



# 工作汇报



孙梓轩



2024.9.16





南京邮电大学  
Nanjing University of Posts and Telecommunications

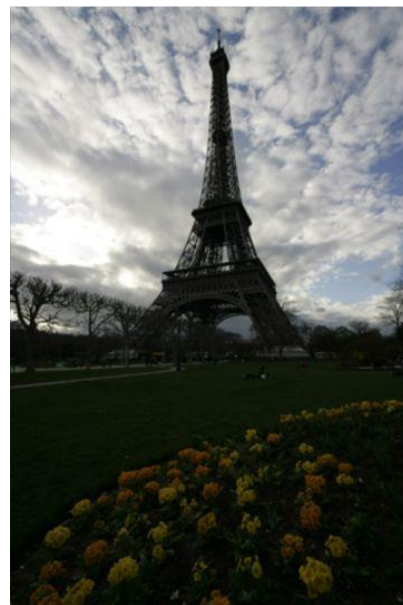
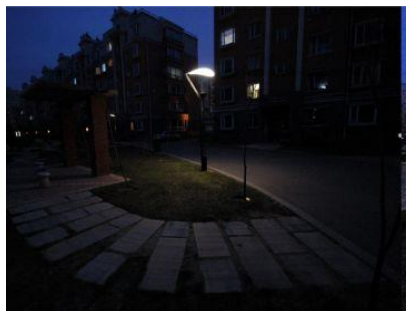
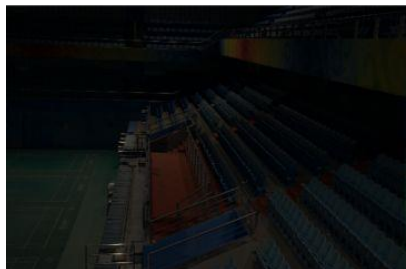
# 1

