第5章

幅-相感知与高分辨率语义生成

考虑到高分辨率语义生成阶段高级语义与精细定位之间的内在矛盾，对比现有场景语义解析模型倾向于二者之间的极限权衡或过度补偿，本章提出了频域下语义-定位的解耦表征，从本质上突破像素级学习的局限性问题。具体而言，本章首先研究并揭示了图像幅度和相位在语义和定位方面的反向对称固有特性。基于该特性，针对高低层特征之间的语义鸿沟问题，本章提出了一种高效的幅度感知模块（magnitude perceptron，MP），通过动态权重机制来自适应地捕获特定频率组合，从而在增强语义表征的同时，避免语义鸿沟。针对不同层级特征之间的定位结构不对齐问题，本章提出了一种简洁的相位修正模块（phase amender，PA），通过简单高效的谱残差映射，显式挖掘并修正每层特征的定位偏移和错位信息，从而促进细粒度的原型定位表征。此外，针对标准分割损失的不平衡优化问题，本章提出了一种新颖的相位敏感性损失（phase-sensitive loss，PSL）作为辅助约束，以促使网络各阶段基于相位来自适应学习无关于语义的原型定位，从而在保证多层级语义多样性的同时，增强了模型的细粒度分辨率重建能力。本章通过多种实验与可视化手段，论证了本章所构建的细粒度高分辨率语义生成思想与方法的优越性，解决了空间像素级学习方法的语义依赖和细节失真等局限性问题。

5.1 概述

场景语义解析的核心目标，在于如何生成同时具有强大语义和精确定位的高分辨率语义结果。然而，这两者之间的内在矛盾促使大多数现有方法在分辨率重建过程中对高级语义和精细定位之间倾向于进行极限权衡或过度补偿，这可能导致性能有限或计算成本巨大。为此，本章致力于从频率空间解耦语义和定位表征

来提高网络的分辨率细节重建能力和泛化能力。具体而言，本章针对5.1.1节提出的科学问题，给出了如5.1.2节所述的研究内容及解决方法。

5.1.1 拟解决的主要问题

本章拟解决的场景语义解析网络在高分辨率语义生成方面的主要问题如下。

（1）**高低层特征之间语义粒度差异产生的语义鸿沟容易导致上下文建模过程中的信息淹没与语义混淆。**首先，不同层次特征之间的语义鸿沟是多层级上下文建模过程中不可避免的挑战。现有方法通常利用像素级注意力（pixel-wise attention）[60,137]或门控机制（gating mechanisms）[62,140]，通过为多层特征分配不同权重来构建上下文关系。然而，这些方法往往由于低层特征的引入而产生语义混淆，且低级特征由于语义较弱容易受光照变化及图像噪声影响，使得有用信息淹没在大量无效信息中，从而损害模型泛化能力。针对该问题，本章强调，每一层级的语义特征在同一特定位置下的贡献是不一致的。因此，得益于频域下图像幅度谱的语义特异性，本章设计了一个具有动态加权机制的幅度感知器，通过自适应调制频谱中每个频率分量的强度，在增强语义表征的同时，避免语义鸿沟问题。

（2）**不同层级或尺度的特征之间的空间偏移或错位，致使特征融合阶段空间定位结构难以对齐。**多尺度特征融合过程中的定位结构对齐问题对精确分割结果至关重要。它主要源于上采样或噪声干扰等原因导致的深浅层特征之间的空间位置偏移或错位。现有方法[11,28,39]往往通过光流或变形网络来对齐高低层特征或单纯拉取浅层特征来增强定位信息等。然而，由于没有直接解决深浅层特征在空间位置上的偏移根源，这些方法难以确保特征在融合时定位结构的一致性和准确性。得益于相位具有语义无关但定位特异性，本章提出利用相位来显式挖掘并修正每层特征的定位信息，促进细粒度重建的原型定位表征。

（3）**标准的语义分割优化目标由于强制让所有网络层致力于拟合最终的语义信息，致使层级语义特征趋于同质化。**辅助监督作为一种常用的训练策略，致力于提高分割性能。然而，值得注意的是，在中间层仅依赖语义标签作为优化目标可能会导致优化不平衡和性能受限，因为它迫使每一层的输出都去拟合相同的语义标签。这一过程从本质上破坏了多层特征之间的语义多样性和互补性，使其趋于同质化。这对解码器中以细粒度分辨率重建为目标的网络层尤其不公平。尽管有些方法[10,62] 建议使用边界监督来提高性能，但它们仍然无法摆脱语义本身带

来的优化约束。得益于解耦的纯相位描述的每层特征的定位信息是类无关且客观存在的，为此，本章特别定义了一个相位敏感性损失作为辅助约束，以促使网络各阶段基于相位来自适应学习无关于语义的原型定位，保证多层级语义多样性和鲁棒性。

5.1.2　研究内容及贡献

针对5.1.1节所提出的科学问题，本章研究内容及贡献如下。

（1）本章探索了频域下图像谱特征的反向对称固有特性，即幅度谱的语义特异性但定位无关性及相位谱的语义无关性但定位特异性。

（2）本章提出了基于动态权重机制的幅度感知模块MP，通过动态学习各频率分量强度，来自适应地捕获特定频率组合，从而在增强语义表征的同时，避免语义鸿沟。

（3）本章提出了一种简洁的相位修正模块PA，通过自适应学习相位偏移，从而使模型保持对定位偏移的敏感度，进而提升了细粒度的原型定位表征。

（4）本章定制了一个相位敏感损失PSL，来促进网络自适应地学习不同层级语义无关的原型定位特征，从而优化了定位信息并增强了分辨率重建细节。

5.2　图像频域表征分析

5.2.1　图像频域表征形式

作为频率的复值函数，（快速）傅里叶变换[134]（FFT）以相同的长度将有限信号从时域变换到频域。换言之，FFT可用于实现原始输入的频率表征变换。图像作为一种只有实数的二维离散有限信号，因此它的频率表征可以通过式(5.1)所示的2D FFT获得。

$$F(m,n)=\sum_{x=0}^{H-1}\sum_{y=0}^{W-1}f(x,y)\cdot \mathrm{e}^{-\mathrm{j}2\pi\left(\frac{mx}{H}+\frac{ny}{W}\right)} \tag{5.1}$$

