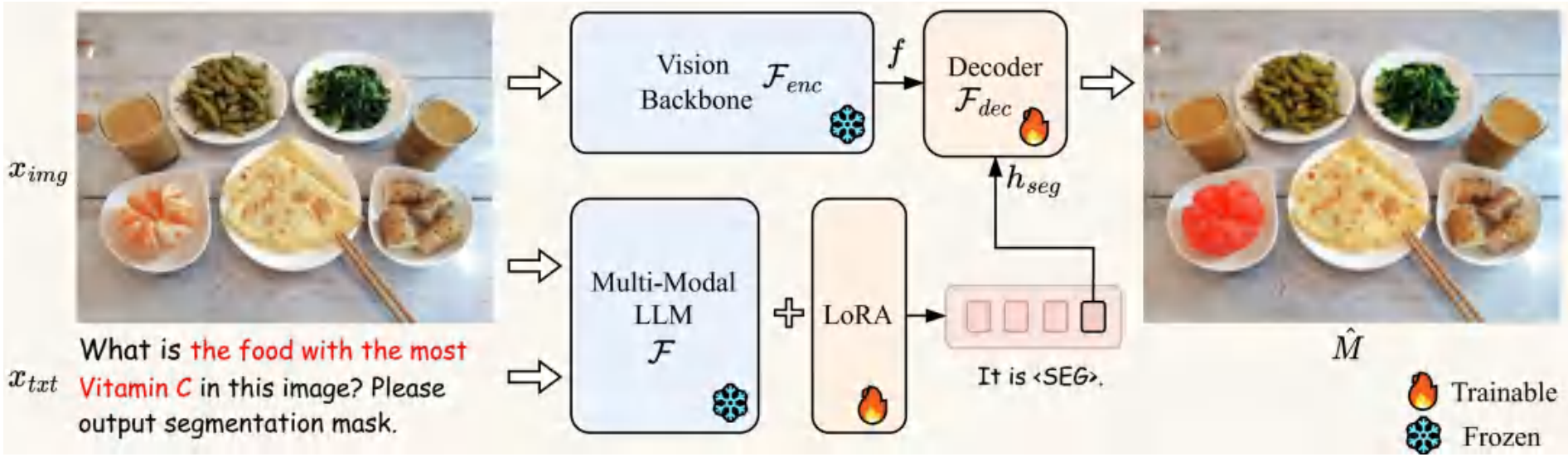

x_{img}
x_{txt}
What is the food with the most Vitamin C in this image? Please output segmentation mask.
Vision Backbone $\mathcal{F}_{enc}$
f
Decoder $\mathcal{F}_{dec}$
h_{seg}
Multi-Modal LLM $\mathcal{F}$
LoRA
It is <SEG>.
$\hat{M}$
Trainable
Frozen


Image
Event
Segmentation
What object is flying in the sky? Please output the segmentation mask.


Image
Event
Segmentation
What object is flying over the building? Please output the segmentation mask.

计算机科学与技术前沿汇报

Thesis report

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目 录

CONTENTS

第一部分 ▷▷

GaussianShader: 3D Gaussian Splatting with Shading Functions for Reflective Surfaces

- 研究背景
- 论文方法
- 实验与结论

- 高反射三维场景重建方法
  - 基于NeRF的方法：将着色函数与隐式表达相结合
    - Ref-NeRF：渲染速度缓慢，优化时间长（数小时）
    - ENVIDR：SDF灵活程度有限，难以重建复杂场景，且相比3DGS渲染缓慢
  - 3D Gaussian Slpatting：球谐函数表示物体的颜色
    - 四阶球谐函数表达能力有限，在面对高反射场景时性能显著降低

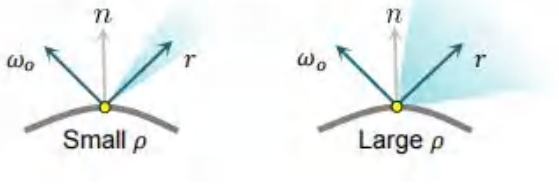
主要贡献

- 通过**简化的着色函数**显式地近似渲染方程，显着增强渲染场景的真实感，特别是对于高度镜面反射和反射的表面。
- 提出了一个新的**3D高斯法线估计框架**，具有新的正则化损失，可以实现精确的法线估计。
- 利用高斯泼溅的效率，提供**实时渲染**功能，使其适合需要高效渲染的交互式应用程序和场景。

- 简化着色函数（替换球谐函数）

$$\mathbf{c}(\omega_o) = \gamma(\mathbf{c}_d + \mathbf{s} \odot L_s(\omega_o, \mathbf{n}, \rho) + \mathbf{c}_r(\omega_o))$$

$$L_s(\omega_o, \mathbf{n}, \rho) = \int_\Omega L(\omega_i) D(\mathbf{r}, \rho)(\omega_i \cdot \mathbf{n}) d\omega_i$$



ω_0：相机视角
γ：伽马调色映射函数
c_d：椭球固有颜色
s：反射表面颜色
n：椭球法向量
ρ：表面粗糙度
c_r：残差颜色，用于表示全局光照，用球谐函数表示
ω_i：入射光方向
L：场景光照信息，由可微分光照贴图给出

- 高斯椭球法向量

$$\mathbf{n} = \begin{cases} \mathbf{v} + \Delta\mathbf{n}_1 & \text{if } \omega_o \cdot \mathbf{v} > 0, \\ -(\mathbf{v} + \Delta\mathbf{n}_2) & \text{otherwise.} \end{cases}$$

$\mathcal{L}_{tvg} = ||\Delta\mathbf{n}||^2$ ：防止迭代出的法向量与椭球极短轴方向偏差过大

$\mathcal{L}_{normal} = ||\bar{\mathbf{n}} - \mathbf{n}||^2$ ：保证局部形状之间法向量的相关性

- 损失函数

$$\mathcal{L} = \mathcal{L}_{\text{color}} + \lambda_{\text{n}}\mathcal{L}_{\text{normal}} + \lambda_{\text{s}}\mathcal{L}_{\text{sparse}} + \lambda_{\text{r}}\mathcal{L}_{\text{reg}}$$

$$\mathcal{L}_{\text{color}} = \|\mathbf{C} - \mathbf{C}_{\text{gt}}\|^2$$

$$\mathcal{L}_{\text{normal}} = \|\bar{\mathbf{n}} - \tilde{\mathbf{n}}\|^2$$

$$\mathcal{L}_{\text{sparse}} = \frac{1}{|\alpha|}\sum_{\alpha_i}\left[\log(\alpha_i) + \log(1-\alpha_i)\right].$$

稀疏损失：鼓励椭球的不透明度向两端接近

$$\mathcal{L}_{\text{reg}} = \|\Delta\mathbf{n}\|^2$$

$$\lambda_{\text{n}} = 0.01, \lambda_{\text{s}} = 0.001, \lambda_{\text{r}} = 0.001.$$

	NeRF Synthetic [32]								
	Chair	Drums	Lego	Mic	Materials	Ship	Hotdog	Ficus	Avg.
	PSNR↑								
NeRF [32]	33.00	25.01	32.54	32.91	29.62	28.65	36.18	30.13	31.01
VolSDF [52]	30.57	20.43	29.46	30.53	29.13	25.51	35.11	22.91	27.96
Ref-NeRF [45]	33.98	25.43	35.10	33.65	27.10	29.24	37.04	28.74	31.29
ENVIDR [27]	31.22	22.99	29.55	32.17	29.52	21.57	31.44	26.60	28.13
Gaussian Splatting [21]	35.82	26.17	35.69	35.34	30.00	30.87	37.67	34.83	33.30
Ours	35.83	26.36	35.87	35.23	30.07	30.82	37.85	34.97	33.38
	SSIM↑								
NeRF [32]	0.967	0.925	0.961	0.980	0.949	0.856	0.974	0.964	0.947
VolSDF [52]	0.949	0.893	0.951	0.969	0.954	0.842	0.972	0.929	0.932
Ref-NeRF [45]	0.974	0.929	0.975	0.983	0.921	0.864	0.979	0.954	0.947
ENVIDR [27]	0.976	0.930	0.961	0.984	0.968	0.855	0.963	0.987	0.956
Gaussian Splatting [21]	0.987	0.954	0.983	0.991	0.960	0.907	0.985	0.987	0.969
Ours	0.987	0.949	0.983	0.991	0.960	0.905	0.985	0.985	0.968
	LPIPS↓								
NeRF [32]	0.046	0.091	0.050	0.028	0.063	0.206	0.121	0.044	0.081
VolSDF [52]	0.056	0.119	0.054	0.191	0.048	0.191	0.043	0.068	0.096
Ref-NeRF [45]	0.029	0.073	0.025	0.018	0.078	0.158	0.028	0.056	0.058
ENVIDR [27]	0.031	0.080	0.054	0.021	0.045	0.228	0.072	0.010	0.067
Gaussian Splatting [21]	0.012	0.037	0.016	0.006	0.034	0.106	0.020	0.012	0.030
Ours	0.012	0.040	0.014	0.006	0.033	0.098	0.019	0.013	0.029

RTX 3090

高反射数据集

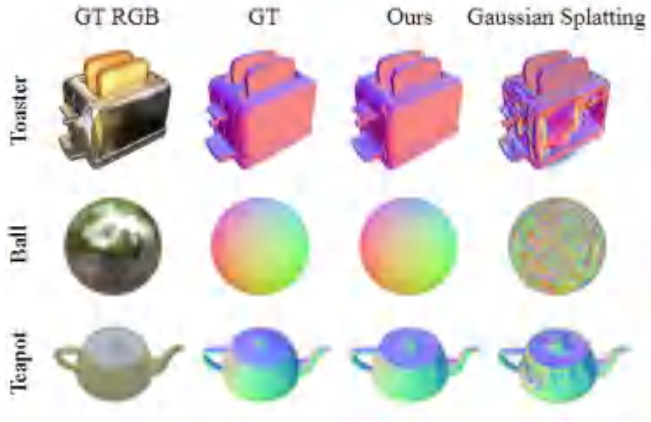
	Shiny Blender [45]						
	Car	Ball	Helmet	Teapot	Toaster	Coffee	Avg.
	PSNR↑						
NVDiffRec [34]	27.98	21.77	26.97	40.44	24.31	30.74	28.70
NVDiffMC [18]	25.93	30.85	26.27	38.44	22.18	29.60	28.88
Ref-NeRF [45]	30.41	29.14	29.92	45.19	25.29	33.99	32.32
NeRO [29]	25.53	30.26	29.20	38.70	26.46	28.89	29.84
ENVIDR [27]	28.46	38.39	32.23	41.59	26.11	29.48	32.88
Gaussian Splatting [21]	27.24	27.69	28.32	45.68	20.99	32.32	30.37
Ours	27.90	30.98	28.32	45.86	26.21	32.39	31.94
	SSIM↑						
NVDiffRec [34]	0.963	0.858	0.951	0.996	0.928	0.973	0.945
NVDiffMC [18]	0.940	0.940	0.940	0.995	0.886	0.965	0.944
Ref-NeRF [45]	0.949	0.956	0.955	0.995	0.910	0.972	0.956
NeRO [29]	0.949	0.974	0.971	0.995	0.929	0.956	0.962
ENVIDR [27]	0.961	0.991	0.980	0.996	0.949	0.949	0.969
Gaussian Splatting [21]	0.930	0.937	0.951	0.996	0.895	0.971	0.947
Ours	0.931	0.965	0.950	0.996	0.929	0.971	0.957
	LPIPS↓						
NVDiffRec [34]	0.045	0.297	0.118	0.014	0.169	0.076	0.119
NVDiffMC [18]	0.077	0.312	0.157	0.014	0.225	0.097	0.147
Ref-NeRF [45]	0.051	0.307	0.087	0.013	0.118	0.082	0.109
NeRO [29]	0.074	0.094	0.050	0.012	0.089	0.110	0.072
ENVIDR [27]	0.049	0.087	0.051	0.011	0.146	0.139	0.072
Gaussian Splatting [21]	0.047	0.161	0.079	0.007	0.126	0.078	0.083
Ours	0.045	0.121	0.056	0.007	0.079	0.078	0.068

渲染速度（包含NeRF Synthetic和Shiny Blender）

	PSNR	Training Time	FPS
Ref-NeRF [45]	31.73	23h	0.03
ENVIDR [27]	30.04	6h	1.33
Gaussian Splatting [21]	32.05	**0.25h**	**274**
Ours	**32.76**	0.58h	97

消融实验

	PSNR↑	SSIM↑	LPIPS↓
w/o $\mathcal{L}_{sparse}$	31.79	0.952	0.056
w/o $\mathcal{L}_{normal}$	30.93	0.941	0.060
w/o $\mathbf{c}_r$	31.49	0.948	0.060
w/o $\mathbf{v}$	31.47	0.951	0.058
MLP Lighting	29.73	0.936	0.075
Full Model	**32.09**	**0.953**	**0.054**