## 研究背景与现存问题

厚德弘毅 求是笃行

# 1

## 低光照图像增强

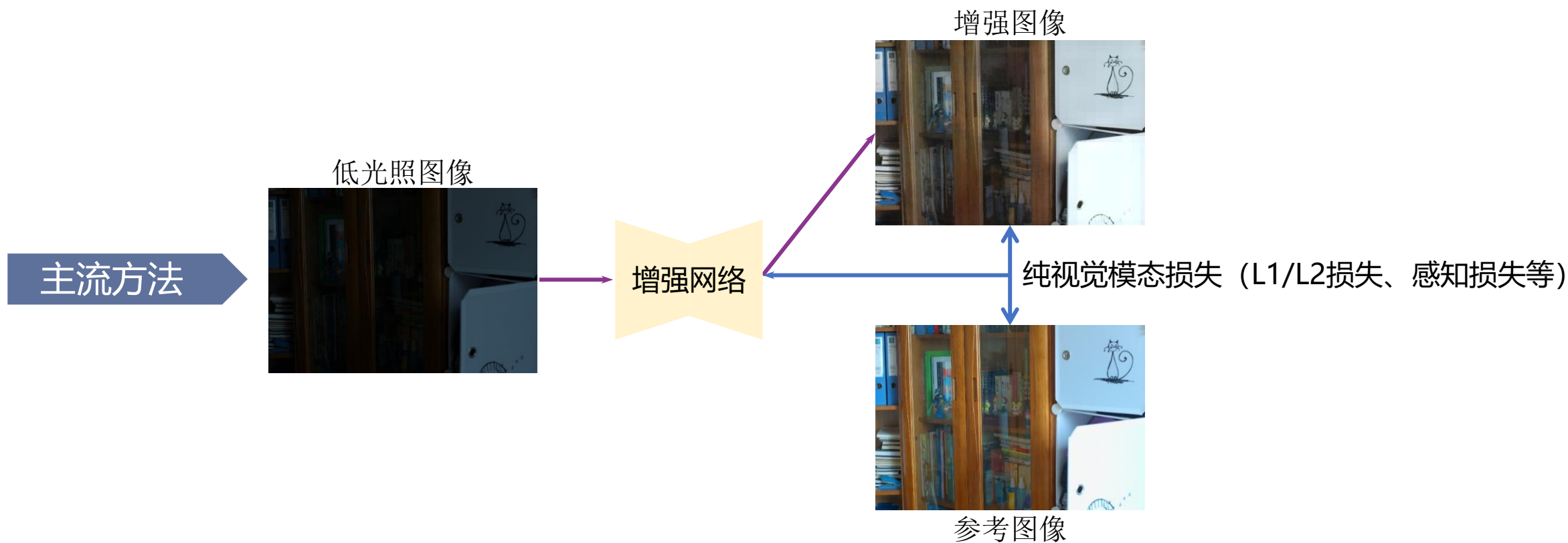


**任务目的：** 将低光照的图像恢复至正常光照图像



## 1

## 现存问题



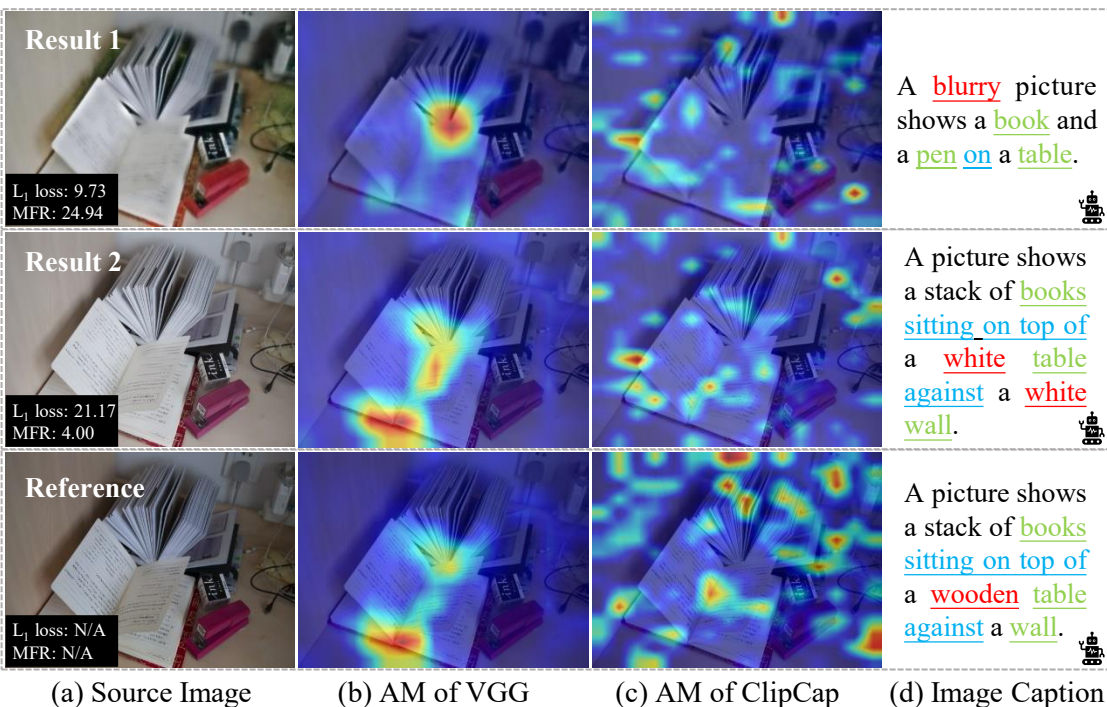
## 提出问题

大多数现有的低光照图像增强深度学习方法通常仅依赖于**视觉线索**来监督网络，忽略了**多模态语义**在指导增强过程中的重要性。这种**单一模态**的监督导致了图像**细节重建不足**。