其中，$f(x,y)$表示大小为$H\times W$的图像在空间坐标(x,y)下像素值；$F(m,n)$是空间频率坐标(m,n)处所对应的频率值。

在给定$F(m,n)$的情况下，可以通过式(5.2)所示的反FFT（IFFT）来恢复原

始图像信号 $f(x,y)$。

$$f(x,y)=\frac{1}{HW}\sum_{m=0}^{H-1}\sum_{n=0}^{W-1}F(m,n)\cdot \mathrm{e}^{\mathrm{j}2\pi\left(\frac{mx}{H}+\frac{ny}{W}\right)} \tag{5.2}$$

根据欧拉公式，式中的 $F(m,n)$ 复数值可以用以下三种形式进行表示。

（1）**实部-虚部表征：** 设 $R(m,n)$ 和 $I(m,n)$ 分别表示 $F(m,n)$ 的实部和虚部，简记 $R(m,n)=A$ 和 $I(m,n)=B$。式(5.1)可被变换为

$$F(m,n)=R(m,n)+\mathrm{j}I(m,n)=A+\mathrm{j}B \tag{5.3}$$

该表达形式比较直观，却由于其物理意义不明确而不利于谱分析和处理。

（2）**幅值-相位表征：** 数值上，利用 $F(m,n)$ 的模和辐角来分别表示其幅度和相位，即

$$\begin{aligned}|F(m,n)|&=\sqrt{R(m,n)^2+I(m,n)^2}=\sqrt{A^2+B^2}\\ \angle F(m,n)&=\arctan\left(\frac{I(m,n)}{R(m,n)}\right)=\arctan\frac{B}{A}\end{aligned} \tag{5.4}$$

幅值代表强度，其数值大小反映了某一空间频率对图像的贡献有多大。它反映了图像特征的几何结构，即空间域的变化。相位反映了每个频率的一个完整正弦波周期的移动，它决定了图像特征的位置。

（3）**极坐标表征：** 该方式利用复平面上的极坐标来表示 $F(m,n)$，其中极半径为 $r_F(m,n)$，极角为 $\theta_F(m,n)$。因此，可以将 $F(m,n)$ 重写为

$$F(m,n)=r_F(m,n)\cdot \mathrm{e}^{\mathrm{j}\theta_F(m,n)} \tag{5.5}$$

其中，

$$\begin{aligned}r_F(m,n)&=|F(m,n)|=\sqrt{A^2+B^2}\\ \theta_F(m,n)&=\angle F(m,n)+2k\pi=\arctan\frac{B}{A}+2k\pi,\quad k\in\mathbb{Z}\end{aligned} \tag{5.6}$$

由于具有周期性 $2k\pi$ 的极角在数值计算后的 F 与利用相位计算的一样，故本书在进行变换时不计入 $2k\pi$。

本章方法主要是通过 PyTorch 中的傅里叶变换函数计算图像或特征图的频谱，得到其实部-虚部表征，而后计算其幅值-相位表征，用于进一步表征学习和处理；最后，通过极坐标表征将学习到的幅度和相位频谱组合成新的频率输出，并利用 PyTorch 进行逆傅里叶变换生成空间特征图。

5.2.2　图像谱特性分析

本章探索了一种重要假设：场景语义解析任务中目标/区域的语义信息与定位信息应是可解耦的。即某一物体/区域是什么，与它在图像中的位置无关。这同样适用于交互性语义，例如人和自行车可以被划分为骑手这一语义类别，而将其作为一个整体时，则该“骑手”的语义信息并不取决于其在图像中的空间位置；同样，图像中的定位信息也并不一定涉及其语义信息。尽管有些形状/纹理本身就是语义的一种，但此处更强调其在图中的空间位置/旋转/缩放等。

受自然图像频率模型的启发，图像的幅度谱反映了图像中每个频率分量的强度，而其相位则代表了对应分量的位置，故通过图5.1的图像谱特性分析验证相关假设。将原始“人”的幅度与频域中随机平移图像的相位相结合，然后逆变换到空间域的结果。很明显，结果与随机平移图像相同，这表明改变相位可以在幅值不变的情况下改变物体位置。即相位决定了物体定位，反映了其语义无关但位置特定的特性。右图表示将“道路”的幅值替换为“天空”的幅值后，“道路”区域变成了类似天空的外观，或者说具有天空语义。这意味着保持原始相位不变时，幅度能够控制图像的语义，体现了其语义特定但位置无关的特性。上述观察结果表明，幅值和相位的独立表征有助于实现“语义”和“定位”之间的解耦。

