产教融合研究与实践
Research and Practice of Industry-Education Integration
Vol. 4, No. 3 May. 2026
DOI: https://10.58244/rpiei.263854

图像生成与街景分割技术研究参与过程的学习经验总结

钟晗熙

西安工业大学　陕西　西安　中国

摘　要　作为计算机科学与技术专业的本科生，我有幸先后参与了《智能艺术创造力：基于人工智能技术的创造性图像生成技术研究》与《物联网智能实时街景监测识别系统：基于改进 DeeplabV3 模型的高效街景图像分割技术研究》两项课题研究。本文系统梳理了从对 Transformer、语义分割仅有模糊认知的初学者，到逐步深入理解 StyTR-2 与改进 DeeplabV3 模型架构的设计哲学，再到亲身参与实验验证与论文撰写的完整历程。文章将为期一年的科研实践划分为“基础理论系统性补课”“模型架构理解与创新探索”“实验验证与结果分析”“总结反思与能力沉淀”四个关键阶段，详细记录了本科生初次涉足深度学习科研领域时，如何跨越理论到实践的鸿沟，从被动接收知识转向主动探索未知的认知路径与成长轨迹。

关键词　学习经验；图像风格迁移；图像分割；本科生科研；深度学习实践；StyTR-2；DeeplabV3
文章编号　056-2026-3854

Learning Experience Summary on Participating in Image Generation and Street Scene Segmentation Research

Hanxi Zhong

Xi'an Technological University, Shaanxi 710021, China

Abstract　As an undergraduate majoring in Computer Science and Technology, I have been fortunate to participate in two research projects successively, namely Intelligent Artistic Creativity: Research on Creative Image Generation Technology Based on Artificial Intelligence and Intelligent Real-time Street Scene Monitoring and Recognition System for the Internet of Things: Research on Efficient Street Scene Image Segmentation Technology Based on the Improved DeeplabV3 Model. This paper systematically reviews my complete research journey. Starting as a beginner with only a superficial understanding of Transformer and semantic segmentation, I gradually gained in-depth insight into the design philosophy of the StyTR-2 and improved DeeplabV3 model architectures, and further engaged in experimental verification and paper writing. The one-year scientific research practice is divided into four key stages: systematic supplementation of basic theories, comprehension of model architectures and innovative exploration, experimental verification and result analysis, as well as summary, reflection and competency accumulation. This article elaborately records the cognitive path and growth trajectory of an undergraduate engaging in in-depth learning research for the first time. It illustrates how I bridged the gap between theory and practice, and transformed from passive knowledge reception to active

收稿日期：2026-01-27　**录用日期**：2026-04-19

通讯作者：钟晗熙；**单位**：西安工业大学　陕西　西安

基金信息：西安工业大学2025年国家级大学生创新创业训练计划项目(编号：202510702030)

exploration of unknown research frontiers.

Keywords: Learning Experience; Image Style Transfer; Image Segmentation; Undergraduate Research; Deep Learning Practice; StyTR-2; DeeplabV3

1 引言

全大二学年结束后的暑期，我有幸加入研究团队，参与两项聚焦计算机视觉前沿领域的课题研究：一项是基于 Transformer 架构的图像风格迁移技术研究，另一项是针对物联网场景的高效街景图像语义分割研究。彼时，我虽已修完《Python 程序设计》与《数据结构》等专业基础课，但对深度学习的认知几乎完全停留在理论介绍层面。我知晓卷积神经网络（CNN）的存在，也听过反向传播算法的原理，但从未真正动手搭建过一个完整的深度学习模型，更遑论参与真实的科研项目。

初次接触导师提供的技术文献——《StyTR-2: Image Style Transfer with Transformers》 与《DeepLab: Semantic Image Segmentation with Deep Convolutional Nets》时，我的内心充满了困惑与茫然。Transformer 的自注意力机制为何能替代卷积？内容感知位置编码（CAPE）是如何工作的？ DeeplabV3 中的空洞卷积与空间金字塔池化究竟解决了什么问题？这些如迷雾般的问题让我深刻意识到自身知识储备的匮乏。然而，正是这种强烈的“知识欠缺的紧迫感”，驱动我制定并执行了一个系统地学习计划，决心从最基础的概念入手，循序渐进地揭开深度学习的神秘面纱。

回顾这为期一年的科研参与经历，我的学习路径可清晰地划分为四个阶段。本文旨在按照这一脉络，详细复盘我在每个阶段的认知突破、技能积累与思维转变，既是对个人成长的总结，也期望为同处于科研起步阶段的同学提供一份真实的参考样本。