## 1

## 分析问题

## 为什么单一模态的监督存在问题？



大多数现有方法倾向于使用**像素级**的目标函数（如 **L1** 和 **L2**）来建立低映射关系。

这些损失函数的**负梯度方向**常常偏离了**视觉感知提升**的方向。

为了缓解上述问题，一些方法引入了**感知损失**作为额外的正则化项，以增强视觉保真度。

感知损失借助预训练的**分类**模型进行特征正则化，关注的是图像**主体的类别**信息。对图像**细节、物体颜色和背景**不敏感，而这些都是视觉感知的重要组成部分。

一些方法利用**CLIP**来探索**语言模态**信息在像素级图像增强中的潜力，使用**CLIP图像编码器**为模型提供视觉-语言信息。

多项研究揭示了多模态对比模型中存在着**模态差距**。这差距阻碍了低光照图像增强模型有效利用视觉编码来捕捉**文本语义特征**。

提出了一种**多模态特征正则化**方法，利用**预训练的图像字幕**模型有效地将丰富的视觉-语言信息注入到像素级图像增强模型中。

通过对增强图像与参考图像之间的**多模态特征**进行**一致性约束**，预训练的 MLLMs 提供了对**局部退化敏感**的信息。



南京邮电大学  
Nanjing University of Posts and Telecommunications

# 2

## 研究方法与实验结果

厚德弘毅 求是笃行



## 2

## Image Caption模型介绍

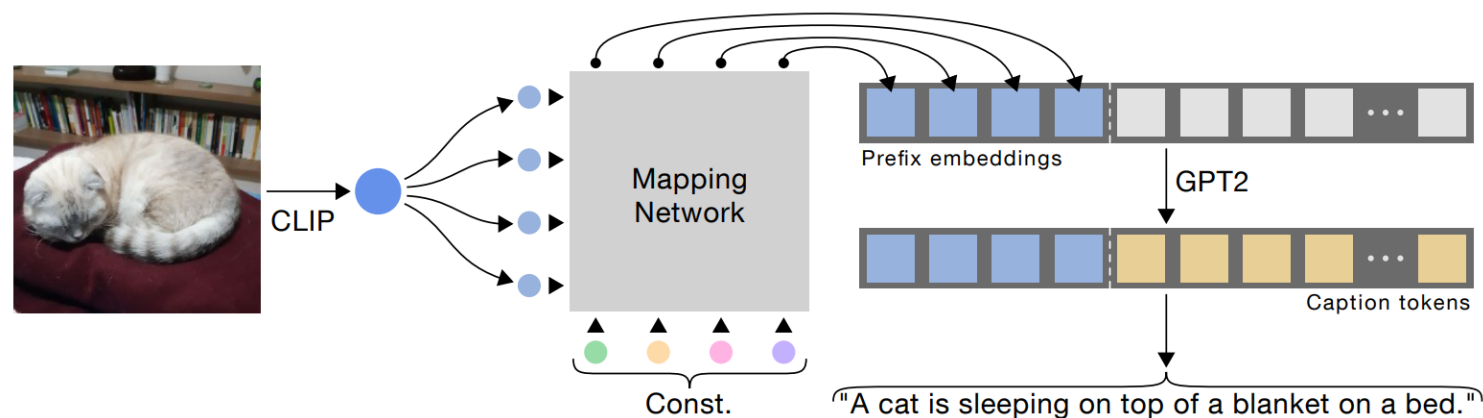
## 定义

In image captioning, the task is to provide **a meaningful and valid caption** for a given input image in a **natural language**.