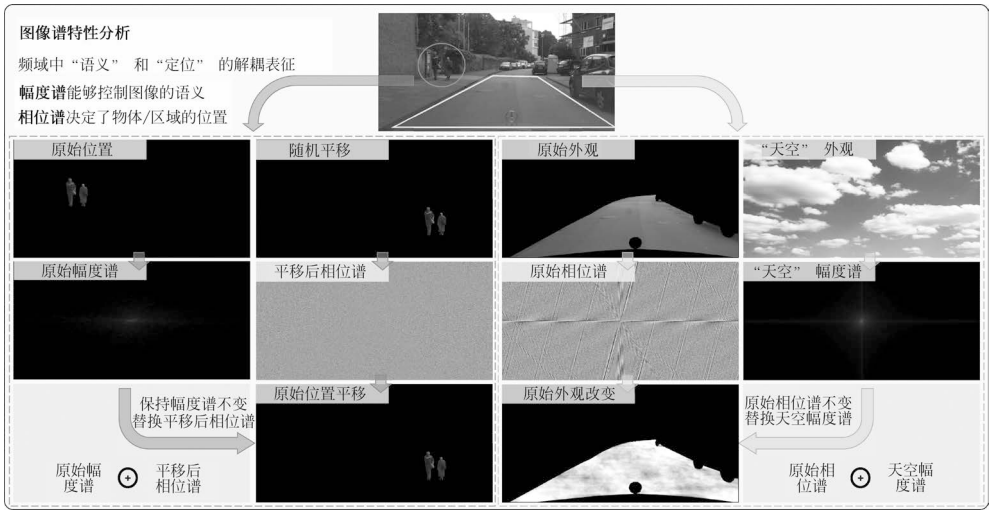

图 5.1　图像谱特性分析示意图

鉴于此，启发于之前的工作[43,135,141-142]，图像的幅值和相位在语义和定位信

息之间具有反向对称特性，即幅度谱具有语义特异性但定位无关性，而相位谱具有语义无关性但定位特异性。

5.2.3 语义-定位解耦表征变换

图5.2描绘了语义-定位解耦表征变换的具体过程，称为自适应频率感知模块（adaptive frequency-aware module，AFM）。该过程主要包含3个步骤：首先对多层级空间特征进行FFT，并对高/低分辨率特征进行尺度对齐预处理，然后对所有特征图进行幅-相感知学习以获取频域下语义-定位解耦表征结果，最后将学习后的结果组合后进行IFFT至图像域进行表征，并使用一个点卷积控制通道数量。其中，幅-相感知过程作为本章关键内容将在后续章节进行详述。

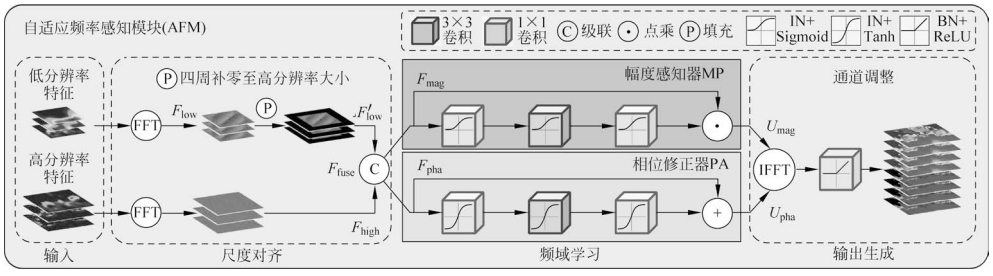

图 5.2 自适应频率感知模块结构示意图

1. 尺度对齐预处理

解码器的设计必然涉及上采样过程，而通用上采样方法无非三种：插值、反卷积或像素混洗。插值计算速度虽快，但精度较差，尤其对物体或区域边界影响较大。另外两种方法计算量过大，效率不高。为解决这一问题，本节探索了一种高效的谱上采样方法对解耦表征前的谱图进行预处理以实现频域内的尺度对齐。

设F_{low}为低分辨率特征输入的FFT结果，F_{high}为高分辨率特征的对应结果。然后，将F_{low}用零填充到F_{high}的大小，就可以得到F'_{low}：

$$F'_{\text{low}} = \text{Pad}(F_{\text{low}}, F_{\text{high}}.\text{shape}) \tag{5.7}$$

高低尺度特征进行尺度对齐后，将两组特征图的频谱图级联为F_{fuse}，即

$$F_{\text{fuse}} = \begin{cases} F_{\text{low}}, & F_{\text{high}} \text{ 为 } \varnothing \\ \text{Concat}\left(F'_{\text{low}}, \quad F_{\text{high}}\right), & \text{其他} \end{cases} \tag{5.8}$$

最后，尺度对齐后的频谱 F_{fuse} 通过式(5.9)转换成幅-相表征形式，以便进一步处理：

$$\begin{cases} F_{\text{mag}} = |F_{\text{fuse}}| \\ F_{\text{pha}} = \angle(F_{\text{fuse}}) \end{cases} \tag{5.9}$$

2. 输出变换

利用式(5.10)，将学习到的 U_{mag} 和 U_{pha} 转换为实部-虚部表征形式为

$$O_{\text{f}} = U_{\text{mag}} \cdot \text{e}^{(1\text{j} \times U_{\text{pha}})} \tag{5.10}$$

然后将结果 O_{f} 进行逆傅里叶变换到图像域，并使用空间点卷积 $W_s^{1\times1}$ 将通道数调整为所需数量。将空间域输出定义为 $O_{\text{s}} \in \mathbb{R}^{[C,H,W]}$，可得

$$O_{\text{s}} = W_s^{1\times1} \otimes (\text{IFFT}(O_{\text{f}})) \tag{5.11}$$

5.3　基于幅度感知的语义多样性表征

5.3.1　设计原理

鉴于语义鸿沟问题的存在，本节认为不同层级特征在融合后特征图的每个位置上的重要性并不相同。因此，鉴于空间上下文关联的限制，纯空间图像中的像素组合很难独立判别准确的语义分布。考虑到频域图像的幅度谱具有语义特异性，但与定位无关，本节提出使用频率级特征提取来学习解耦的幅度谱的分布，有助于学习空间图像的语义分布，进而辅助缓解语义鸿沟问题。

为验证该假设，本节考虑一个简单的相邻层级特征图融合问题来说明幅度感知器的设计原理，如图5.3所示。其中，第一行特征图分别来自 ResNet-18[1]网络的 1/32 和 1/16 尺度，提取它们的幅度谱，分别使用两个手工设计的掩膜与之相乘，最后再变换回图像域。最右侧上下两张图分别展示了幅度谱调整前后的融合效果。由图5.3可知，来自 ResNet-18 骨干网络最后一层的 1/32 尺度特征突出了“车辆”类的语义，而 1/16 尺度输出则主要提取了边界。直接融合这些特征得到的（c）列上方结果表现出明显的语义混淆。而经过掩码处理的幅度谱与原始相位谱结合后被转换回空间域，其输出结果如图第四行所示。结果显示，由于部分频率被移除，转换后的图谱保留了相对较强的激活特征，即掩码后的 1/32 尺度输出减少了一些高频干扰，而 1/16 尺度输出则保留了部分边缘。至此，假设成立。总结而言，直