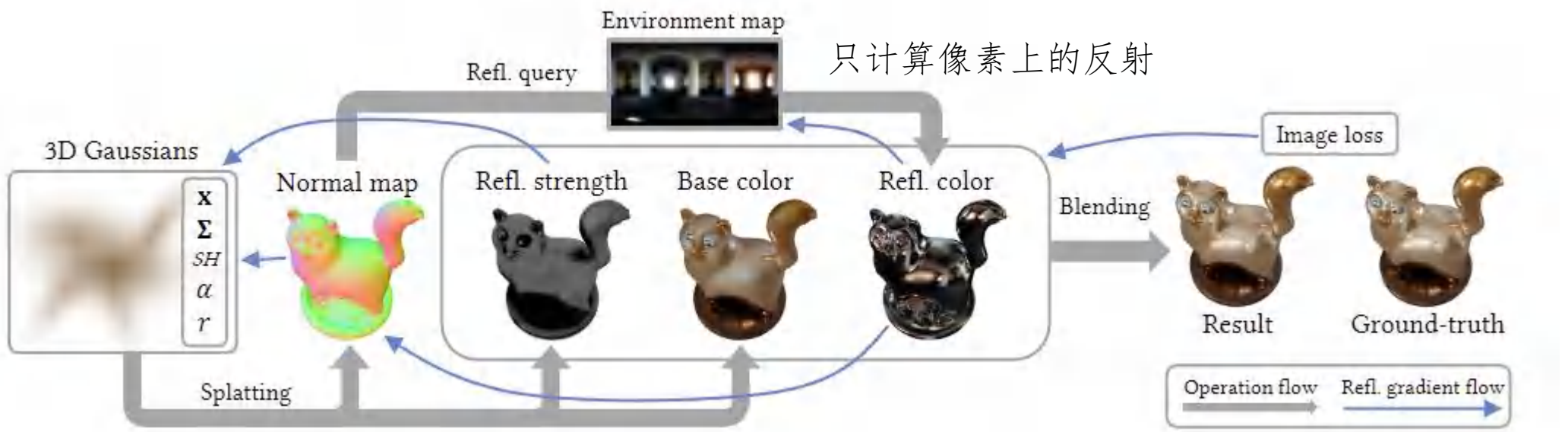
第二部分 ▷▷

3D Gaussian Splatting with Deferred Reflection

- 研究背景
- 论文方法
- 实验与结论

主要贡献

- 介绍了一种**延迟着色**方法，可以通过高斯泼溅有效渲染镜面反射。
- 出了一种具有**优化传播**的训练算法，使得正确的法线信息能够在高斯之间传播，提高高斯法线精度。
- 保持与常规高斯泼溅相近的**实时渲染**速度，并且能够生成更准确的法线和环境贴图数据



3DGS模块，将反射强度r、法线n和基色烘焙到屏幕空间贴图。对于每个像素，使用法线贴图来计算反射方向并查询环境贴图以获取反射颜色。然后使用反射强度将基色和反射颜色混合成最终结果

$$C(\mathbf{v}) = \sum_i c_i(\mathbf{v}) G(\Theta_i, \mathbf{v}).$$

$$N(\mathbf{v}) = \sum_i n_i G(\Theta_i, \mathbf{v}),\ R(\mathbf{v}) = \sum_i r_i G(\Theta_i, \mathbf{v}),$$

$$C'(\mathbf{v}) = (1 - R(\mathbf{v}))C(\mathbf{v}) + R(\mathbf{v})E\left(\frac{2\mathbf{v} \cdot N(\mathbf{v})N(\mathbf{v})}{\|N(\mathbf{v})\|} - \mathbf{v}\right)$$

V是相机视角，G是GS泼溅至屏幕空间的运算，E是在环境贴图上的一个双线性查询

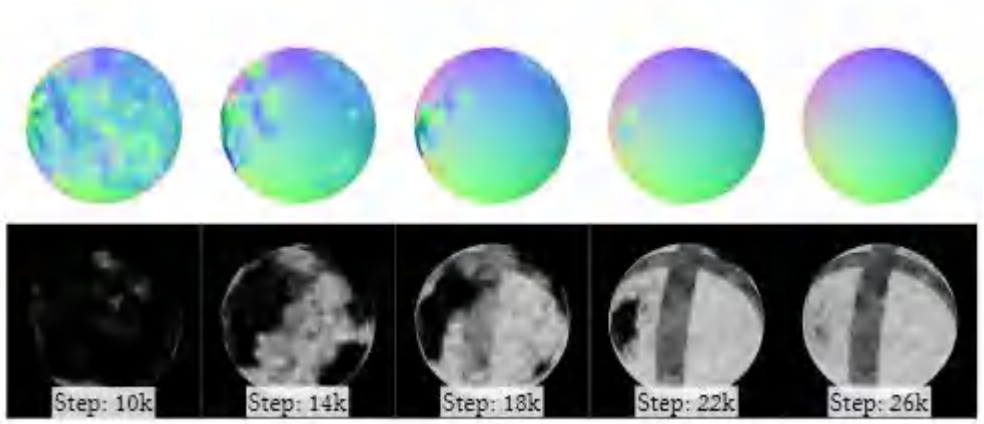
训练细节

- 将SH阶数设置为0，反射强度r_i设置为0，即关闭颜色和反射强度优化进行前导训练（数千次）

- 开启反射强度优化，定期将所有高斯不透明度设为0.9，反射强度提高到至少0.001，将两最长轴扩展1.5倍，对法线和环境贴图进行优化，称为**正常传播**。

- 为了抵消漫反射颜色$c_{i,0}$过拟合，在应用法向传播时添加±10% 的噪声，故意破坏 $r_i \leqslant 0.1$ 的尚未反射的高斯的颜色，称为**颜色破坏**。

- 将反射率提升周期与反射率抑制周期错开，当$r_i > 0.1$的高斯数量在一定迭代周期内停止增加，则停止正常传播与颜色破坏，得到**漫反射环境贴图**。再将SH阶数提高，开始正常训练。



实验与结论

- 损失函数

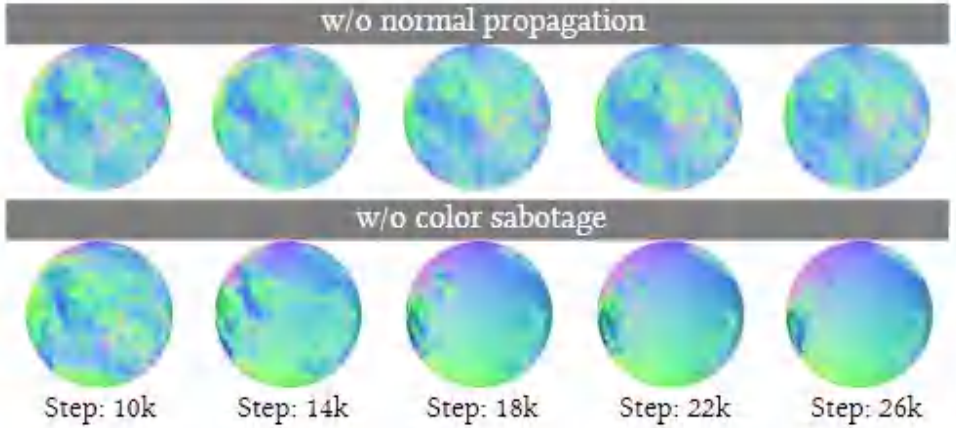
$$\mathcal{L} = (1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{D-SSIM}$$

- Ours, forward

直接将反射强度等信息跟着原始高斯一起训练

Datasets		Shiny Blender [Verbin et al. 2022]						Glossy Synthetic [Liu et al. 2023]						Real		
		ball	car	coffee	helmet	teapot	toaster	bell	cat	luyu	potion	tbell	teapot	garden	sedan	toycar
PSNR ↑	Ref-NeRF	33.16	30.44	33.99	29.94	45.12	26.12	30.02	29.76	25.42	30.11	26.91	22.77	22.01	25.21	23.65
	NPC	23.76	24.19	30.39	25.59	41.22	19.76	22.41	25.35	23.68	23.09	19.03	18.21	21.01	24.77	22.84
	3DGS	27.65	27.26	32.30	28.22	45.71	20.99	25.11	31.36	26.97	30.16	23.88	21.51	21.75	26.03	23.78
	GShader	30.99	27.96	32.39	28.32	45.86	26.28	28.07	31.81	27.18	30.09	24.48	23.58	21.74	24.89	23.76
	ENVIDR	41.02	27.81	30.57	32.71	42.62	26.03	30.88	31.04	28.03	32.11	28.64	26.77	21.47	24.61	22.92
	Ours, forward	27.64	28.99	31.61	28.01	45.68	24.83	25.74	32.22	27.01	30.25	24.11	23.13	21.49	26.05	23.49
	Ours, deferred	33.66	30.39	34.65	31.69	47.12	27.02	31.65	33.86	28.71	32.29	28.94	25.36	21.82	26.32	23.83
SSIM ↑	Ref-NeRF	0.971	0.950	0.972	0.954	0.995	0.921	0.941	0.944	0.901	0.933	0.947	0.897	0.584	0.720	0.633
	NPC	0.908	0.898	0.955	0.938	0.994	0.835	0.892	0.921	0.854	0.877	0.742	0.762	0.558	0.711	0.547
	3DGS	0.937	0.931	0.972	0.951	0.996	0.894	0.908	0.959	0.916	0.938	0.900	0.881	0.571	0.771	0.637
	GShader	0.966	0.932	0.971	0.951	0.996	0.929	0.919	0.961	0.914	0.936	0.898	0.901	0.576	0.728	0.637
	ENVIDR	0.997	0.943	0.962	0.987	0.995	0.922	0.954	0.965	0.931	0.960	0.947	0.957	0.561	0.707	0.549
	Ours, forward	0.939	0.941	0.968	0.947	0.996	0.919	0.909	0.964	0.911	0.938	0.904	0.891	0.566	0.767	0.626
	Ours, deferred	0.979	0.962	0.976	0.971	0.997	0.943	0.962	0.976	0.936	0.957	0.952	0.936	0.581	0.773	0.639
LPIPS ↓	Ref-NeRF	0.166	0.050	0.082	0.086	0.012	0.083	0.102	0.104	0.098	0.084	0.114	0.098	0.251	0.234	0.231
	NPC	0.237	0.120	0.119	0.156	0.013	0.226	0.203	0.121	0.101	0.174	0.243	0.246	0.302	0.311	0.347
	3DGS	0.162	0.047	0.079	0.081	0.008	0.125	0.104	0.062	0.064	0.093	0.125	0.102	0.248	0.206	0.237
	GShader	0.121	0.044	0.078	0.074	0.007	0.079	0.098	0.056	0.064	0.088	0.122	0.091	0.274	0.259	0.239
	ENVIDR	0.020	0.046	0.083	0.036	0.009	0.081	0.054	0.049	0.059	0.072	0.069	0.041	0.263	0.387	0.345
	Ours, forward	0.156	0.044	0.081	0.082	0.008	0.091	0.104	0.059	0.068	0.096	0.124	0.096	0.252	0.221	0.249
	Ours, deferred	0.098	0.033	0.076	0.049	0.005	0.081	0.046	0.040	0.053	0.075	0.067	0.067	0.247	0.208	0.231

i7-13700KF CPU, NVIDIA RTX 4090 GPU



- 训练资产

	ball	car	helmet	toaster
Ours, deferred	27.2k	117k	36k	100k
Ours, forward	76.4k	205k	62k	236k
GShader	76.5k	316k	106k	336k
3DGS	199k	307k	100k	342k

- 消融实验

Ablations	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	33.66	0.979	0.098
w/o propagation	27.85	0.938	0.159
w/o sabotage	30.00	0.959	0.128

- 非高反射平面

	NeRF Synthetic					
	drums	ficus	hotdog	lego	mic	ship
	PSNR ↑					
3DGS	25.10	28.14	35.52	32.94	31.55	29.06
Ours	25.31	28.03	35.58	32.94	31.97	29.07
	SSIM ↑					
3DGS	0.947	0.965	0.983	0.979	0.986	0.897
Ours	0.946	0.963	0.982	0.978	0.987	0.894
	LPIPS ↓					
3DGS	0.055	0.540	0.032	0.025	0.028	0.124
Ours	0.055	0.055	0.033	0.026	0.028	0.129

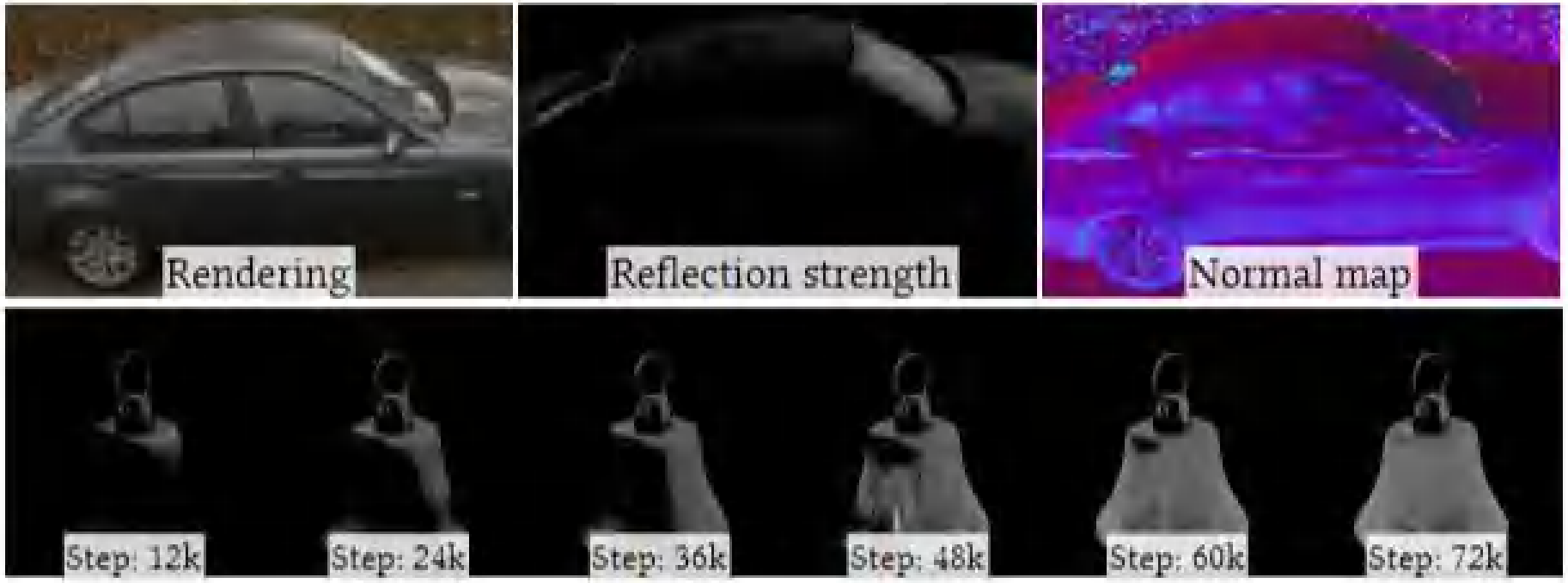


Figure 7: Limitations. Top row: inconsistent treatment of car windows. Bottom row: slow convergence on concave bell.