2 学习路径与过程

2.1 第一阶段：基础理论的系统补课——跨越“看不懂”的门槛

参与科研的第一道门槛，是大量陌生概念带来的认知冲击。我意识到，若不先填补基础理论的缺口，后续的实验与论文工作无异于空中楼阁。因此，我花费了近一个月的时间，有针对性地开展了深度学习基础的“扫盲”工作。

首先是卷积神经网络原理的深化。课堂上的 CNN 知识相对浅显，而科研要求我准确理解特征图、感受野、通道数等概念背后的具体物理含义。我重新研读了经典文献，并对照 PyTorch 官方教程，逐行梳理了简单 CNN 模型的构建逻辑。我弄懂了卷积层如何提取特征、池化层如何降低维度、激活函数如何引入非线性。更重要的是，我理解了深层网络能够提取更抽象语义特征的原理——这为我后续理解为何 StyTR-2 需要分层设计、为何 DeeplabV3 需要多尺度特征融合奠定了坚实的认知基础。

其次是 Transformer 架构的攻克。 在风格迁移课题中，StyTR-2 模型完全摒弃了 CNN，转而采用纯 Transformer 架构。这与我们熟悉的卷积范式截然不同。我从自然语言处理领域的原始论文《Attention Is All You Need》切入，逐步拆解了 Transformer 的核心组件：多头自注意力机制允许模型在处理当前图像块时，动态计算其与全局所有图像块的关联权重，从而打破了 CNN 局部感受野的限制，能够捕捉长距离依赖关系。我也理解了其计算复杂度为，这很好地解释了论文中提及的为何 StyTR-2 在处理高分辨率图像时会面临效率挑战。

再次是语义分割与轻量化网络的学习。针对街景分割课题，我梳理了语义分割“为每个像素分配类别标签”的核心定义。为了理解 DeeplabV3 的优越性，我重点攻克了两个概念：一是空洞卷积，它通过在卷积核中插入“空洞”来在不增加参数量的前提下指数级扩大感受野；二是空间金字塔池化（ASPP），它通过并行使用不同扩张率的空洞卷积，让模型能够同时洞察近处行人的细节与远处建筑的轮廓。此外，我还深入学习了 Xception 架构的核心——深度可分离卷积。通过对比计算（如 3×3 卷积核、256 输入通道、512 输出通道下，标准卷积参数量约 118 万，而深度可分离卷积仅约 13 万），我直观地理解了课题二为何能通过替换主干网络实现 30% 的推理速度提升。

核心体会：基础不牢，地动山摇。每一个看似简洁的公式背后，都凝聚着研究者对特定痛点的深刻思考。只有彻底弄懂“是什么”与“为什么”，才能在后续的模型改进中找到正确的方向。

2.2 第二阶段：模型理解与创新探索——从“看热闹”到“看门道”

掌握了基础工具后，我开始深入研读两项课题的模型架构，试图解构每一项设计背后的动机与创新之处。

2.2.1 StyTR-2 模型的设计哲学：全局协调与结构保真

在风格迁移课题中，核心痛点是传统 CNN 因局部感受野限制，导致风格迁移后出现“风格碎片化”与“内容结构扭曲”。针对此，StyTR-2 采用了三大创新设计：

内容感知位置编码（CAPE）：起初我困惑于 Transformer 已有注意力为何还要位置编码。后来我明白，注意力是“位置无关”的，它只看内容相似度，不知道图像块在哪。传统正弦编码虽能提供位置，却与图像语义脱节。CAPE 的精妙之处在于将位置编码与图像语义内容动态关联。具体实现中，它利用公式 来计算相对位置关联强度，并通过学习固定大小矩阵并双线性插值的方式实现尺度不变性。这意味着，即便“天空”区域的像素块在空间上相隔甚远，CAPE 也能赋予它们相似的位置表征，从而在风格注入时保持语义区域的一致性，这正是论文中内容损失降至 1.51 的关键所在。

双路 Transformer 编码器：为何不让内容和风格共享编码器？因为二者分属不同特征域：内容需要保留精确空间结构，风格需要提取抽象色彩纹理。共享编码器会导致特征污染。双路设计从根源上实现了内容与风格的彻底解耦，为后续的交叉注意力融合扫清了障碍。

2.2.2 改进 DeeplabV3 模型的实用主义：精度与效率的平衡

在街景分割课题中，目标是将模型部署于资源受限的物联网设备。原始 DeeplabV3 虽精度较高，但参数量大、推理慢。我们的改进紧紧围绕“降本增效”展开：

Xception 架构替换主干：通过深度可分离卷积，在几乎不损失特征表达能力的前提下，将模型参数量从 42.6M 降至 29.8M。

注意力机制融合：创新性地将通道注意力与空间注意力并行融合。通道注意力通过全局平均池化与全连接层学习特征通道的重要性权重，空间注意力则通过最大 / 平均池化拼接卷积生成空间权重图。二者的结合使得模型能精准聚焦于街景中的关键目标（如行人、交通标志），交并比因此提升了 5.2 个百分点。