接融合（在本例中使用随机3×3卷积）原始特征图可能会造成语义混淆，而融合幅度掩码图则可以获得更为准确的语义和精确的边界。

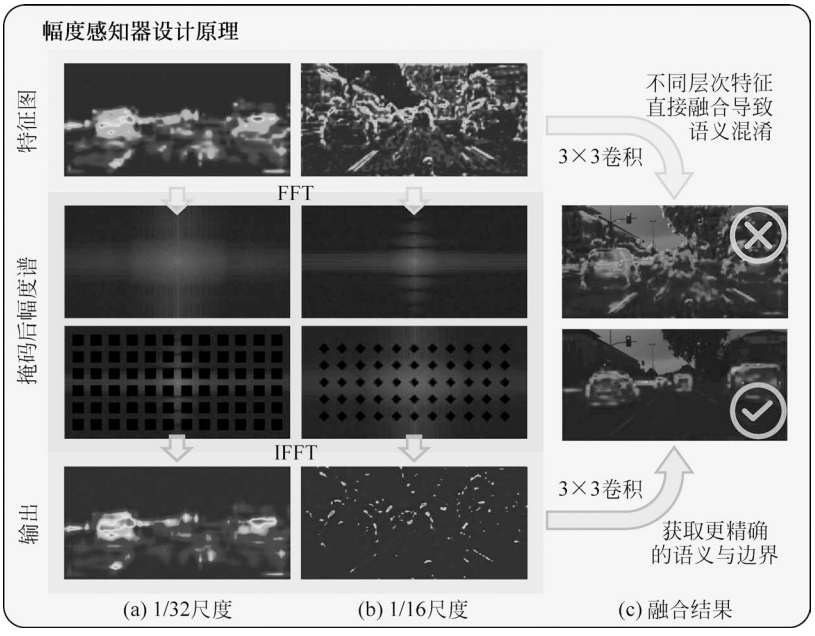

图 5.3　幅度感知器的设计原理示意图

上述讨论揭示，在幅度图上学习频谱特性，能够自适应地评估频谱中各频率成分的重要性，这不仅对于语义控制具有可行性，且有利于特征的有效融合和噪声的有效屏蔽。

5.3.2　网络结构设计

幅度感知器的目标是确保网络能够提高对各频率分量的敏感度，以便在后续变换中对重要的频率分量加以利用和分析，并对价值较低的部分进行抑制。为此，本节提出通过对幅度权重进行显式建模以实现这一目标。图5.2的蓝色方框部分给出了幅度感知器的结构。具体地，MP是一个具有动态加权机制$\mathcal{T}_{\text{mag}}$的幅值计算单元。假设w_{mag}表示学到的权重，则MP的输出U_{mag}可表示为

$$U_{\text{mag}} = \mathcal{T}(F_{\text{mag}}, W_{\text{mag}}) \cdot F_{\text{mag}} = w_{\text{mag}} \cdot F_{\text{mag}} \tag{5.12}$$

由于大多数自然图像的幅度谱在统计上呈$1/f^2$幂律分布[143]，故幅度谱经过

对数变换后会呈现巨大的尺度差异。为解决该问题并提高网络学习效率，本节利用Sigmoid函数对幅度谱进行归一化到[0,1]范围，即

$$\hat{F}_{\mathrm{mag}} = \sigma(\ln(F_{\mathrm{mag}})) \tag{5.13}$$

其中，σ指的是Sigmoid函数。然后，考虑到模型的复杂性，MP利用$1\times1-3\times3-1\times1$的卷积层（其中通道数变化为$C-C/4-C$）形成一个瓶颈来制定动态权重机制。特别是，每个卷积层之后都有一个实例归一化（IN）层和Sigmoid激活层。由此可得下式

$$\begin{cases} w_{\mathrm{mag}} = \mathcal{T}_{\mathrm{mag}}(F_{\mathrm{mag}}, W_{\mathrm{mag}}) = \Pi_{i=1}^{3} W_{\mathrm{mag}}^{i} \otimes \hat{F}_{\mathrm{mag}} \\ W_{\mathrm{mag}}^{i} = \sigma(\mathrm{IN}(W_{\mathrm{mag}}^{k\times k}(\cdot))), \quad k = 1 \text{ 或 } 3 \end{cases} \tag{5.14}$$

其中，$\otimes$表示卷积运算符，Π指连续的卷积运算，$W_{\mathrm{mag}}^{k\times k}$是一个大小为$k\times k$的卷积核。根据式(5.12)～式(5.14)，可以得出MP最终输出的完整形式，即

$$\begin{aligned} U_{\mathrm{mag}} &= \mathcal{T}(F_{\mathrm{mag}}, W_{\mathrm{mag}}) \cdot F_{\mathrm{mag}} \\ &= (\Pi_{i=1}^{3} W_{\mathrm{mag}}^{i} \otimes [\sigma(\ln F_{\mathrm{mag}})]) \cdot F_{\mathrm{mag}} \end{aligned} \tag{5.15}$$

最终输出U_{mag}可以解释为特定频率成分的集合，其统计量对整个语义表征具有重要价值。

5.4　基于相位修正的定位原型优化

5.4.1　设计原理

场景语义解析任务中，生成高级语义时，往往存在空间定位结构不对齐问题。空间定位结构不对齐问题是指在不同层级的特征图之间，存在着空间尺度和定位位置的不匹配，导致生成的高级语义缺乏语义一致性和定位准确性。例如，在场景语义解析任务中，由于下采样和上采样的操作，高分辨率的特征图和低分辨率的特征图之间的对应关系会发生偏移，从而影响目标的边界和细节的恢复。考虑到图像的相位频谱保留了每个频率成分的位置信息，但没有保留其语义标签的信息，因此本节提出：能否通过简单地调节低分辨率特征图的相位来修正上采样特征图的定位？