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

计算机科学与技术前沿

基于路径生成的全景图像质量评价

汇报人：刘勇奇

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

- 全景图像
- 路径生成
- Assensor360
- 实验结果
- 前景展望

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

- 全景图像
- 路径生成
- Assensor360
- 实验结果
- 前景展望

1 全景图像

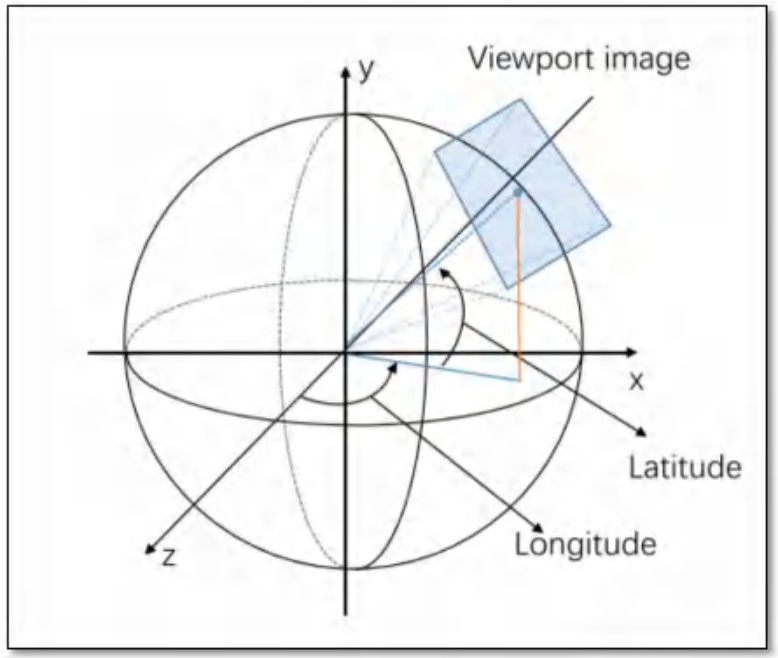
全景图像也就是360°全景图像，即通过对专业相机捕捉整个场景的图像信息或者使用建模软件渲染过后的图片，使用软件进行图片拼合，并用头戴式显示器(HMD)进行播放。

全景图像提供了全方位视角，互动性强，在虚拟现实(VR)，三维动画，三维导航等方面有广泛应用。

1 全景图像

由于全景图像在应用中是360°球形图像，为了方便通过客观方法对图像进行分析，首先通过投影方法将全景图像投影至平面。

视口(viewport)：用户在通过头戴式设备观察全景图像时，观察到的一种视图形式，由于视野的局限性，往往只能看到全景图像的局部。



1 全景图像

目前， 广泛使用的投影模式有ERP投影与CMP投影，其他投影方式包括条带投影(SSP)等。由于信息量的缘故，全景图像的像素量往往是普通图像的10倍以上。这对全景图像的压缩，传输以及质量控制提出了更高的要求。

ERP图像

CMP图像

1 全景图像

出于全景图像的视觉体验的考虑，全景图像对于质量的要求相比传统图像更为严格，这也为全景质量评估(OIQA)提出了更多的挑战。

目前全景质量评估的困难主要有以下几点：

- 全景图像通常以等矩形投影（ERP）格式存储，这种格式在不同纬度上显示出相当大的几何变形。这种扭曲会对质量评估产生负面影响。而CMP图像由于对图像进行了分割，会损失部分的全局信息。
- 用户观察体验会受到观察路径的影响，全景图像的观察往往集中于某些视口，这些视口根据用户的观察路径呈现出很强的相互依赖性。
- 目前客观的质量评估具有低成本，高效率等优点，但高效模型的建立依赖于主观质量评估建立的数据库，全景图像由于拍摄和放映成本的问题数据库的构建成本高，数据量少。
- 全景图像相对传统图像信息量大，像素量多，质量评估计算速度较慢。

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

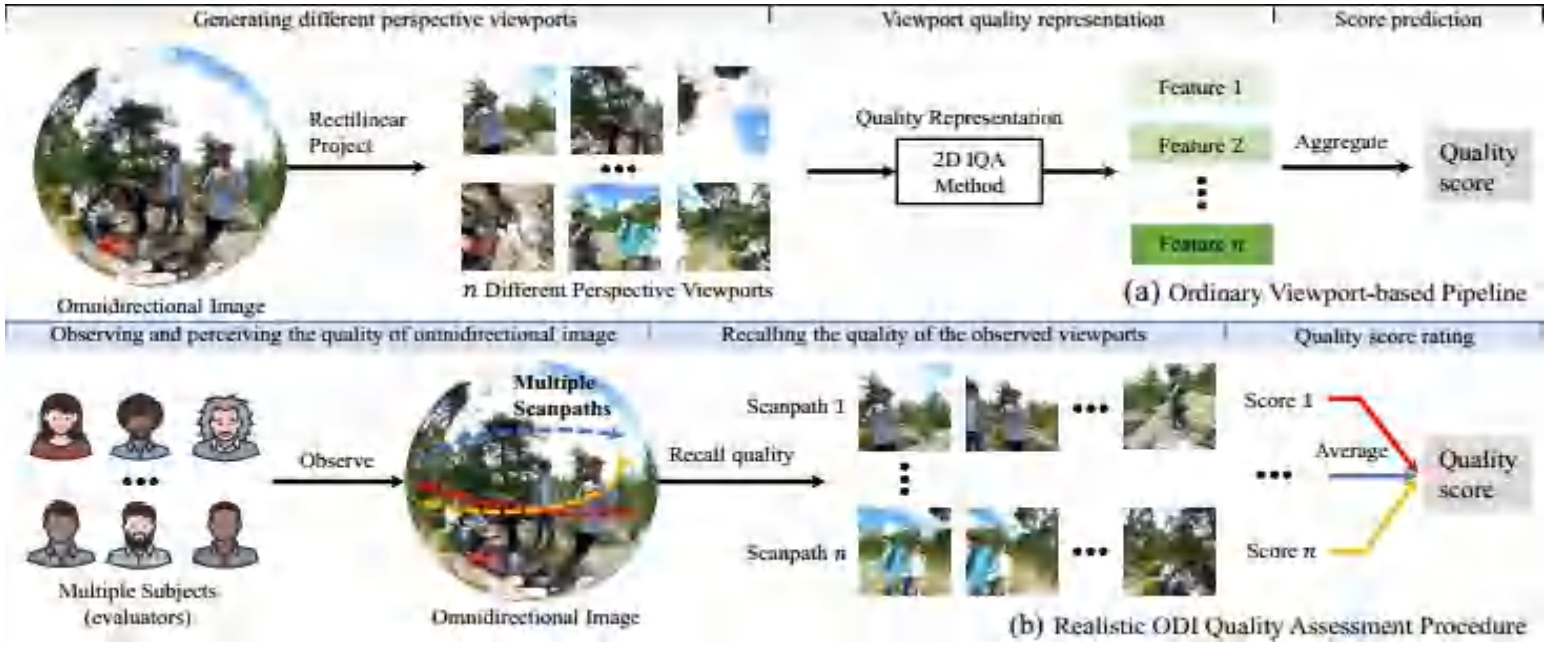
- 全景图像
- **路径生成**
- Assensor360
- 实验结果
- 前景展望

2 路径生成

全景图像本身具备曲面结构，用户的观察往往由一系列视口组成，实际观察的视口序列因人而异。

扫描路径是指在离散的时间序列上对直线视口序列进行采集(一般使用ERP重投影)，最终得到的是平面的图像序列，就可以使用传统的2D图像质量评估方法来处理。

然而基于固定视口序列的与实际的图像评价过程大有不同，如下图所示，图像质量评估往往会收集多个用户的评价，他们的视口序列不是固定的，这使得固定视口序列的扫描路径与实际情况产生较大的偏差。



2 路径生成

为了解决这个问题，一种合乎逻辑的方法是使用扫描路径预测模型在全景图像上生成伪视口序列，而不使用真实的扫描路径，即路径生成。

然而，目前的扫描路径生成方法是针对未失真的全景图像开发的，主要集中在高层区域。但是在主观全景图像质量评估任务中，由于评估人员需要为全景图像提供准确的质量分数，因此他们的扫描路径分布在低质量和高细节区域。这使得目前可用的路径生成方法都不适合用来完成客观质量评估的任务的路径生成。

针对以上的这些问题，此处介绍Assensor360这种基于概率生成扫描路径的方法，该方法在多个全景图像评估数据集上都取得了最优的结果。

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

- 全景图像
- 路径生成
- **Assensor360**
- 实验结果
- 前景展望

3 Assensor360

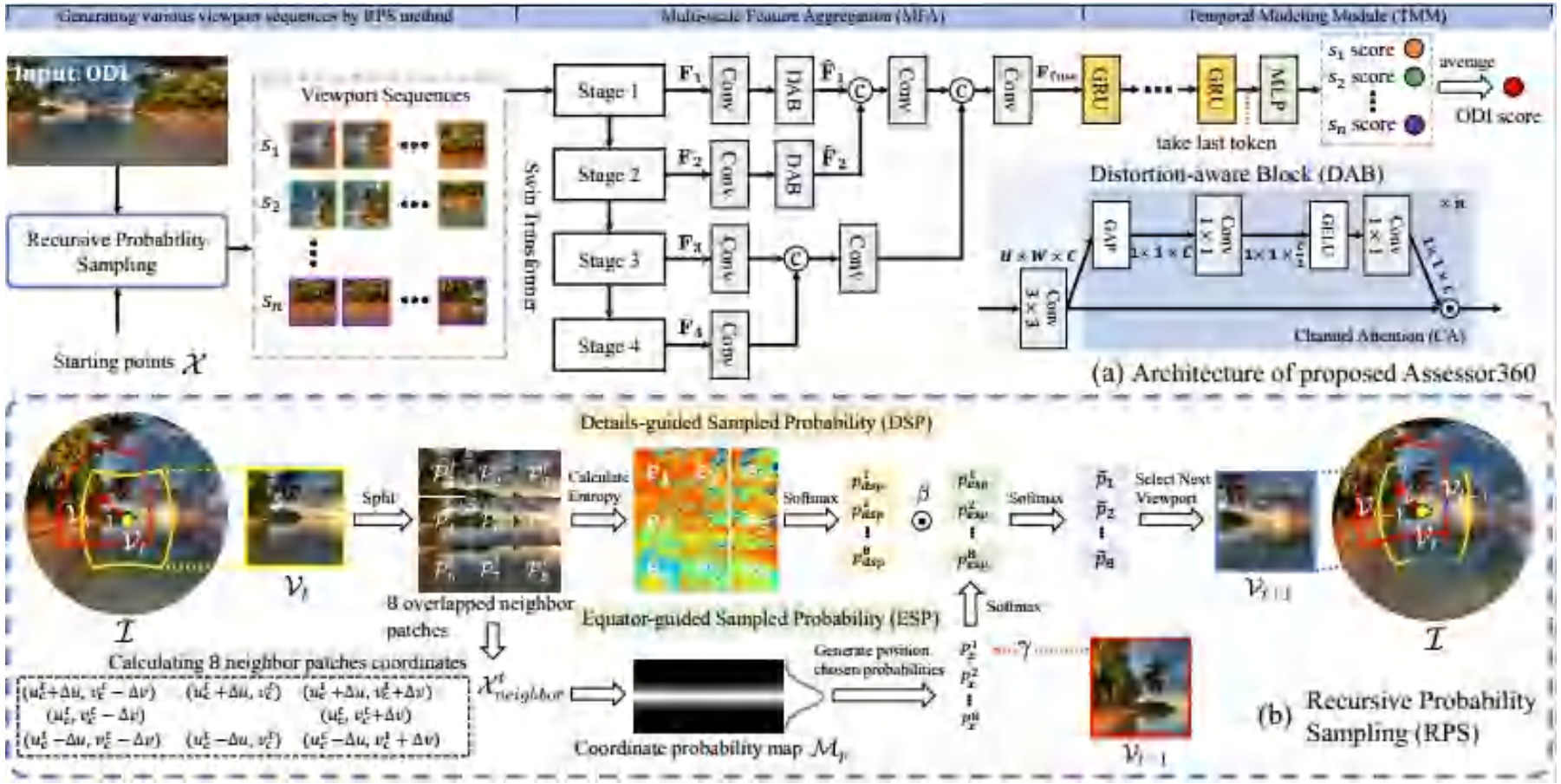
路径扫描算法关注于模仿人类视觉扫描路径的模式，但依赖人工采集扫描路径繁琐且耗时，这里介绍的多序列图像质量评价网络(Assensor360)由Wu Tianhe于提出，在多个全景图像质量评估数据集获得了最优性能。