多尺度特征融合优化：结合多级处理与深度分解策略，强化了模型在动态光照和遮挡场景下对小目标的识别鲁棒性。

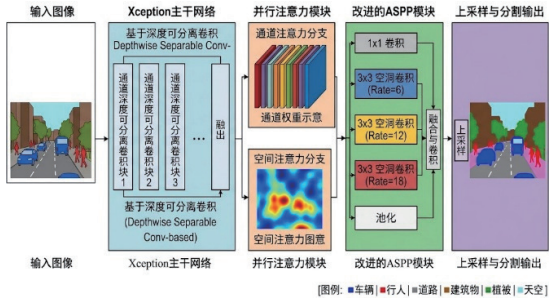

图 1. 改进 DeeplabV3 模型整体架构图

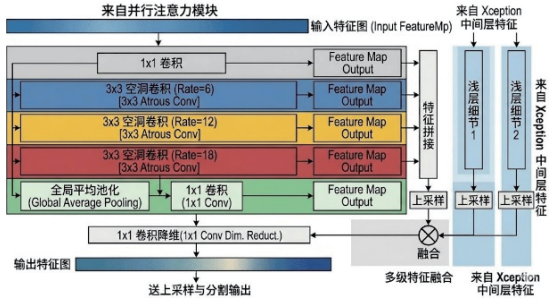

图 2. 改进后的 ASPP 模块结构图

核心体会：科研创新并非总是颠覆性的。StyTR-2 并未发明 Transformer，而是将其巧妙适配到视觉任务；改进 DeeplabV3 也并未抛弃 ASPP，而是注入了注意力机制与轻量化模块。这种“站在巨人肩膀上”的增量式改进，同样具有重大的学术与应用价值。

2.3 第三阶段：实验验证与分析——在试错中逼近真相

理论学习之后，实践是检验真理的唯一标准。在这一阶段，我全程参与了从环境搭建到结果分析的全流程，将抽象的公式转化为具体的代码与图表。

2.3.1 环境搭建与数据准备：细节决定成败

两个课题均基于 PyTorch 框架。我掌握了使用 Anaconda 创建独立虚拟环境的方法，避免了库版本的冲突噩梦。在街景分割课题中，我们使用 Cityscapes 数据集（包含 50 城、5000 张精细标注图像）。我亲手完成了图像尺寸统一（512×1024）、像素归一化，并实施了随机裁剪、翻转、颜色扰动等数据增强策略。这些看似枯燥的操作，实际上是提升模型泛化能力的基石。

2.3.2 训练与调试：直面理论与工程的鸿沟

在训练过程中，我切身体会到了理论与实践的差距。课堂上讲的损失函数，在代码里需要设置具体的权重与优化器参数。我遭遇过训练初期 Loss 剧烈震荡、显存溢出（OOM）、模型过拟合等一系列工程难题。在导师指导下，我学会了使用 TensorBoard 可视化训练曲线，通过观察损失值与评估指标（如 mIoU、LPIPS）的动态变化来诊断模型状态，并据此调整学习率衰减策略与批量大小。

2.3.3 结果分析与论文撰写：用数据讲好科研故事

实验数据是对模型最好的辩护。在 StyTR-2 课题中，我们的风格损失（1.93）与内容损失（1.51）均优于 Gatys、AdaIN 等对比方法，这直观印证了全局注意力机制在风格协调上的巨大优势。在街景分割课题中，改进后模型的 mIoU 达到 92.5%，推理速度提升 30%。特别是消融实验让我印象深刻：移除 CAPE 后内容损失上升 17.9%，替换 Xception 后推理耗时增加 35%。这些冰冷的数字背后，是模块有效性的铁证。在撰写论文时，我学会了如何用准确的语言描述 CAPE 的工作流程，如何绘制清晰网络结构图来展示双路编码器，以及如何通过对比表格让实验结论一目了然。

2.4 第四阶段：总结反思与能力沉淀

两项课题的结题并不意味着学习的结束，而是新认知构建的开始。

知识与能力的跃迁：经过一年的科研训练，

表 1. 不同模型在 Cityscapes 数据集上的性能对比

模型类型	交并比 (%)	像素准确率 (%)	推理时间 (ms)	模型参数量 (M)
ENet	88.0	94.2	65	3.6
U-Net	90.0	95.1	98	31.0
原始 DeeplabV3	87.3	93.6	125	42.6
改进 DeeplabV3	92.5	96.8	87.5	29.8