为此，本节进行了一个简单实验来说明相位修正器的设计原理，如图5.4所示。从ResNet-18[1]的最后一层获取一张特征图 $\mathrm{Feat}_{\mathrm{a}}$，并简单地用FFT计算其相位谱

Phase_a。我们定义一个随机相位偏移 $\Delta P \in [-\pi, \pi]$。令 $\text{Phase}_\text{z} = \text{Phase}_\text{a} + \Delta P$。然后，进行IFFT，生成修正后的特征图 Feat_z。可以看出，原始的特征图 Feat_a 是随机相移的，生成的图更加模糊，甚至部分偏离定位。相反，若已知 Feat_z，令 $\text{Phase}_\text{a} = \text{Phase}_\text{z} - \Delta P$，那么就能生成相对精确的定位特征图 Feat_a。

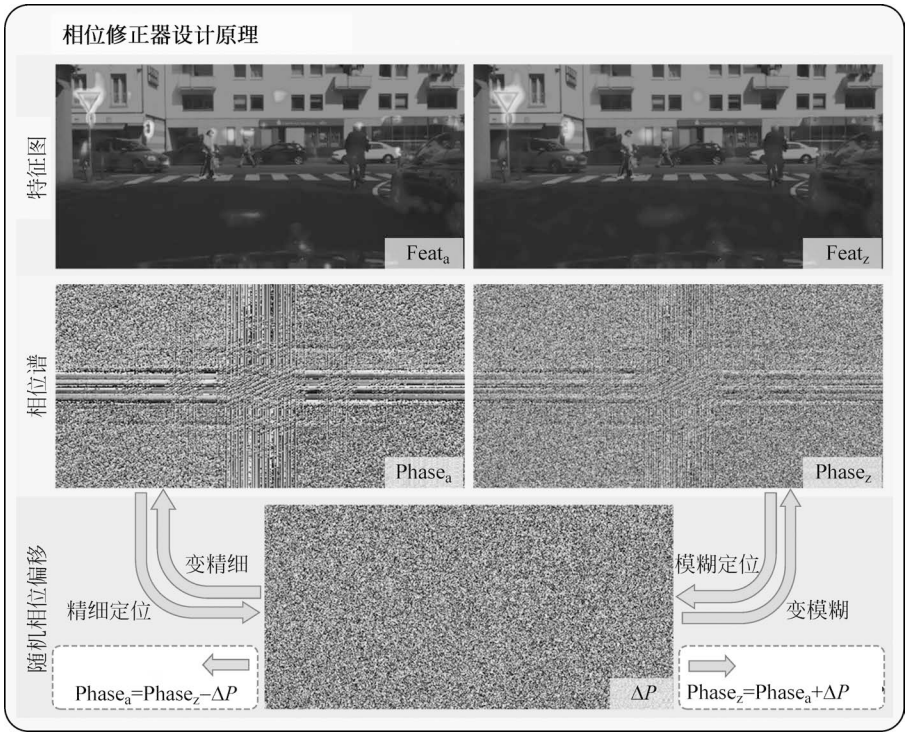

图 5.4 相位修正器的设计原理示意图

基于上述观察结果，可以认为，对不同尺度的特征相位谱进行自适应相位偏移（ΔP）学习能够产生更好的语义定位并促进空间位置对齐。

5.4.2 网络结构设计

基于以上分析，本节提出一种相位修正器（PA）来进行语义定位的优化与对齐，其目标是使模型对偏离规范的定位特征保持敏感性。为此，本节提出一种谱残差映射 $\mathcal{T}_\text{pha}(F_\text{pha}, W_\text{pha})$ 来学习定位的估计误差。定义PA的输入输出如下：

$$U_\text{pha} = F_\text{pha} + \mathcal{T}_\text{pha}(F_\text{pha}, W_\text{pha}) \tag{5.16}$$

与MP类似，同样设计PA模块为一个通道数变化为$C-C/4-C$的$1\times1-3\times3-1\times1$卷积瓶颈结构作为残差相位映射$\mathcal{T}_{\text{pha}}(F_{\text{pha}},W_{\text{pha}})$。注意，与MP不同的是，PA结构中，每个卷积层之后都有一个实例归一化（IN）层和Tanh激活层。

$$\begin{cases}\mathcal{T}_{\text{pha}}(F_{\text{pha}},W_{\text{pha}})=\Pi_{i=1}^{3}W_{\text{pha}}^{i}\otimes F_{\text{pha}}\\ W_{\text{pha}}^{i}=\delta(\text{IN}(W_{\text{pha}}^{k\times k}(\cdot))),\quad k=1\text{ 或 }3\end{cases}\tag{5.17}$$

其中，δ表示Tanh激活函数。

至此，PA的输出可表示为

$$U_{\text{pha}}=F_{\text{pha}}+\Pi_{i=1}^{3}W_{\text{pha}}^{i}\otimes F_{\text{pha}}\tag{5.18}$$

最终输出U_{pha}可以解释为特定频率的校正定位结果，这对图像的显式原型定位表征非常有价值。

5.5　相位敏感性约束

5.5.1　设计原理

本节考虑了网络训练过程中的辅助约束条件，以提高最终性能。本节考虑在网络中间层引入一个新的辅助损失来缓解标准语义分割的不平衡优化问题。该问题主要表现在以下两方面：一方面，网络的不同尺度具有不同的特征信息，强制不同尺度学习所有尺度的信息是不公平的；另一方面，各层级特征融合的目的之一就是取长补短，这意味着需要保证各层级特征的多样性，而对各层网络使用与最终输出相同的损失函数，可能导致各层特征的同质化，从而削弱特征融合效果。为了同时保证各层级特征语义信息的多样性，同时保证最终分割结果的边界准确性，本节提出利用相位谱的语义无关性，在网络中间层引入相位敏感性损失（PSL）。