该文章提出了一种广义递归概率抽样（RPS）方法，该方法结合语义场景和局部的失真特征，从给定的起点生成多个伪视口序列。

此外，该文章设计了一个带有失真感知块（DAB）的多尺度特征聚合（MFA）模块来融合每个视口的扭曲特征和语义特征。

本文同时使用了时域建模模块（TMM）来学习时域内的视口转换，对视口的转换的时间相关性进行建模。

该模型的整体结构如下图所示：



3 广义递归概率抽样

递归概率采样（RPS）策略可以基于语义上下文的先验知识和全景图像中的退化特征自适应生成场景过渡方向的概率，视口序列就是根据这个概率对不断过渡的视口进行采样来生成的。

对于每次生成的视口，计算向下个视口的过渡方向和距离，直到达到所需的采样视口数。而选择下个视口的过程主要由赤道引导采样概率（ESP）和细节引导采样概率（DSP）两部分组成。

在实际操作过程中，我们首先遵循摩尔邻域理论，从视口中心坐标定义K个邻域（K = 8）过渡方向。之后，将过渡距离(Δu, Δv)设置为（24°，24°），以避免采样过度重叠的视口。

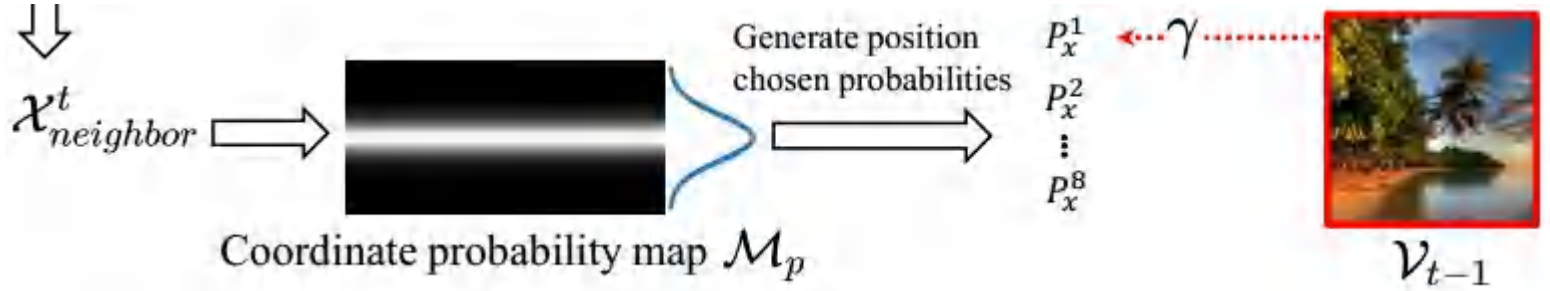
在浏览过程中，评估者不仅被吸引到高级场景中，还会关注低级纹理和细节区域，从而给出合理的质量评分。因此，根据广义先验内容信息和像素级细节度量，计算ESP与DSP，根据它们选择下一个采样的视口位置。

3 广义递归概率抽样

赤道引导采样概率(ESP)基于一个前提，即全景图像中的信息主要集中于赤道部分，这也是对人类观察方式的一种模拟。

假设样本相对赤道的采样偏差服从高斯分布。靠近赤道的区域具有较高的采样权值，而靠近赤道的区域具有较高的采样权值，使用Softmax进行映射得到不同坐标处的采样概率。

此外，在进行概率计算时，也需要考虑抑制返回(IOR)，即已经被眼睛注视过的区域再次被注视的概率降低，通过关注度递减因子γ对概率加权来得到ESP。

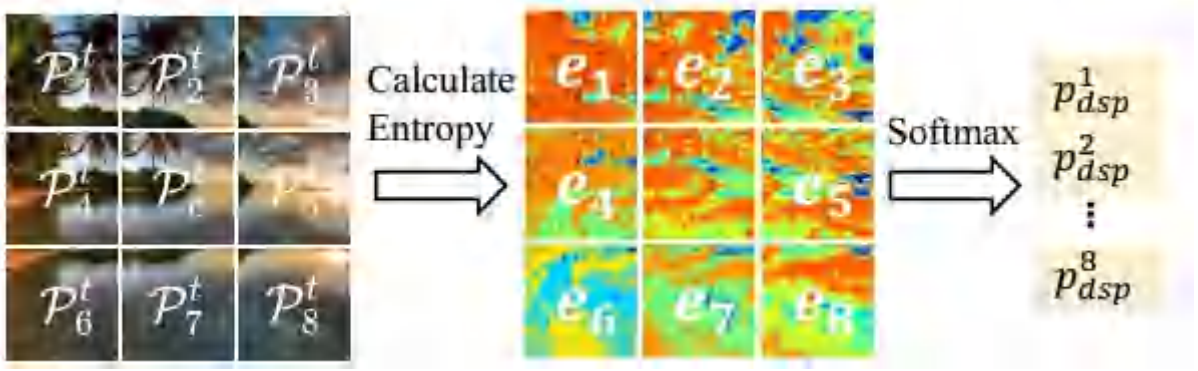


3 广义递归概率抽样

细节引导采样概率(DSP)则基于人类对于复杂纹理位置的关注的前提，对视口内的纹理细节复杂度进行量化，这里使用了像素级别的信息熵ε，计算方法如下：

$$\varepsilon(p_i^t) = \sum_{m=1}^{H}\sum_{n=1}^{W} -p(p_i^t[m,n])log_2(p(p_i^t[m,n]))$$

此处计算的$\varepsilon(p_i^t)$即为第t个视口的第i个邻域内的像素级信息熵，对于8个视口得到的信息熵进行Softmax归一化，得到DSP。



3 广义递归概率抽样

总体的RPS算法流程如下：

Algorithm 1 Viewport Sequence Generation (RPS Algorithm)

Input: N starting points $\{x_i\}_{i=1}^{N}$; an ODI $\mathcal{I}$; rectilinear projection $\mathcal{R}$; ESP calculation function $\mathcal{F}_{esp}(\mathcal{X})$; DSP calculation function $\mathcal{F}_{dsp}(\mathcal{P})$; selecting function $\Gamma(\mathcal{X}|\tilde{p})$

Output: A set of N length M viewport sequences $\{s_i = \{\mathcal{V}_t\}_{t=1}^{M}\}_{i=1}^{N}$;

for $i = 1 \rightarrow N$ **do**
  Initialize the current coordinate $x \leftarrow x_i$
  for $t = 1 \rightarrow M$ **do**
    Generate viewport by the current coordinate $\mathcal{V} \leftarrow \mathcal{R}(x, \mathcal{I})$
    Split $\mathcal{V}$ to obtain overlapped neighbor patches $\mathcal{P}$ and calculate sampled coordinate $\mathcal{X}$
    Calculate ESP and DSP $p_{esp} \leftarrow \mathcal{F}_{esp}(\mathcal{X}), p_{dsp} \leftarrow \mathcal{F}_{dsp}(\mathcal{P})$
    Generate next viewport coordinate $x \leftarrow \Gamma(\mathcal{X}|\text{Aggregate}(p_{esp}, p_{dsp}))$
  Sequentially gather generated M viewports $\{\mathcal{V}_t\}_{t=1}^{M}$ as a viewport sequence s_i
Gather generated N viewport sequences $\{s_i\}_{i=1}^{N}$ as the output

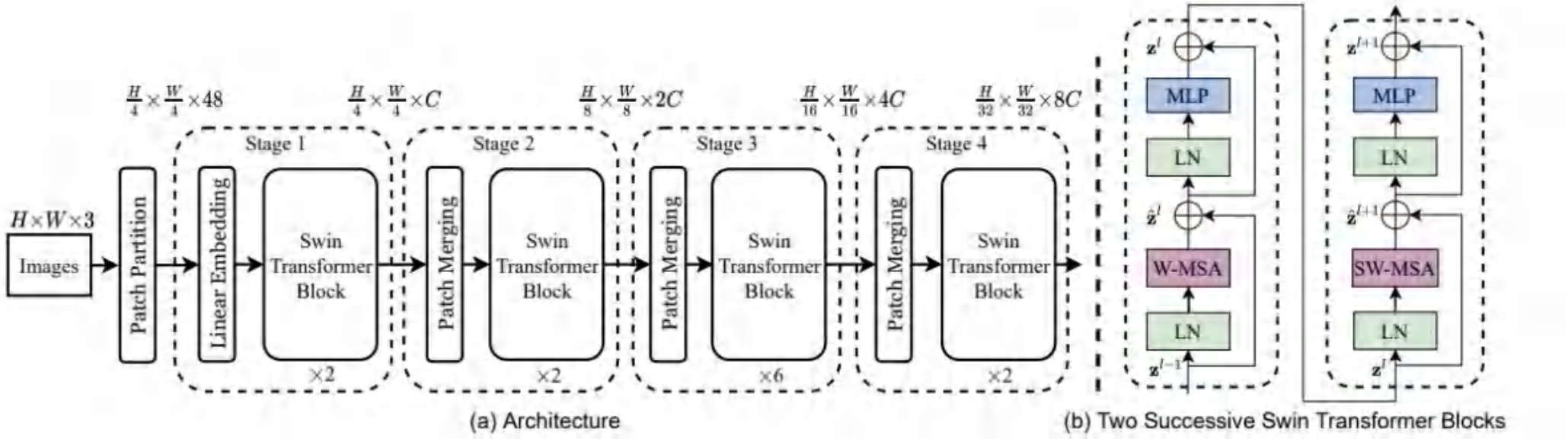
3 多尺度特征聚合

为了表示每个视口的语义信息和扭曲模式，文中对视口的多尺度特征进行了聚合。这里文章采用了预训练的Swin Transformer来得到四个阶段的不同尺度的特征。

采用的Swin Transformer的主干结构如下图：

一个 Swin Transformer block由一个基于移位窗口的 MSA 模块，后接一个带有GeLU非线性层的2层MLP。每个MSA和每个 MLP 前都有LayerNorm (LN)层，采用残差连接避免梯度退化。

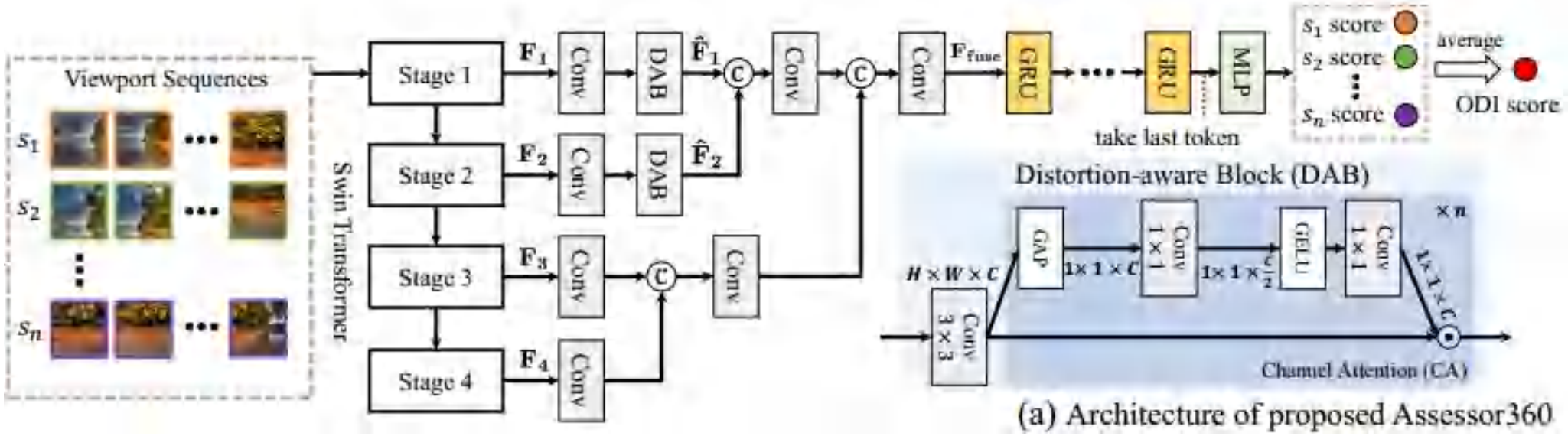
Swin Transformer 相比于ViT，将标准多头自注意力模块 (MSA) 替换为基于移位窗口的多头自注意力模块(W-MSA/SW-MSA)且保持其他部分不变。