生成自然语言描述需要结合考虑图片的信息，涉及到更加复杂的**视觉模式**。因此，Image Caption模型需要在复杂的多元数据中寻找**关联性和细节**。

## ClipCap: CLIP Prefix for Image Captioning

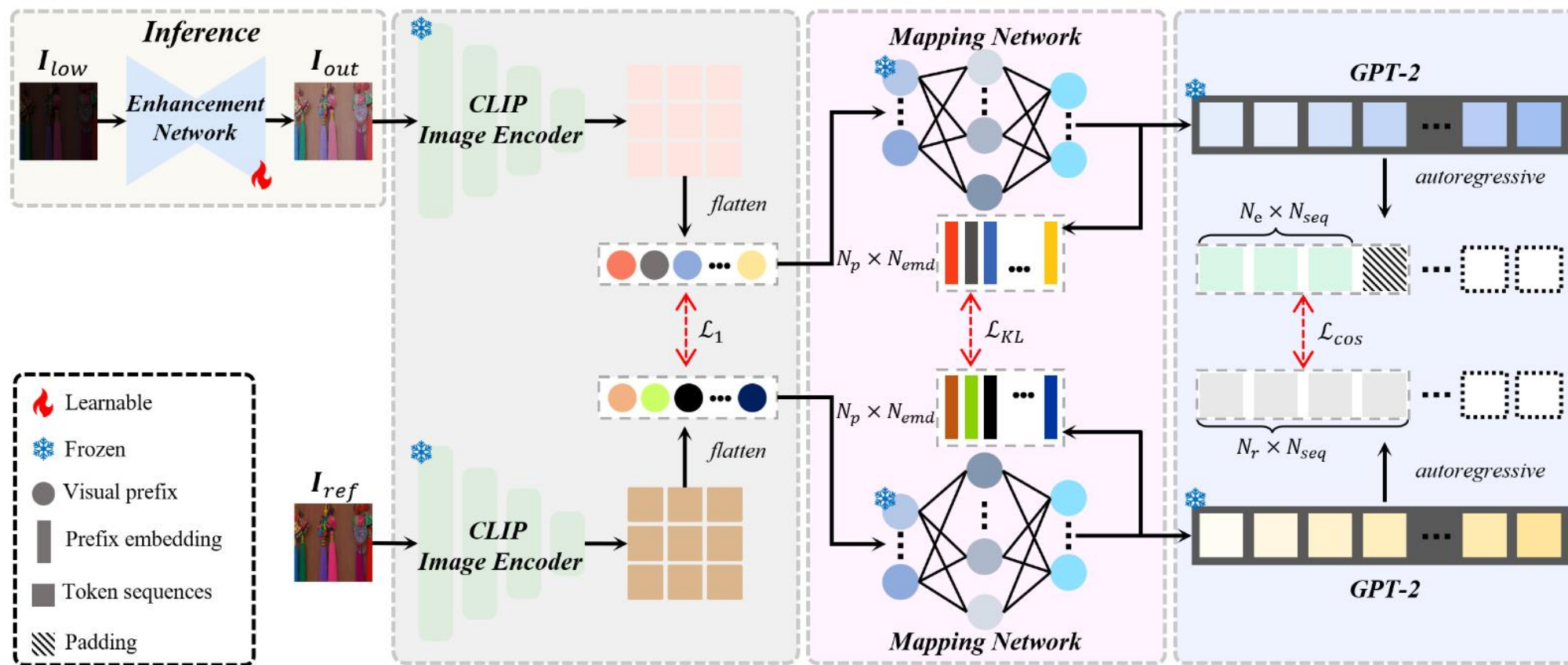
Ron Mokady\* Amir Hertz\* Amit H. Bermano  
The Blavatnik School of Computer Science, Tel Aviv University