为了验证这一想法，本节进行了如图5.5所示的实验来说明相位敏感性约束的设计原理。图中第一行左图的分割结果来自UNet-ResNet18[22]，右图是该场景语义分割的真值。在该案例中，主要关注person这一类别的分类结果，如第二行所示。对比可见，UNet-ResNet18的预测结果在分割边界上并不精确。实验发现，如果将真值的相位直接与预测图的幅值拼接，则可以得到如第四行所示的较为精准的分割结果。这意味着，通过相位监督能够有利于直接优化中间层的特征定位，而

不对各特征的语义层级造成影响。

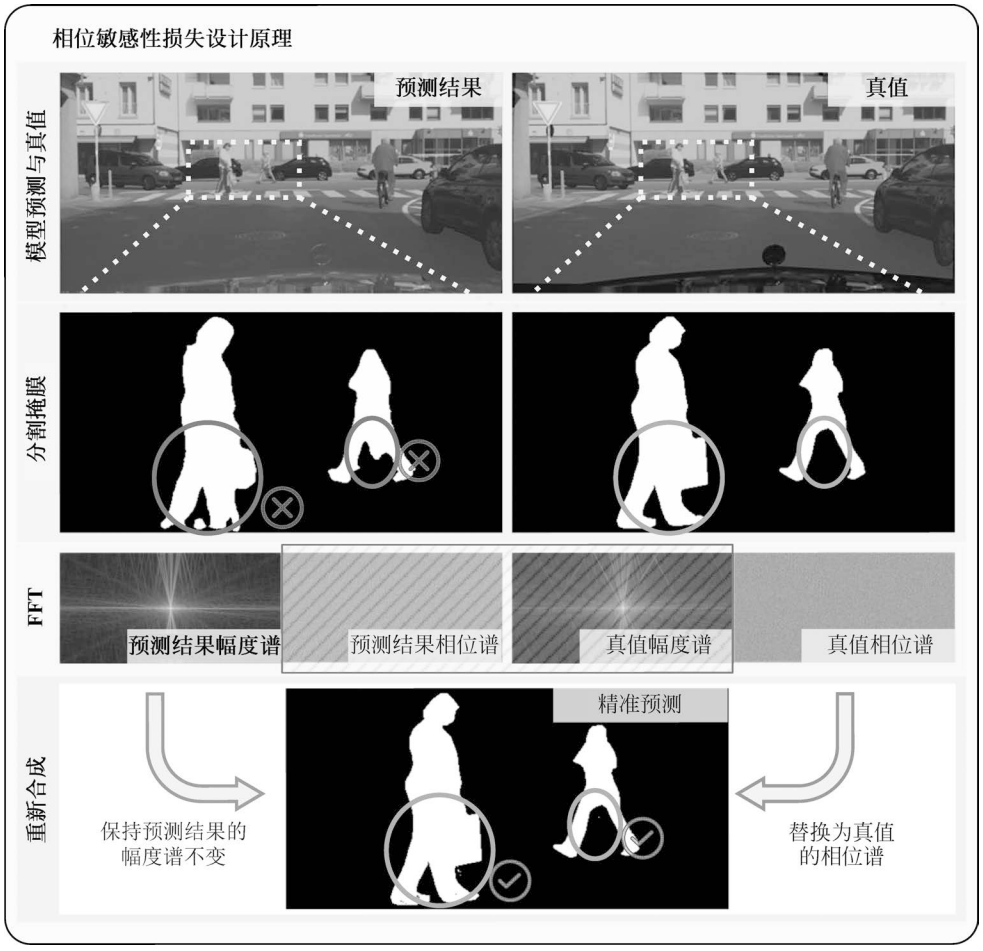

图 5.5 相位敏感性损失的设计原理示意图

基于以上观察，本节在解码器的各个尺度上添加了像素级的分类头，分别计算其预测结果和真值的相位谱之间的距离，并将其作为优化目标的辅助约束，以提高各层特征的定位性能。

5.5.2 设计细节

本节提出的相位敏感性损失（PSL）计算流程如图5.6所示。

算法3给出了该计算过程的伪代码。PSL 的详细计算过程如下。

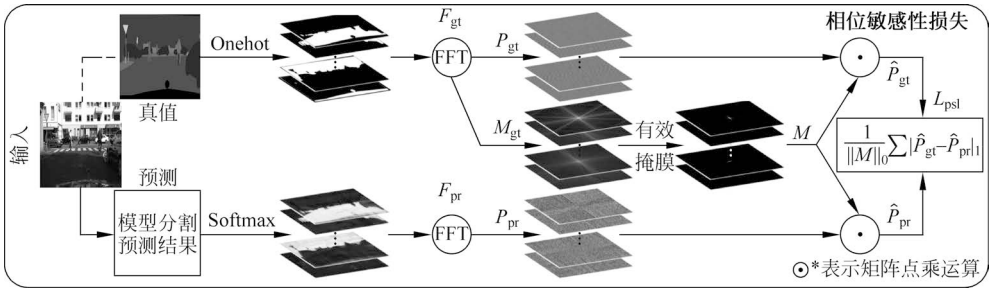

图 5.6　相位敏感损失计算流程

算法 3: PSL 计算过程伪代码

输入: 预测结果 pr$\in \mathbb{R}^{[b,n_c,h,w]}$, 真值 gt$\in \mathbb{R}^{[b,1,h,w]}$，其中 b 为批大小，n_c 为类别数，h 为高度，w 为宽度。

输出: 计算损失 L_{psl}

$\hat{\text{gt}} \leftarrow \text{Onehot}(\text{gt}), \hat{\text{pr}} \leftarrow \text{Softmax}(\text{pr})$;