3 多尺度特征聚合

对于Swin Transformer前两个阶段F1， F2的输出对失真模式更敏感，而最后两个阶段F3， F4的输出强调抽象特征。在融合多尺度特征之前，首先使用4个1 × 1的卷积层将输出的特征维数降至d。此外，为了进一步强调局部退化，文章设计并应用了失真感知块（DAB），该块包含一个3 × 3的卷积层，并进行了n次通道注意（CA）操作。该操作可以帮助模型更好地感知信道维度上的失真模式，实现失真感知能力。

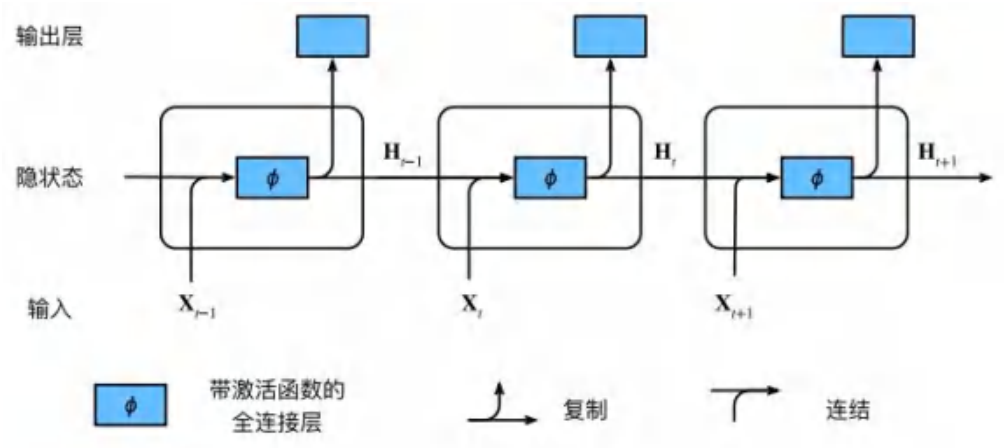
最后，将这些特征与全局平均池化（GAP）和多个1 × 1卷积层连接和集成，以获得与质量相关的表示。之后，每个聚合的特征将被发送给TMM进行视口序列质量评估。



(a) Architecture of proposed Assessor360

全景图像的浏览过程自然会导致时间相关性。近因效应表明用户更有可能评估受他们最近查看的视口影响的整体图像质量，特别是在长时间的探索期间。

为了对这种关系建模，引入门控循环单元(GRU)模块来学习视口转换。由于最后一个token编码了最近的信息，并且最后一个时间步的表示涉及序列的整个时间关系，使用MLP层将GRU模块输出的最后一个特征回归到序列质量分数。图为GRU模块的简图，将视口特征作为时间序列进行建模，得到最后一个token来进行质量回归。



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

- 全景图像
- 路径生成
- Assensor360
- **实验结果**
- 前景展望

3 实验结果

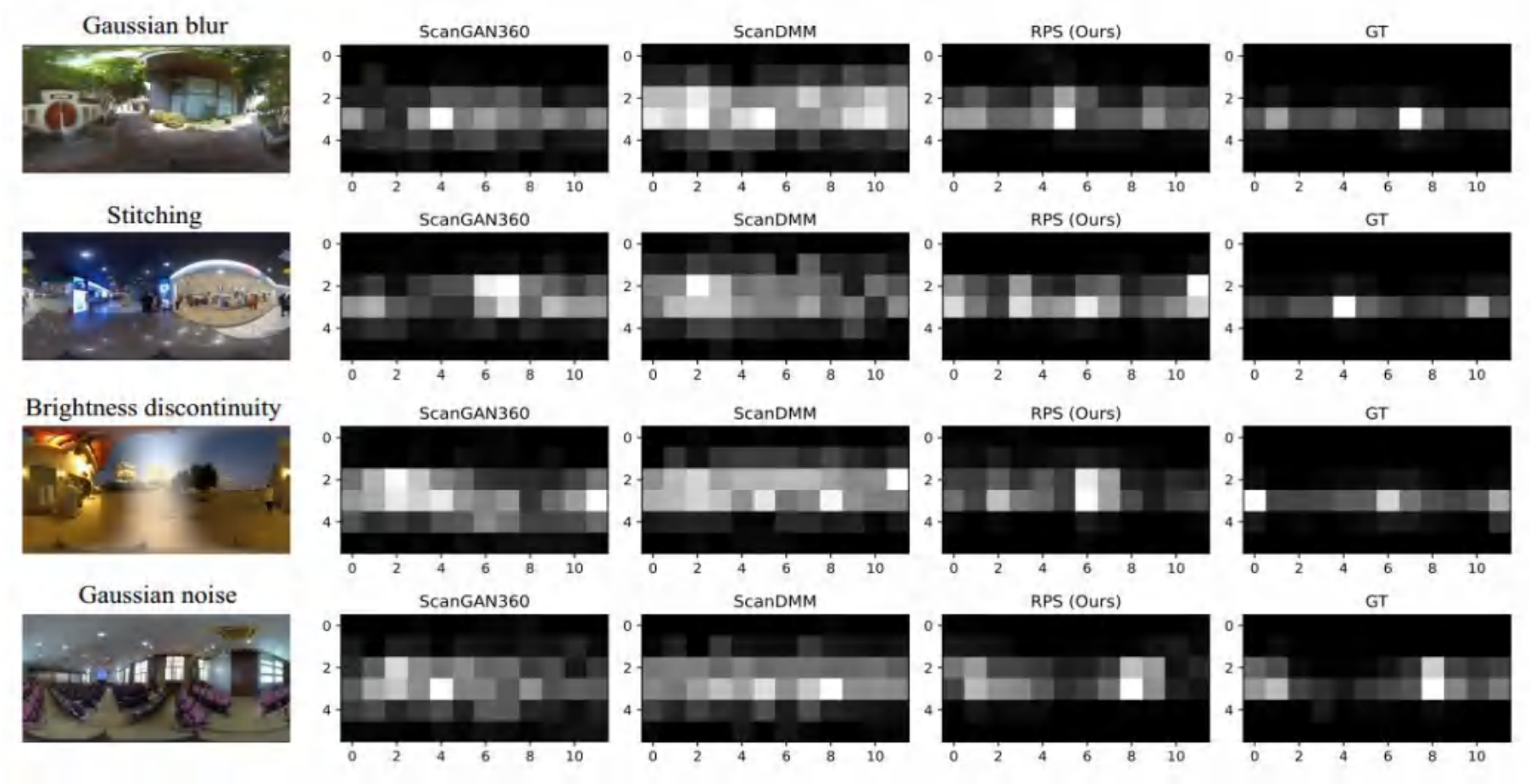
与其他全景图像质量评估模型的对比。

Type	Method	MVAQD		OIQA		IQA-ODI		CVIQD	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
FR-IQA methods	PSNR	0.8150	0.7591	0.3929	0.3893	0.4018	0.4890	0.8015	0.8425
	SSIM [39]	0.8272	0.7202	0.3402	0.2307	0.5014	0.5686	0.6737	0.7273
	MS-SSIM [40]	0.8032	0.7136	0.5750	0.5084	0.7434	0.8389	0.9218	0.9272
	WS-PSNR [37]	0.8152	0.7638	0.3829	0.3678	0.3780	0.4708	0.8039	0.8410
	WS-SSIM [51]	0.8236	0.5328	0.6020	0.3537	0.5325	0.7098	0.8632	0.7672
	VIF [35]	0.8687	0.8436	0.4284	0.4158	0.7109	0.7696	0.9502	0.9370
	DISTS [8]	0.7911	0.7440	0.5740	0.5809	0.8513	0.8723	0.8771	0.8613
	LPIPS [55]	0.8048	0.7336	0.5844	0.4292	0.7555	0.7411	0.8236	0.8242
NR-IQA methods	NIQE [26]	0.6785	0.6880	0.8539	0.7850	0.6645	0.5637	0.9337	0.8392
	BRISQUE [25]	0.8408	0.8345	0.8213	0.8206	0.8171	0.8651	0.8269	0.8199
	PaQ-2-PiQ [50]	0.3251	0.3643	0.1667	0.2102	0.0201	0.0419	0.7376	0.6500
	MANIQA [48]	0.5534	0.5718	0.4555	0.4171	0.2642	0.2776	0.6013	0.6142
	MUSIQ [19]	0.5436	0.6117	0.3216	0.3087	0.0565	0.0983	0.3483	0.3678
	CLIP-IQA [36]	0.5862	0.4941	0.2330	0.2551	0.0927	0.1929	0.4884	0.4347
	LIQE [57]	0.8837	0.7539	0.7634	0.7419	0.8851	0.9020	0.8594	0.8086
	SSP-BOIQA [58]	0.7838	0.8406	0.8650	0.8600	-	-	0.8614	0.9077
	MP-BOIQA [17]	0.8420	0.8543	0.9066	0.9206	-	-	0.9235	0.9390
	MC360IQA [30]	0.6605	0.6977	0.9071	0.8925	0.8248	0.8629	0.8271	0.8240
	SAP-net [46]	-	-	-	-	0.9036	0.9258	-	-
	VGCN [43]	0.8422	0.9112	0.9515	0.9584	0.8117	0.8823	0.9639	0.9651
	AHGCN [44]	-	-	0.9647	0.9682	-	-	0.9617	0.9658
	baseline w/ ERP	0.9076	0.9240	0.8961	0.8857	0.9098	0.9196	0.9330	0.9485
	baseline w/ CMP	0.8966	0.9324	0.9216	0.9170	0.9105	0.9122	0.9390	0.9442
	Assessor360	**0.9607**	**0.9720**	**0.9802**	**0.9747**	**0.9573**	**0.9626**	**0.9644**	**0.9769**

Generation Method	OIQA		MVAQD	
	SRCC	PLCC	SRCC	PLCC
Random Generation	0.9461	0.9444	0.9359	0.9543
StanGAN360	0.9705	0.9670	0.9493	0.9692
ScanDMM	0.9652	0.9634	0.9558	0.9612
RPS (Ours)	**0.9802**	**0.9747**	**0.9607**	**0.9720**

左表为与普通方法的对比，而上表为与同样使用路径生成的算法的对比。
可以看出，针对OIQA任务设计的RPS模块比随机生成或是其他任务的路径生成算法具备明显优势。

路径生成算法效果的对比可视化图。



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

- 全景图像
- 路径生成
- Assensor360
- 实验结果
- **前景展望**

5 前景展望

本文提出了基于实际评估流程的Assessor360，设计了递归概率采样（RPS）来生成基于语义场景和失真的视口序列。此外，我们还提出了带有扭曲感知块（DAB）的多尺度特征聚合（MFA）来结合视口的扭曲特征和语义特征。引入时序建模模块来学习视口的时序转换。

在基于路径生成的框架下，可能还有以下几种优化方向。

- RPS中的过渡方向和距离是固定的，导致视口之间的距离相等。而实际的观察路径中观察者对于不同位置的观察时间不同，并且视线并不是匀速移动的，更合理的视口生成建模会更好的模拟实际的观察过程。
- 本文中路径的生成只受到局部的纹理复杂度和赤道引导的控制，实际场景中人类的注意力转移存在多种诱导因素，这些都会影响到观察路径的生成。

图像数据增强

Image Data Augmentation

A8-王禹桥

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目 录

CONTENTS

数据增强定义

Data augmentation is a statistical technique which allows maximum likelihood estimation from incomplete data. Data augmentation has important applications in Bayesian analysis, and the technique is widely used in machine learning to reduce overfitting when training machine learning models, achieved by training models on several slightly-modified copies of existing data.