表 2. 消融实验结果

模型变体	平均交并比 (%)	参数量 (M)	推理时间 (ms)
完整模型	92.5	29.8	87.5
无注意力机制	89.3	29.8	87.5
无 Xception 优化	88.4	42.6	118.1
无多尺度融合	89.7	29.8	87.5

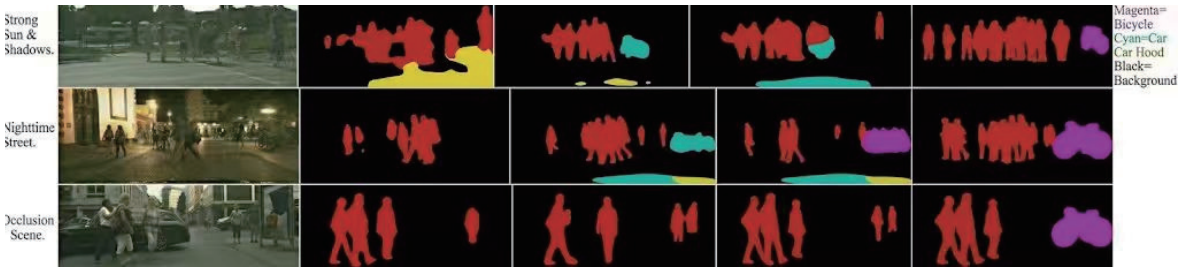

图 3. 复杂场景下分割效果对比图

我不仅掌握了深度学习的前沿知识，更习得了查阅文献、搭建环境、调试复杂模型、撰写技术文档等无法在课堂获得的“硬实力”。我从一个只会运行 Demo 的初学者，成长为能够复现并改进模型的入门研究者。

对“创新”概念的祛魅与重构：我曾天真地认为创新就是发明全新的算法。但在实践中我领悟到，有价值的创新更多是针对特定痛点的深度优化。StyTR-2 的 CAPE 是为了解决 Transformer 在视觉任务中的“语义失明”，改进 DeeplabV3 的注意力融合是为了解决复杂场景下的特征聚焦。这种基于深刻理解问题的“增量式创新”，同样是推动技术进步的核心动力。

正视自身的不足：我也清醒地认识到自身在数学推导严谨性、大规模工程架构设计能力以及学术前沿追踪敏锐度方面的不足。这些将是我未来攻读研究生阶段需要重点攻克的方向。

3 学习心得与成长感悟

3.1 从“记忆公式”到“还原思考过程”

在最初学习 CAPE 公式时，我习惯于死记

硬背数学表达式。然而，当代码实现出现偏差导致 Loss 不收敛时，我才意识到仅记住公式毫无用处。我开始尝试还原作者的思考过程：作者面对的是“Transformer 在视觉任务中无法感知位置关系”的问题，传统的正弦位置编码虽然提供了坐标信息，却将视觉内容与位置割裂开来。作者为了解决这个问题，设计了一个既能反映相对距离、又能与图像语义内容产生关联的编码机制。我顺着这个思路推导公式的物理意义，发现实际上是在通过傅里叶特征来隐式地度量相对位置，并通过学习的权重自适应地调整对不同尺度距离的敏感度。当我理解到这一层，再去看代码中关于矩阵计算和张量广播的操作，瞬间豁然开朗。这让我明白，深度学习的学习，核心在于理解每个数学符号背后的“动机”，而非符号本身。

3.2 敬畏每一行代码与每一个比特

参与项目前，我认为科研就是写论文、画图表。亲身经历告诉我，科研的底座是厚重的工程代码。在调试街景分割模型时，我曾因为一个 `torch.Tensor` 和 `numpy.ndarray` 的类型未对齐，导致归一化出错，模型输出的掩码全黑，而我花了整整一天才定位到这个不起眼的 bug。还有一次，在服务器上跑实验，忘记设置 `torch.no_grad()` 导致验证集计算图累积，显存直接爆掉，拖慢了整个课题组的进度。这些“血的教训”让我建立了严格的代码规范意识：变量命名必须清晰、张量形状必须写注释、实验前必须用 `assert` 检查尺寸。我也因此养成了写实验日志的习惯，记录每一次参数改动、每一次报错与解决方案。这种对细节的敬畏，是我在课堂作业中从未体验过的。

3.3 从“逐字翻译”到“批判性对话”