$F_{\text{gt}} \leftarrow \text{FFT}(\hat{\text{gt}}), F_{\text{pr}} \leftarrow \text{FFT}(\hat{\text{pr}})$;

$P_{\text{gt}} \leftarrow \angle F_{\text{gt}}, M_{\text{gt}} \leftarrow |F_{\text{gt}}|, P_{\text{pr}} \leftarrow \angle F_{\text{pr}}$;

$M \leftarrow \text{zeros}(b, n_c, h, w)$;

while $0 \leqslant i < h$ **do**

  while $0 \leqslant i < h, 0 \leqslant j < w$ **do**

    if $M_{gt}(i,j) \geqslant 1$ **then**

      $M(i,j) = 1$;

    end

  end

end

$\hat{P}_{\text{gt}} \leftarrow P_{\text{gt}} \times M, \hat{P}_{\text{pr}} \leftarrow P_{\text{pr}} \times M$;

$N \leftarrow ||M||_0$;

$L_{\text{psl}} \leftarrow \dfrac{1}{N} \sum |\hat{P}_{\text{gt}} - \hat{P}_{\text{pr}}|_1$;

（1）将语义标签的真值编码为独热（Onehot）向量，从而单独生成每个类别的真值。随后对其进行 FFT，得到 F_{gt}。

（2）对模型的预测输出，执行 Softmax 操作，生成每个类别的概率图。随后对其进行 FFT，得到 F_{pr}。

（3）将 F_{gt} 转化为幅值-相位表示，得到 P_{gt} 和 M_{gt}；由 F_{pr} 得到其相位谱 P_{pr}，即

$$\begin{cases} M_{\mathrm{gt}} = |F_{\mathrm{gt}}|, \quad P_{\mathrm{gt}} = \arctan(F_{\mathrm{gt}}) \\ P_{\mathrm{pr}} = \arctan(F_{\mathrm{pr}}) \end{cases} \tag{5.19}$$

（4）生成频谱掩膜 M 以避免无关频率影响。当某一语义类别在真值中不存在时，会产生全为0的Onehot结果，且该情况尤为常见。根据式(5.3)～式(5.5)，当幅值为0时，任何相位值都可使频谱为0。然而，在幅值为0的频谱上进行零相位监督可能会导致训练不稳定。故有必要将此类频率视为无关频率进行滤除。由式(5.1)，只要Onehot向量中至少有一个像素的值为1，其幅度值就一定大于或等于1，故选取幅度值大于或等于1的频率部分形成频谱掩膜 M。

$$M_{ij} = \begin{cases} 1, & M_{ij}^{\mathrm{gt}} \geqslant 1 \\ 0, & \text{其他} \end{cases} \tag{5.20}$$

（5）将 P_{gt} 和 P_{pr} 分别与 M 相乘，从而提取与分割结果相关的相位信息 $\hat{P}_{\mathrm{gt}}$ 和 $\hat{P}_{\mathrm{pr}}$。

（6）计算 $P_{\mathrm{gt}}^{'}$ 和 $P_{\mathrm{pr}}^{'}$ 上有效相位间的平均曼哈顿距离，作为损失函数 L_{psl}，其数学表达式如下所示：

$$L_{\mathrm{psl}} = \frac{1}{||M||_0} \sum |\hat{P}_{\mathrm{pr}} - \hat{P}_{\mathrm{gt}}|_1 \tag{5.21}$$

其中，$||M||_0$ 是 M 的 L_0 范数，用于计算 M 中非零元素的个数。它表示为频谱中具有有效相位的频率数量。

5.6 实验结果与分析

5.6.1 实验模型构建

基于本章所述模块，本节构建了名为幅相学习分割网络（magnitude-phase learning segmentation network，MPLSeg）的实验模型，其采用典型的编码器-解码器结构，编码器采用通用的骨干网络，网络结构如图5.7所示。给定RGB图像输入 $I \in \mathbb{R}^{[3,H,W]}$，从 ResNet[1] 等骨干网络中提取第 l 层（$l \in [2,5]$）特征图表示为 Enc_l，其表达式为

$$\mathrm{Enc}_l | l \in [2,5] = \mathrm{Net}(I) \tag{5.22}$$

其中，第2、3、4、5级特征分别对应1/4、1/8、1/16、1/32尺度输出。

然后，相邻的低分辨率解码特征和高分辨率编码特征会逐渐共同输入 AFM，从而得到l_{th}级特征图$\text{Dec}_l(l \in [2,5])$，即

$$\begin{cases} \text{Dec}_5 = \text{AFM}\left(\text{Enc}_5\right), & l = 5 \\ \text{Dec}_l = \text{AFM}\left(\text{Enc}_l, \text{Dec}_{l+1}\right), & l \in [2,\ 4] \end{cases} \tag{5.23}$$

最后，将Dec_2输入分割头，并将结果直接进行4倍上采样，以获得最终输出。值得注意的是，除了将标准二值交叉熵作为最终输出（Dec_2）的分割损失外，本章所提出的 PSL 还被用作解码器中间输出（$\text{Dec}_{3\sim5}$）的辅助约束，以提高定位效果。

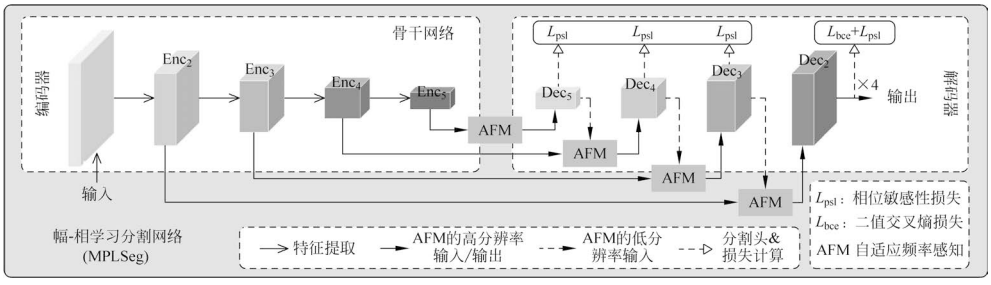

图 5.7　MPLSeg网络结构示意图

整体网络架构采用UNet[22]形式，基于通用的骨干网络模型，形成自下而上的网络架构。为了充分验证本章方法的普适性和有效性，实验在多种代表性骨干网络上进行了验证，其中包括经典的ResNet系列，高效CNN ConvNeXt系列，以及最近兴起的 Swin Transformer系列。在每个系列的骨干网络中，又分别选取了小模型以追求速度-精度的权衡，以及大模型以追求准确率的提升。MPLSeg网络结构参数设置如表5.1所示。