——Wikipedia

在机器学习和深度学习的领域中，数据增强（Data Augmentation）是一个重要的概念。它主要指的是通过增加原始数据集的多样性和丰富性，来提升模型的学习能力和泛化能力。

图像数据增强解决的问题

- **数据有限性**
  - 扩充训练数据集的大小和质量，帮助构建更好的深度学习模型。
- **过拟合问题**
  - 引入多样性来减少模型对训练数据的过度依赖，从而减少过拟合。
- **类别不平衡**
  - 生成少数类的样本，平衡类别分布，提高模型对少数类的识别能力。
- **数据标注成本高**
  - 在不增加标注成本的情况下，通过生成新的数据样本来扩充数据集。
- **图像识别中的挑战**
  - 图像识别任务中存在视角、光照、遮挡、背景、尺度等多种挑战。数据增强技术通过引入这些变化，帮助模型在面对这些挑战时表现更好。

图像数据增强方法类别

基于图像处理的数据增强技术

- **几何变换（Geometric Transformations）**
  - 包括平移、翻转、旋转、缩放、剪切等操作。
- **颜色空间变换（Color Space Transformations）**
  - 包括亮度调整、对比度调整、颜色抖动等操作。
- **核滤波（Kernel Filters）**
  - 通过高斯滤波器来模糊图像或通过锐化操作增强图像的边缘细节。
- **图像混合（Mixing Images）**
  - 将两张或多张图像按一定比例混合在一起或将多张图像拼接成一张大图像。
- **随机擦除（Random Erasing）**
  - 随机选择图像的一部分并将其擦除，以模拟遮挡情况。

基于深度学习的数据增强技术

- **特征空间增强（Feature Space Augmentation）**
  - 通过卷积神经网络提取图片的低维特征向量，再在特征向量上进行添加噪声、插值等操作。
- **对抗训练（Adversarial Training）**
  - 通过添加噪声生成对抗样本。
- **生成对抗网络（GAN-based Augmentation）**
  - 使用生成对抗网络生成新的图像样本。
- **神经风格迁移（Neural Style Transfer）**
  - 将一种图像的风格迁移到另一种图像上，生成新的图像。
- **元学习方案（Meta-learning Schemes）**
  - 一种学习如何学习的方法，旨在使模型能够快速适应新任务或新环境。

Image-Level Automatic Data Augmentation for Pedestrian Detection

Yunfeng Ma , Min Liu , Member, IEEE, Yi Tang , Xueping Wang , and Yaonan Wang

IEEE Transactions on Instrumentation and Measurement 2024

任务场景

- 拥挤场景下的行人检测（Pedestrian detection in crowded scenes）

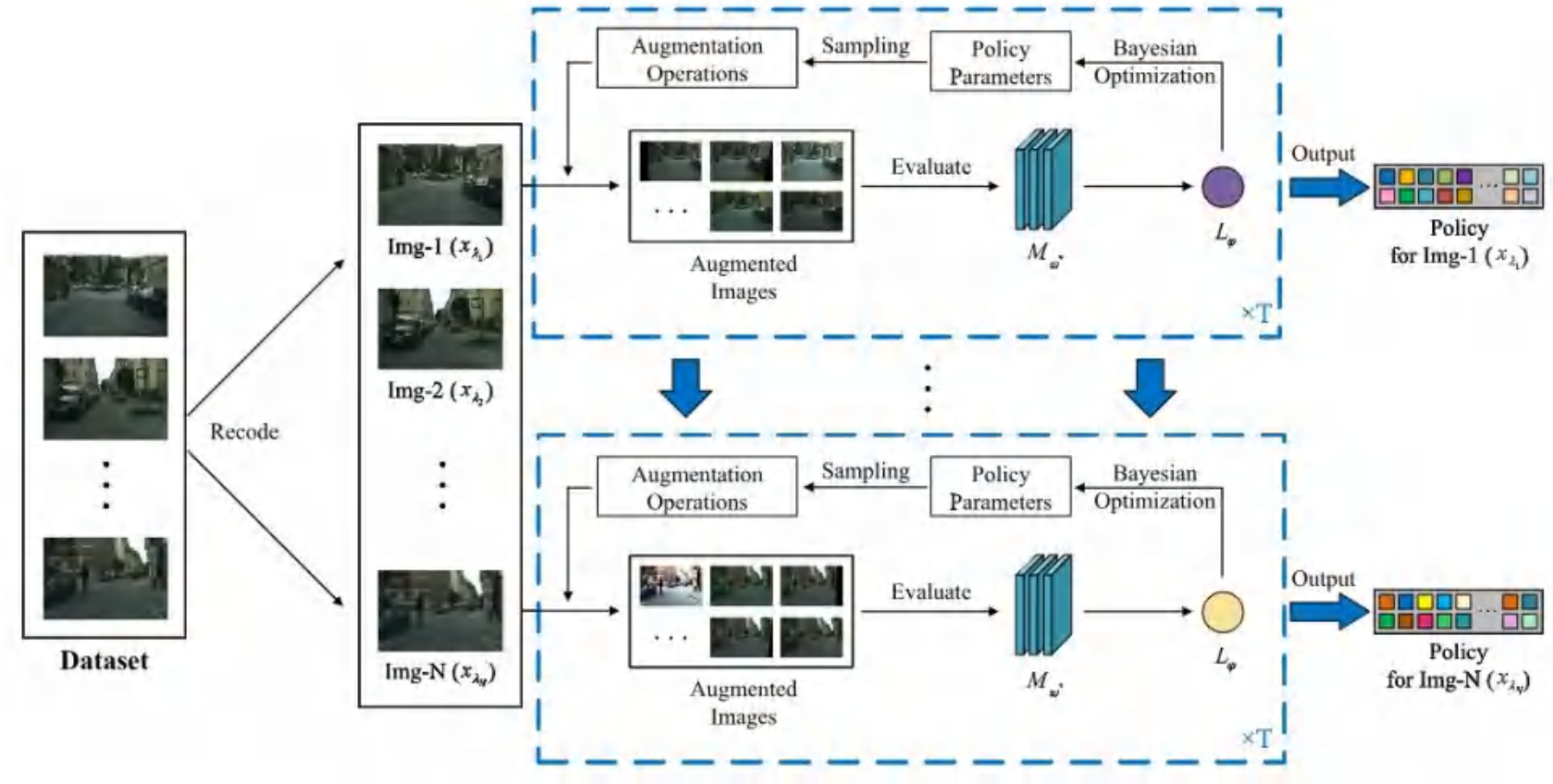
现存方法的不足

- 手动选择适合不同数据集的增强操作及相关参数，并最终将这些操作组合成有效的增强策略，需要大量专业知识并进行多次参数调整，耗时且困难。
- 现有的数据集级别的数据增强方法无法根据图像之间的差异自动调整增强策略，这会导致生成异常数据并影响模型的稳定性。

本文贡献

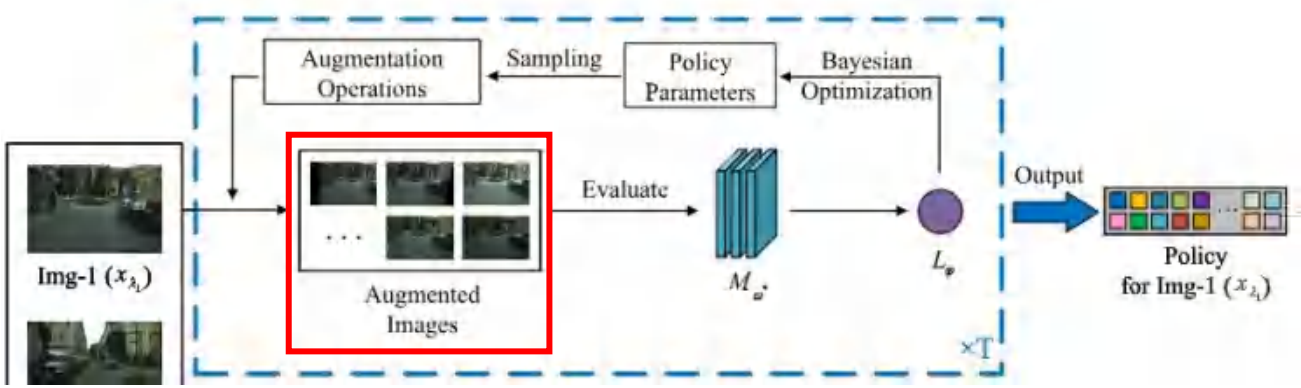
- 提出了一种新颖的图像级数据增强方法，该方法可以根据每个图像各自的特征搜索该图片进行数据增强的最佳策略。与对整个数据集使用相同的增强策略相比，为每张图像定制更细粒度的增强策略可以避免异常数据的产生，提高模型的稳定性。
- 首先通过构建基于分类分布的操作搜索空间，其中操作被采样的概率由它们各自的有效性决定。随后，设计了一种编码方法来重新编码图像和策略的索引，以实现它们之间稳定的匹配关系。最后，将贝叶斯优化集成到搜索框架中，以进行有效的策略挖掘，从而为每个图像生成最佳的增强策略。
- 在行人检测数据集CityPersons 和 CrowdHuman上进行了广泛的实验，，证明了该方法及其各个组成部分的有效性。

方法框架


Augmentation Operations
Sampling
Policy Parameters
Bayesian Optimization
Augmented Images
Evaluate
Output
Policy for Img-1 (x_{λ_1})
×T
Dataset
Recode
Img-1 (x_{λ_1})
Img-2 (x_{λ_2})
Img-N (x_{λ_N})
Augmentation Operations
Sampling
Policy Parameters
Bayesian Optimization
Augmented Images
Evaluate
Output
Policy for Img-N (x_{λ_N})
×T

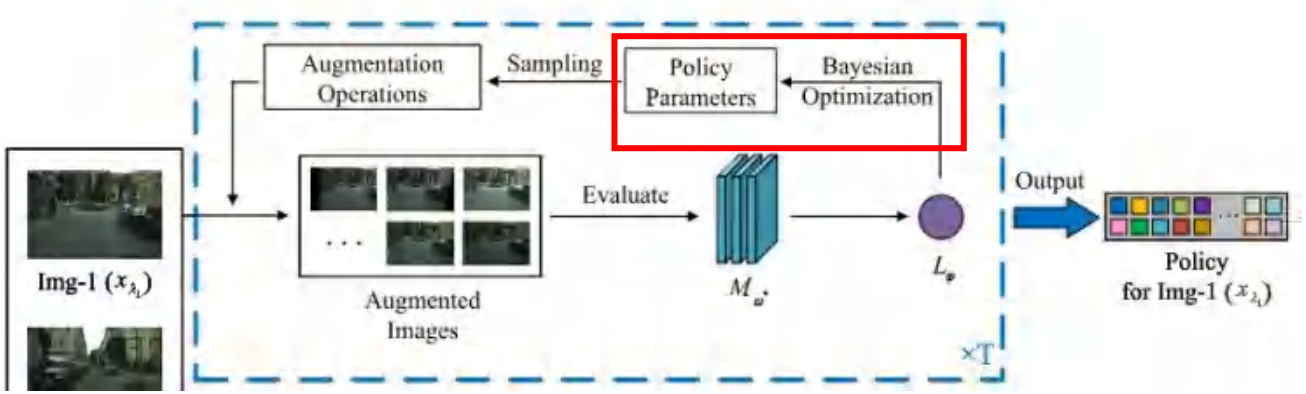
搜索空间（Search Space）

- 策略搜索空间由各种操作构成，每个操作有两个参数P、M，P为选择该操作的概率，M为该操作的幅度。
- 初始时每个操作的概率相同，所有操作概率之和相加为1。
- 在每一次的迭代优化中，效果较好的的操作所对应的概率增大，效果较差的操作所对应的改率降低。



优化流程（Optimization Process）

- 在训练阶段，数据集的打乱（Shuffle）操作会导致图片与策略之间的对应关系打乱。因此，本文提出了一种独立编码方法，能够作为中间编码，实现图片与策略的精确匹配。
- 采用贝叶斯优化作为优化方法，该方法迭代次数少，优化速度快。同时，可以利用CPU的多线程来进行并行计算，从而进一步显着提高效率。



优化目标

$$L_\psi = L_\tau = L_{\text{RPN}} + L_{\text{FRCNN}}$$

$$L_{\text{RPN}} = \frac{1}{N_{\text{cls}}^{\text{rpn}}} \sum_i -\log\left[g_i^* p_i + (1-g_i^*)(1-p_i)\right] + \lambda \frac{1}{N_{\text{reg}}^{\text{rpn}}} \sum_i g_i^* \text{SmoothL1}\left(t_i - t_i^*\right)$$

$$L_{\text{FRCNN}} = \frac{1}{N_{\text{cls}}^{\text{frcnn}}} \sum_i -\log\left[g_i^* p_i + (1-g_i^*)(1-p_i)\right] + \lambda \frac{1}{N_{\text{reg}}^{\text{frcnn}}} \sum_i g_i^* \text{SmoothL1}\left(t_i - t_i^*\right).$$

数据集（Dataset）

- CrowdHuman
- CityPerson

Datasets	Images	Pedestrians	Pedestrians/Img
CrowdHuman [23]	15000	339565	22.64
CityPersons [24]	2975	19238	6.47

评估指标（Metrics）

- 平均精准率（Average Precision，AP）
- 召回率（Recall）
- 每幅图像的误报率的对数平均值（MR^{-2}）

对比试验

Methods	MR^{-2}(%)	AP(%)	Recall(%)
DETR [48]	73.2	75.9	–
FSAF [45]	62.7	83.0	90.4
Grid R-CNN [46]	52.9	79.6	82.7
F-RCNN (Official) [23]	50.4	85.0	90.2
Adaptive NMS [30]	49.7	84.7	91.3
F-RCNN (Our impl.)	46.7	87.2	92.5
Repulsion Loss [11]	45.7	85.6	88.4
Deformable DETR [49]	43.7	91.5	–
PBM [47]	43.4	89.3	93.3
CrowdDet [10]	41.4	90.7	–
Iter-E2EDET [9]	37.7	94.1	–
FSAF + Ours	61.1	84.0	90.9
Grid R-CNN + Ours	52.0	80.3	83.7
F-RCNN + Ours	45.0	88.6	94.0
Iter-E2EDET + Ours	**36.8**	**94.4**	**97.4**