将增强图像和参考图像在三个子空间中的多模态特征进行一致性约束

- 1、CLIP模型提取的图像visual cues
- 2、映射空间投影的Prefix embeddings
- 3、GPT2模型编码的文本序列特征

## Pre-trained Image Caption Model Guided Multi-modal Feature Regularization for Low-light Image Enhancement





Methods	PSNR	SSIM
RetinexNet <small>BMVC'18</small>	16.63	0.495
<b>RetinexNet-TL(Ours)</b>	18.92 (+2.29)	0.726 (+0.231)
MIRNet <small>ECCV'20</small>		
<b>MIRNet-TL(Ours)</b>		
RCTNet <small>ICCV'21</small>		
<b>RCTNet-TL(Ours)</b>		
IAT <small>BMVC'22</small>	20.74	0.732
<b>IAT-TL(Ours)</b>	21.55 (+0.81)	0.799 (+0.067)
SNR <small>CVPR'22</small>	22.59	0.827
<b>SNR-TL(Ours)</b>	23.76(+1.17)	0.830(+0.003)
LLFormer <small>AAAI'23</small>	22.94	0.799
<b>LLFormer-TL(Ours)</b>	23.43(+0.49)	0.811(+0.012)
UHDFour <small>ICLR'23</small>	22.05	0.801
<b>UHDFour -TL(Ours)</b>	22.90(+0.85)	0.814(+0.013)
Retinexformer <small>ICCV'23</small>	23.48	0.826
<b>Retinexformer-TL(Ours)</b>	24.25(+0.77)	0.838(+0.012)



南京邮电大学  
Nanjing University of Posts and Telecommunications

3

总结

厚德弘毅 求是笃行

## 现存问题

大多数现有方法仅依赖于**视觉线索**来监督网络，忽略了**多模态语义**在指导增强过程中的重要性。这种单模态的监督导致了图像**细节重建**的质量较差。

## 解决方法

引入了预训练的**图像字幕**模型，并提出了一种**多模态特征正则化**方法。预训练的多模态大型语言模型能够实现**视觉与语言**之间的细粒度对齐。

## 算法细节

(1) 对由 **CLIP** 图像编码器提取的**视觉前缀**进行正则化，以增强模型的视觉-语言感知能力；(2) 对由**映射网络**投影的**前缀嵌入**进行正则化，以保留并传递原始图像的语义信息；(3) 对在 **GPT-2** 语言空间中编码的**序列特征**进行正则化，以提高增强图像细节的准确性和完整性。

## 实验结果




广泛的实验表明，配备有多模态特征正则化方法的模型在**多个数据集**上都显著**优于基线模型**。



进一步的思考，引申出来的一个问题：

当前像素级图像增强评价指标也存在片面性！

经过调研，发现有论文提出这个问题，利用MLLM量化评分，  
获得与人类视觉更为一致的评分标准

Reference:  Image A:  Image B: 

User

Compare the quality of Image A with Image B and provide a comprehensive explanation for your comparison.

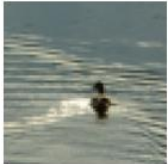
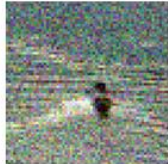



DepictQA

Both images are nearly identical in terms of brightness distortion, color distortion, artifacts, and blurring, with no significant differences. However, Image A *significantly underperforms* Image B in terms of *noise*, which greatly impacts the overall image quality. Particularly in representation of *architectural textures*, the texture in Image A is *completely damaged*, while that in Image B remains *clear and distinguishable*. Therefore, *Image B is clearly superior to Image A* in overall quality.

	PSNR↑	SSIM↑	LPIPS↓	PieAPP↓	Our DepictQA	Humans
Img.A	✓ 20.60	✓ 0.47	✓ 0.34	✓ 0.66		
Img.B	18.10	0.12	0.41	1.18	✓	✓

(a)

Reference:  Image A:  Image B: 

User

Which image has a better quality, Image A or Image B? Explain the comparison result with detailed factors.



DepictQA

Although Image A *slightly outperforms* Image B in terms of *color distortion and artifacts*, it *falls short* in the aspect of *noise*. Despite this, Image A still demonstrates a slight advantage over Image B in terms of overall distortion. Regarding the degree of texture damage, both images exhibit similar performance, with the *water surface texture barely recognizable*. Therefore, despite Image A's inferior performance in noise, its advantages in color distortion and artifacts contribute to its *slight superiority over Image B*.

	PSNR↑	SSIM↑	LPIPS↓	PieAPP↓	Our DepictQA	Humans
Img.A	19.03	0.37	0.45	1.83	✓	✓
Img.B	✓ 21.45	✓ 0.67	✓ 0.29	✓ 1.22		

(b)



# 感谢批评指正！

---

WE  
LOVE  
NJUPT

厚德弘毅 求是笃行