大二读论文时，我抱着一种“瞻仰权威”的心态，认为顶会论文无懈可击。但在深入研究 StyTR-2 后，我发现其在处理极小目标或纹理过于复杂的风格图像时，仍会出现伪影。我尝试在组会上提出：“既然 CAPE 解决了位置语义感知，为何在面对高频纹理时表现下降？”导师引导我查阅了后续的 StyleGAN 相关文献，我才明白是 Transformer 在解码阶段的 Patch 合并操作导致了高频信息的丢失。这让我学会了与论文作者进行“批判性对话”：不仅要吸收文章的优点，更要敏锐地察觉其局限性。这种批判性思维，是进行下一阶段创新研究的必要前提。

4 未来展望与自我期许

一年的科研训练虽然告一段落，但它为我打开了深度学习研究的大门。站在新的起点，我对未来的学习和研究有了更清晰的规划。

4.1 短期目标：夯实理论根基，拓宽技术视野

在接下来的半年里，我计划系统性地补强数学基础，特别是矩阵论和优化理论，以更好地理解扩散模型（Diffusion Models）和神经辐射场（NeRF）等新兴方向背后的数学原理。同时，我将继续追踪 CVPR、ICCV 等顶会的最新成果，保持对前沿技术的敏感度。

4.2 中期规划：深化特定领域研究，准备研究生过渡

在攻读研究生阶段，我希望能够将本次科研经历中的两个方向——生成与分割——进行有机结合。例如，探索利用生成式模型（如 GAN 或扩散模型）来扩充稀缺的街景分割数据，以解决极端天气、夜间光照不足场景下分割精度下降的问题。我也期望能在导师指导下，独立完成一篇高质量的一作论文，完成从“参与者”到“主导者”的身份转变。

4.3 长期愿景：做有温度的计算机视觉研究

通过图像风格迁移项目，我看到了 AI 在

艺术辅助领域的创造力；通过街景分割项目，我看到了 AI 在智慧城市与弱势群体出行（如视障人士导航）中的社会价值。我希望未来的研究不只是追求 SOTA 的冰冷数字，而是能够真正解决现实世界的痛点，做出既有技术深度又有人文温度的科研工作。

5 结语

从最初对深度学习仅有模糊概念，到如今能够参与完成两项前沿课题研究，这一年的经历是我本科阶段最宝贵的财富。它让我完成了从知识被动接收者向主动探索者的身份蜕变，也让我真切感受到了计算机视觉技术在艺术创作与智慧城市建设中的巨大魅。

图像风格迁移展现了 AI 在美学领域的无限可能，街景语义分割则彰显了技术服务社会的实用价值。两项研究虽方向各异，但其底层逻辑共享深度学习的理论基石。这段经历教会我的不仅是 StyTR-2 或 DeeplabV3 的具体算法，更是一种面对未知领域的系统性学习能力与解决问题的科研素养。

更重要的是，我学会了如何在枯燥的公式推导中寻找逻辑之美，在棘手的代码调试中锤炼耐心，在激烈的组会讨论中磨砺思维。这些隐性的收获，远比学会使用某个框架更为珍贵。

道阻且长，行则将至；行而不辍，未来可期。我期待在计算机视觉领域继续深耕，本次科研经历所沉淀的学习能力与研究思维，将成为我未来学术道路上最坚实的基石。

参考文献

[1] 陈淑環 , 韦玉科 , 徐乐 , 等 . 基于深度学习的图像风格迁移研究综述 [J]. 计算机应用研究 , 2019, 36(08): 2250-2255.

[2] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.

[3] 陈淮源 , 张广驰 , 陈高 , 等 . 基于深度学习的图像风格迁移研究进展 [J]. 计算机工程与应用 , 2021, 57(11): 37-45.

[4] 廉露 , 田启川 , 谭润 , 等 . 基于神经网络的图像风格迁移研究进展 [J]. 计算机工程与应用 , 2024, 60(09): 30-47.

[5] Deng Y, Tang F, Dong W, et al. StyTR-2: Image Style Transfer with Transformers[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[6] Cordts M, Omran M, Ramos S, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 3213-3223.

[7] Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 1251-1258.

[8] Johnson J, Alahi A, Fei-Fei L. Perceptual Losses for Real-time Style Transfer and Super-resolution[C]//European Conference on Computer Vision (ECCV), 2016.

[9] Park D Y, Lee K H. Arbitrary Style Transfer with Style-Attentional Networks[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[10] 张文海 . 街景小目标识别算法研究 [D]. 合肥 : 中国科学技术大学 , 2019.

[11] 董康龙 . 基于多尺度特征融合的街景图像分割算法研究 [J]. 计算机应用 , 2020, 40(5): 1345-1350.