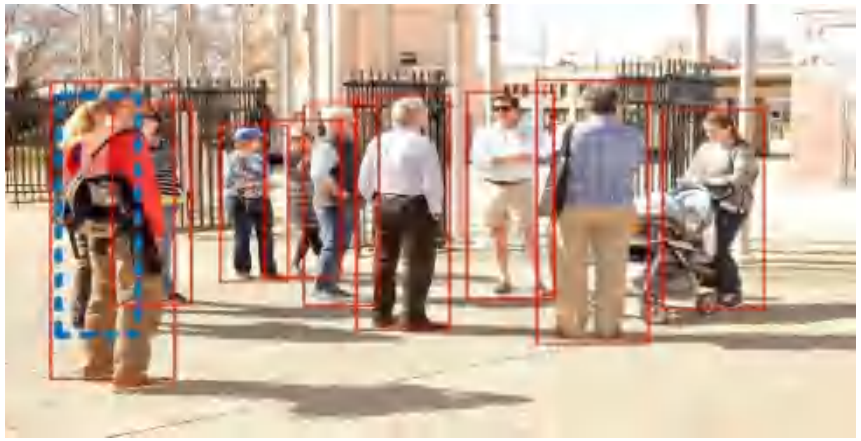
Methods	Backbone	R	HO
ATT-part [26]	VGG-16	16.0	56.7
Adapted Faster R-CNN [24]	VGG-16	15.4	–
Repulsion Loss [11]	ResNet-50	13.2	56.9
Adaptive NMS [30]	VGG-16	12.9	56.4
OR-CNN [12]	VGG-16	12.8	55.7
LBST [50]	ResNet-50	12.8	53.7
Faster R-CNN (Our impl.)	VGG-16	13.0	54.5
Faster R-CNN + Ours	VGG-16	**11.4**	**52.7**
CSP(Official) [51]	ResNet-50	11.0	–
CSP(Our impl. *)	ResNet-50	11.0	–
CSP(Our impl.) + Ours	ResNet-50	**10.2**	–

消融试验

Methods	ILA	SS	MR^{-2}(%)	AP(%)	Recall(%)
Baseline [23]	-	-	50.42	84.95	90.24
Baseline (Ours)	-	-	46.73	87.24	92.48
	✓	-	45.94	87.89	93.46
	-	✓	45.27	88.48	93.84
	✓	✓	**44.97**	**88.59**	**93.96**

可视化结果






低光图像增强

Low-Light Image Enhancement

黄梓颖　week13　计算机科学与技术前沿

任务概述

- **背景介绍**

  原因：环境或技术局限性，如照明不充分不均匀、曝光时间有限等

  表现：对比度降低、噪声和色偏等

  影响：影响美学质量，以及高层的任务

- **低光增强在干啥**

  形式上，低光带噪图像 → 正常光干净图像

  目的上，改善在光照不足环境中拍摄图像的感知或可解释性...

- **任务分解具体分析**

  低光 → 正常光

  带噪 → 干净

  RAW → RGB （*如果是涉及RAW域，而且需要转换的话）

Low-light image and video enhancement using deep learning: A survey, Li C, TPAMI2021

任务概述

- **应用及其需求的挑战性**

  应用上，计算摄影（如夜景模式）、计算感知（如自动驾驶）等

  一般的需求：改善图像质量，保证信息尽可能准确

  特定的需求：如果考虑到智能设备上，实时处理以及资源消耗也是需要关注的

- ~~其他需要注意的：~~

Low-light image and video enhancement using deep learning: A survey, Li C, TPAMI2021

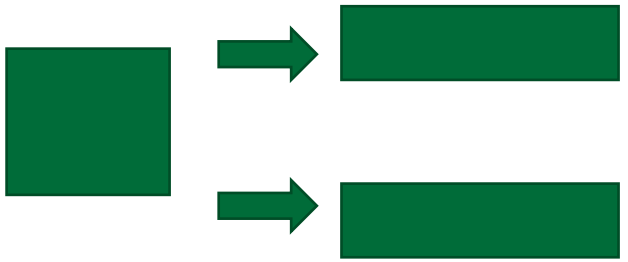
低光增强

任务：低光带噪图像→正常光干净图像

普通方法：直方图均衡化、伽马校正
（光照不一致）

传统认知方法：Retinex理论
即图像分解为反射图和照度图的乘积
反射图反映物体的固有属性，照度图仅与环境光照相关
（噪声和颜色失真，泛化能力差）

深度学习方法：
CNN、CNN+Retinex
（繁琐，远程依赖局限性）

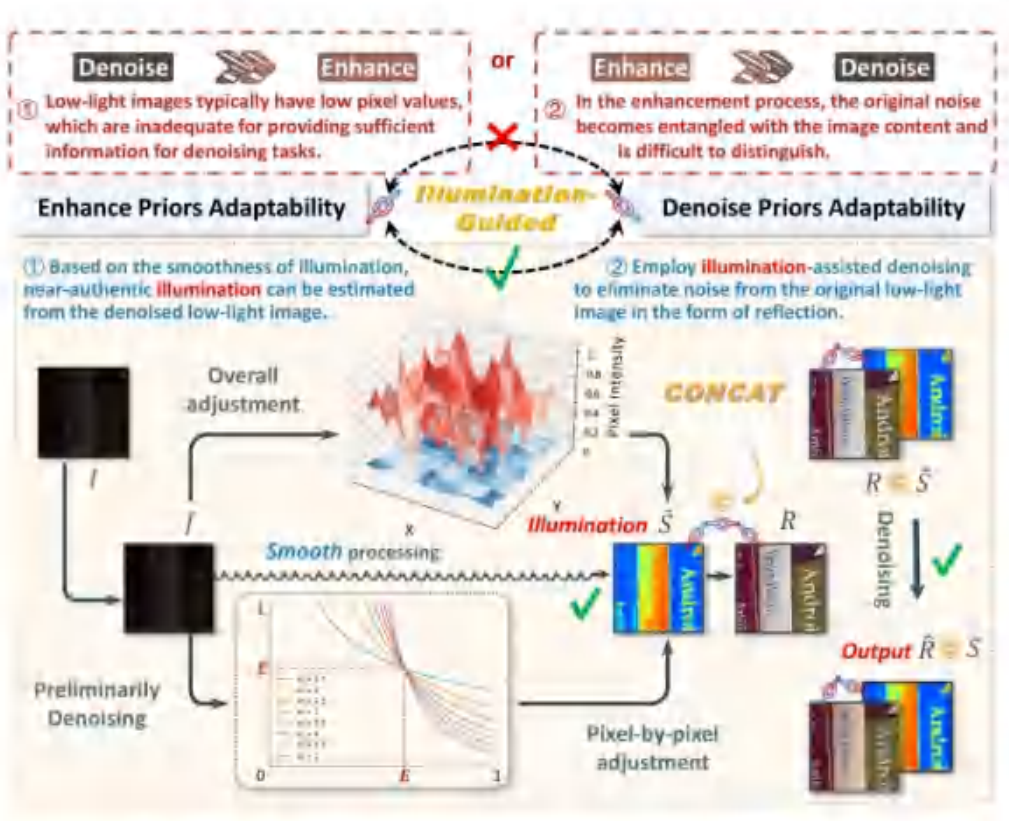


ZERO-IG方法

目前方法的局限性：
（1）Retinex方法依赖手工设计的约束，自适应性不够
（2）依赖成对/非成对训练数据，噪声情况复杂

☞zero-shot（零样本）

ZERO-IG整体流程思路
（I）ZS-N2N初步去噪
（II）估计光照
（III）去噪增强

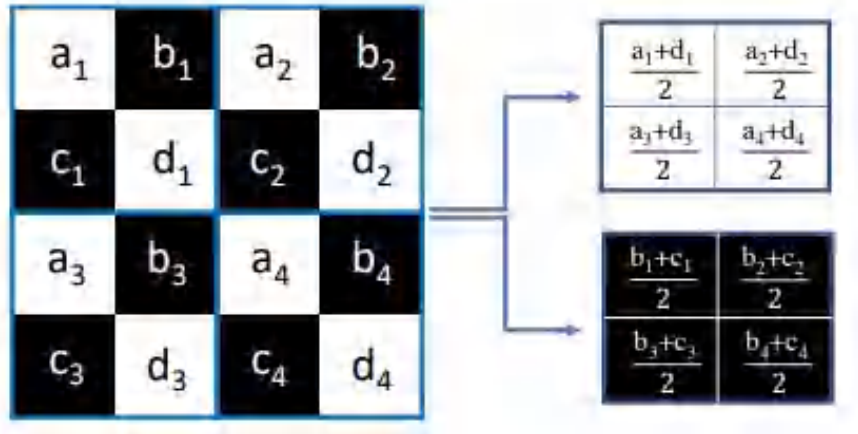


ZERO-IG: Zero-Shot Illumination-Guided Joint Denoising and Adaptive Enhancement for Low-Light Images

补充：N2N、ZS-N2N

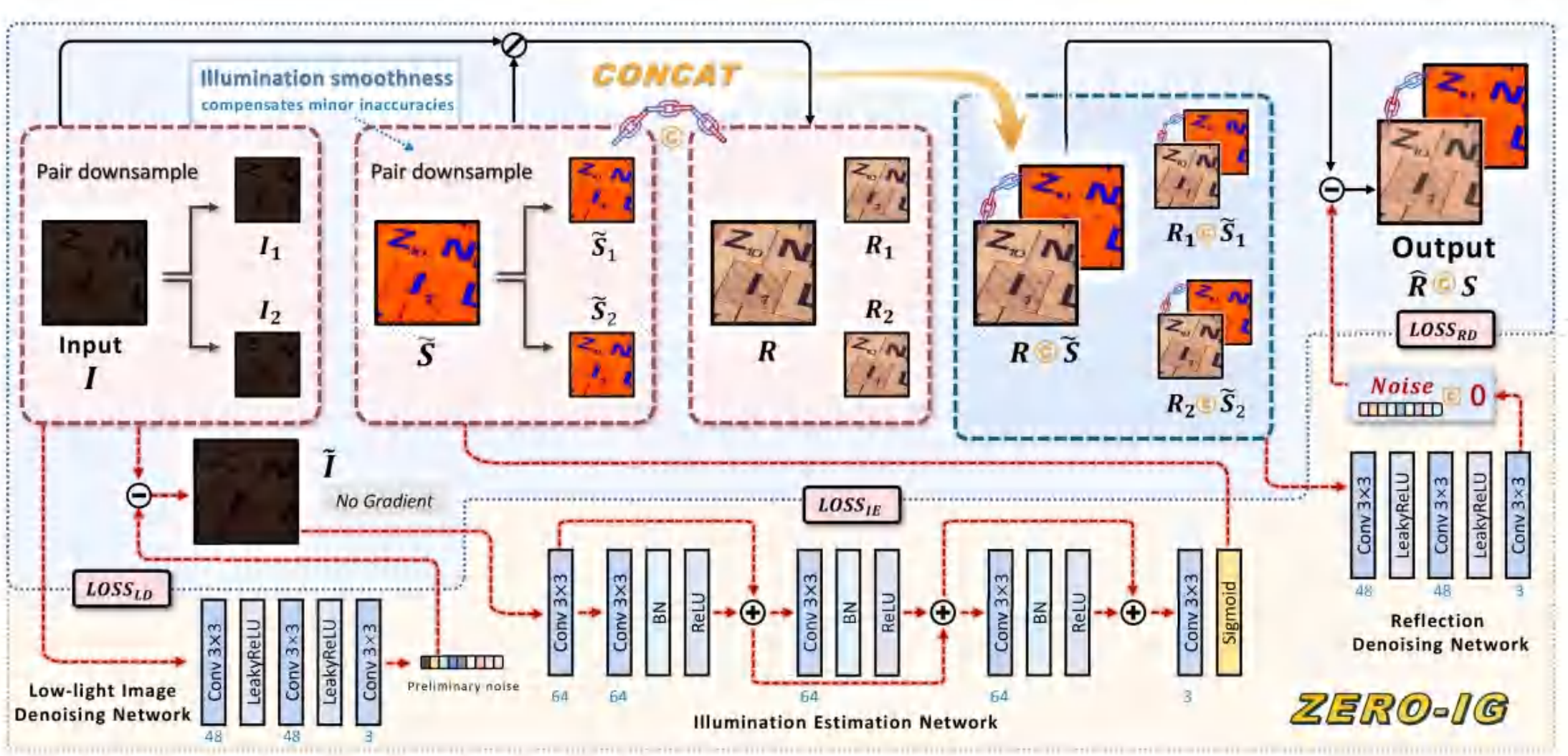
不需要GT的去噪方法
Noise2Noise只需要两张相同场景的独立噪声图像（ICML2018）
ZS-N2N只使用一张带噪图像（CVPR2023）

Noise2Noise理解：
（统计学问题）
样本足够多时，最小化loss，
倾向于输出所有可能的期望，即信号本身



Noise2Noise: Learning Image Restoration without Clean Data
Zero-Shot Noise2Noise: Efficient Image Denoising without any Data

ZERO-IG整体结构



ZERO-IG: Zero-Shot Illumination-Guided Joint Denoising and Adaptive Enhancement for Low-Light Images

ZERO-IG损失函数

LD-Net

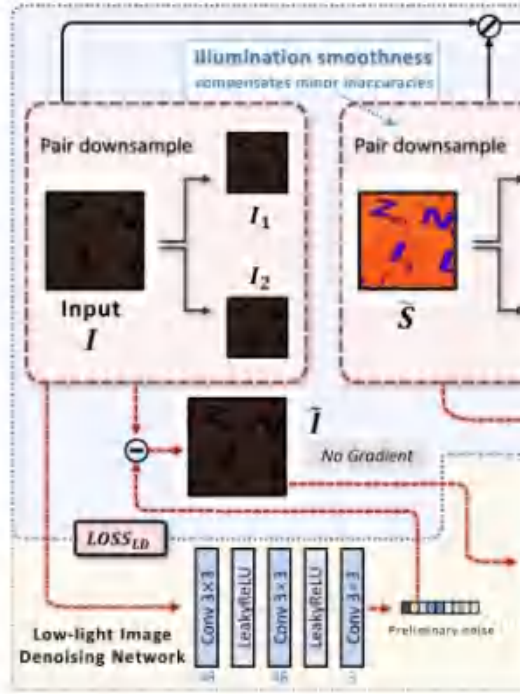
$$\mathcal{L}_{res}(\theta) = \|I_1 - f_\theta(I_1) - I_2\|_2^2 + \|I_2 - f_\theta(I_2) - I_1\|_2^2$$

残差损失：
I1和I2表示两个下采样的图像；
fθ表示噪声，网络估计的噪声

$$\mathcal{L}_{cons}(\theta) = \|I_1 - f_\theta(I_1) - G_1(I - f_\theta(I))\|_2^2 + \|I_2 - f_\theta(I_2) - G_2(I - f_\theta(I))\|_2^2$$

一致性损失：
I1和I2表示两个下采样的图像；
fθ表示噪声，网络估计的噪声；
G1和G2表示两个下采样；
先下采样后去噪 和 先去噪后下采样保持一致性

$$\mathcal{L}_{res}(\theta) + \mathcal{L}_{cons}(\theta).$$



ZERO-IG: Zero-Shot Illumination-Guided Joint Denoising and Adaptive Enhancement for Low-Light Images

N2N、L2损失

ZERO-IG损失函数

IE-Net

$$\mathcal{L}_{mar} = \|\hat{S} - \alpha^{-1}\|_2^2$$

整体调整损失：
亮度系数α；正常光亮度平面均值YH，设置0.5；
一阶段结果（低光）亮度平面均值YL；

$$\alpha = \bar{Y}_H / \bar{Y}_L$$

$$\mathcal{L}_{pix} = \|\hat{S} - \beta(\alpha\tilde{I})^{\alpha}\|_2^2$$

像素调整损失：
缩放因子β，良好水平E应该在调整后不变！设置为0.7

$$\tilde{I}(x,y) \circ \tilde{S}(x,y)^{-1} = \tilde{I}(x,y) \circ (\beta(\alpha\tilde{I}(x,y))^{\alpha})^{-1}$$

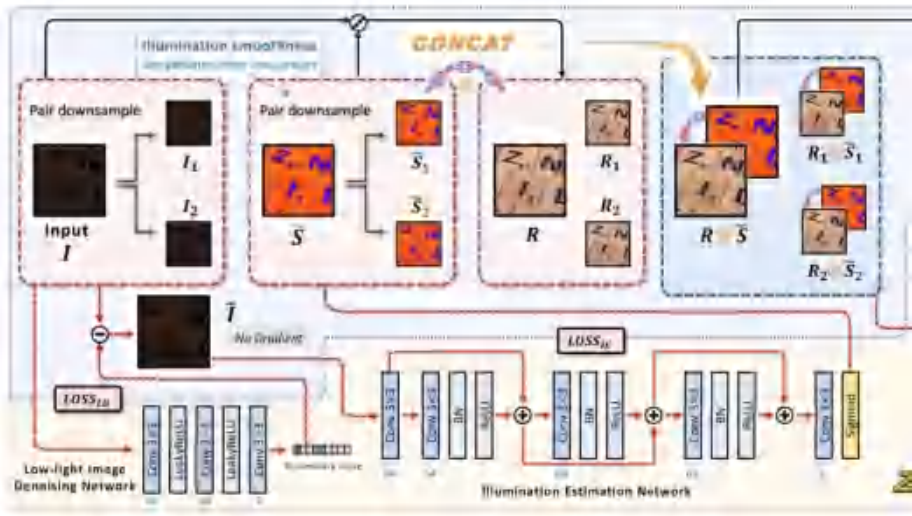
$$E = (\alpha^{-1}E) \circ (\beta E^{\alpha})^{-1}$$

$$\beta = \alpha^{-1}E^{-\alpha}.$$

$$\mathcal{L}_{smooth} = \sum_c (|\nabla_x \tilde{S}_c| + |\nabla_y \tilde{S}_c|)^2 + \sum_{i=1}^{N} \sum_j w_{i,j} |\tilde{S}_i - \tilde{S}_j|$$

平滑损失：
C表示三通道，N是像素总数，w是5×5高斯核

$$\mathcal{L}_{over} + \mathcal{L}_{pix} + \mathcal{L}_{smooth}$$



ZERO-IG: Zero-Shot Illumination-Guided Joint Denoising and Adaptive Enhancement f

ZERO-IG损失函数

RD-Net

$$\mathcal{L}_{res}(\hat{\theta}) = \|R_1 \oplus \bar{S}_1 - f_{\hat{\theta}}(R_1 \oplus \bar{S}_1) - R_2 \ominus \bar{S}_2\|_2^2 + \|R_2 \oplus \bar{S}_2 - f_{\hat{\theta}}(R_2 \oplus \bar{S}_2) - R_1 \ominus \bar{S}_1\|_2^2$$

残差损失：
1和2表示两个下采样的图像；
fθ表示噪声，网络估计的噪声
与第一部分类似

$$\mathcal{L}_{cons} = \|G_1(\hat{R}) - R_1\|_2^2 + \|G_2(\hat{R}) - R_2\|_2^2$$

一致性损失：
1和2表示两个下采样的图像；
fθ表示噪声，网络估计的噪声；
G1和G2表示两个下采样操作；
先下采样后去噪 和 先去噪后下采样 保持一致性
也是和第一部分类似，但是把fθ换了种写法...

$$\mathcal{L}_{\quad} = \|\quad - \quad\|_2^2$$

照度一致损失：
本阶段不应该改变照度

ZERO-IG: Zero-Shot Illumination-Guided Joint Denoising and Adaptive Enhancement for Low-Light Images

ZERO-IG损失函数

RD-Net

$$\mathcal{L}_{inter}=\sum_{i=1,2}\|G_i(\tilde{R})-(D\circ G_i(\tilde{R})+(1-D)\circ\mu_i)\|_2^2$$

交互去噪损失：
μ是5×5均值模糊版本，D表示G1和G2差距
σ是标准差，C避免除0
核心就是两个下采样保持一致

$$2\sigma_1\sigma_2(\sigma_1^2+\sigma_2^2+C)^{-1}.$$

$$\mathcal{L}_{var}=\frac{1}{M}\|\sum_{j=1}^{M}\sigma_j^2-\sum_{j=1}^{M}\sigma_j^2\|_2^2$$

局部方差损失：
M是5×5窗口数，σ分别表示R和n的方差
（带噪图像R局部方差与噪图n局部方差一致，
即小的局部区域接近)

$$\mathcal{L}_{color}=\|R_b-R_b\|_2^2$$

色彩损失：
B代表模糊化版本，G(k,l)是二维高斯模糊算子
色彩一致

$$R_b(i,j)=\sum_{k,l}R(i+k,j+l)\cdot G(k,l)$$

ZERO-IG: Zero-Shot Illumination-Guided Joint Denoising and Adaptive Enhancement for Low-Light Images

ZERO-IG损失函数

RD-Net

$$\mathcal{L}_{rec}(\hat{\theta}) = \|R_1 \oslash \bar{S}_1 - f_\theta(R_1 \oslash \bar{S}_1) - R_2 \oslash \bar{S}_2\|_2^2 + \|R_2 \oslash \bar{S}_2 - f_\theta(R_2 \oslash \bar{S}_2) - R_1 \oslash \bar{S}_1\|_2^2$$

$$\mathcal{L}_{cons} = \|G_1(\hat{R}) - R_1\|_2^2 + \|G_2(\hat{R}) - R_2\|_2^2$$

$$\mathcal{L}_{inter} = \sum_{i=1,2} \|G_i(\hat{R}) - (D \circ G_i(\hat{R}) + (1-D) \circ \mu_i)\|_2^2$$

$$\mathcal{L}_{var} = \frac{1}{M}\|\sum_{j=1}^{M} \sigma_j^2 - \sum_{j=1}^{M} \bar{\sigma}_j^2\|_2^2$$

$$\mathcal{L}_{color} = \|R_0 - R_\theta\|_2^2$$

$$\mathcal{L}_{rec}(\hat{\theta}) + \mathcal{L}_{cons} + \mathcal{L}_{ill} + \mathcal{L}_{inter} + \mathcal{L}_{var} + \mathcal{L}_{color}$$

ZERO-IG: Zero-Shot Illumination-Guided Joint Denoising and Adaptive Enhancement for Low-Light Images

定量结果

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

Datasets	Type	Metrics	*Supervised Learning Methods*				*Unsupervised Learning Methods*					
			URtinexNet (cvpr2022)	LLFlow (AAAI2022)	SNR-aware (cvpr2022)	KinD++ (IJCV2021)	SCI (cvpr2022)	ZeroDCE++ (TPAMI2021)	PairLIE (cvpr2023)	RUAS (cvpr2021)	ZERO-IG IE-Net	ZERO-IG
SIDD	–	PSNR ↑	16.2600	14.6107	14.8907	16.5524	15.5257	15.6757	17.0254	12.5997	15.1154	18.9849
		SSIM ↑	0.4247	0.4401	0.6010	0.5799	0.3537	0.3561	0.5266	0.4287	0.2865	0.6253
	Followed by ZS-N2N	PSNR ↑	16.5879	17.2651	14.8641	16.6056	15.7053	16.4329	17.1241	12.7437	16.0232	18.9849
		SSIM ↑	0.4715	0.5097	0.5994	0.5982	0.3610	0.45883	0.5497	0.4459	0.4577	0.6253
LOL	–	PSNR ↑	20.1405	24.0841	24.6097	17.6476	14.7839	15.1416	18.4684	16.5976	17.6255	22.1751
		SSIM ↑	0.8221	0.8601	0.8494	0.7714	0.52544	0.5657	0.7426	0.6559	0.4566	0.7719
LSRW-Huawei	–	PSNR ↑	18.1566	19.2005	17.6203	17.0251	15.7003	16.3821	18.9887	15.7422	17.6842	19.5414
		SSIM ↑	0.5464	0.5419	0.5781	0.4993	0.4279	0.4696	0.5502	0.4976	0.4101	0.5944
LSRW-Nikon	–	PSNR ↑	15.9870	15.3675	15.9362	15.4706	14.5542	15.2770	15.5219	12.2104	15.3901	16.6157
		SSIM ↑	0.4425	0.4491	0.4691	0.4411	0.4065	0.4129	0.4271	0.4394	0.3848	0.4706

比其他无监督方法更好，同时也与有监督方法相当

定性结果

Figure 6. Visual comparisons on low-light images with uneven brightness from LIME [7], DICM [13], LOL [28] and LSRW [8] datasets.

亮度合适不过曝，
噪声有效抑制

ZERO-IG: Zero-Shot Illumination-Guided Joint Denoising and Adaptive Enhancement for Low-Light Images

未来发展

- **目前存在的问题：**

  （1）宏观层面，未知的退化信息

  比如噪声、伪影和色偏等；还有泛化

  （2）微观层面，差异性的增强

  比如过饱和欠曝问题

  （3）跳出图像之外，其他数据或模态信息利用不完全

  比如视频用相邻帧等

  （4）固有问题/语义相关，网络区分信息

  比如黑发和黑夜、沥青路和噪声等

未来发展

- **未来研究方向**

  整体上来看，有不少工作还可以展开。

  （1）学习策略（更加泛化）

  例如无监督、零参考

  （2）网络结构（并非主要目标，而是应该当成工具即可）

  例如Transformer、深度可分离、NAS神经结构搜索

  （3）数据集（不应该盲目增大规模）

  体现真实性和**多样**性

  （4）损失函数

  更适合LLE的、神经网络近似人类感知

  （5）评价指标

  更好地衡量增强结果，图像质量评价IQA

  IQA中也有两方面，主观质量，以及高级视觉任务的影响

未来发展

- **未来研究方向**

  整体上来看，有不少工作还可以展开。

  （6）视频

  除了图像处理关注点之外，相邻帧间的关系需要额外考虑

  （7）语义信息

  超分和人脸修复等有类似工作

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

Thanks

汇报人：董静涛 王梓臣 刘勇奇 王禹桥 黄梓颖

时　间：2024-11-18