表 5.1　MPLSeg网络结构参数表

模型系列		ResNet[1]		ConvNeXt[144]		Swin Transformer[145]	
骨干网络		ResNet-18	ResNet-101	ConvNeXt-T	ConvNeXt-L	Swin-T	Swin-L
网络名称		MPLSeg-Res18	MPLSeg-Res101	MPLSeg-ConvT	MPLSeg-ConvL	MPLSeg-SwinT	MPLSeg-SwinL
通道数	Enc_2	64	256	96	192	96	192
	Enc_3	128	512	192	384	192	384
	Enc_4	256	1024	384	768	384	768

续表

模型系列		ResNet[1]		ConvNeXt[144]		Swin Transformer[145]	
通道数	Enc_5	512	2048	768	1536	768	1536
	Dec_5	256	1024	384	768	384	768
	Dec_4	256	1024	384	768	384	768
	Dec_3	192	768	288	576	288	576
	Dec_2	128	512	192	384	192	384

5.6.2 消融研究

消融研究主要使用MPLSeg-Res18模型在Cityscapes[107]数据集上进行。为便于描述和对比，消融实验将U-Net作为baseline，即将图5.7中网络架构中的AFM替换为普通3×3卷积后的形式。为公平对比，在对PSL进行消融实验时，分别对比了直接去除PSL（没有辅助损失）和将PSL替换为标准二值交叉熵损失（binary cross entropy，BCE）的结果。MPLSeg各组成部分消融研究结果如表5.2所示。后文将进一步对该结果进行详细分析。

表 5.2 MPLSeg 各组成部分消融研究结果

MP	PA	PSL	BCE	#Params	GFLOPs	mIoU(%)
—	—	—	—	14.59M	168.35G	71.5
✓	—	—	—	12.42M	112.65G	74.3
—	✓	—	—	12.42M	112.65G	75.1
—	—	✓	—	14.59M	168.35G	73.3
—	—	—	✓	14.59M	168.35G	72.6
—	✓	✓	—	12.42M	112.65G	76.8
✓	—	✓	—	12.42M	112.65G	76.2
✓	✓	—	—	13.21M	121.38G	76.8
✓	✓	—	✓	13.21M	121.38G	77.3
✓	✓	✓	—	13.21M	121.38G	78.6

1. 基于幅度感知的语义多样性表征的实验分析

由表5.2可知，MP大幅提升了模型的准确率（74.3% vs. 71.5% (2.9%↑) & 76.8% vs. 75.1% (1.7%↑)），且相比于baseline模型，降低了参数量（14.9%↓）和计算量（33.1%↓）。此外，图5.8提供了MP消融研究的三个典型案例，通过放大被圈出区域来比较激活特征图和最终分割结果，从而揭示MP的内在机制。

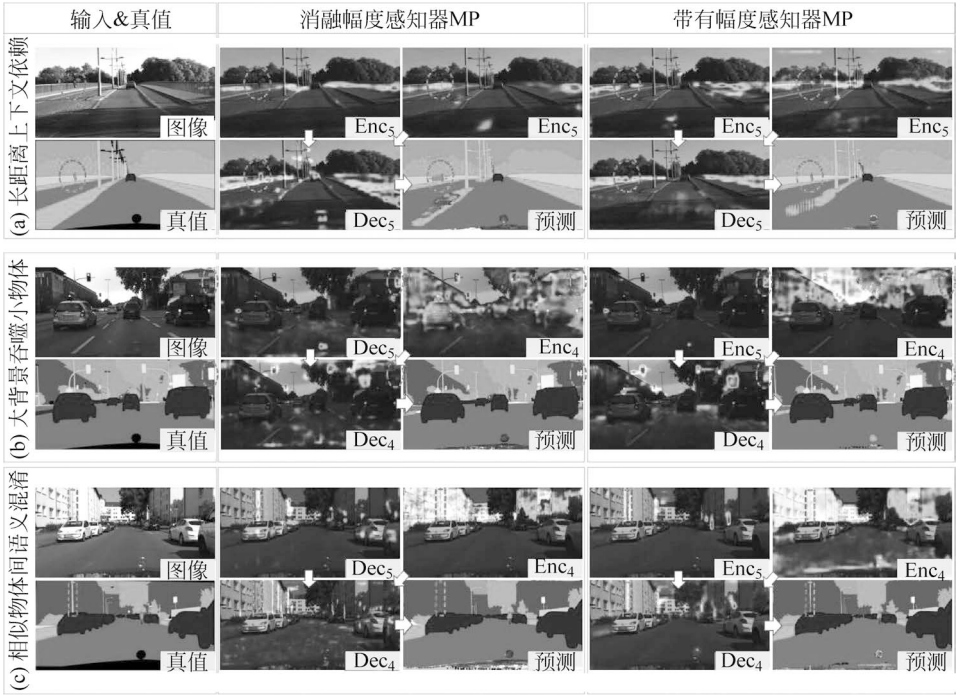

图 5.8　基于幅度感知的语义特征多样性表征方法消融研究的典型案例

示例 (a) 描述了典型纹理重复区域（如栅栏）的长距离上下文依赖关系的构建。不难看出，激活的栅栏特征总是出现在解码器中。由于栅栏的范围较大，但卷积的局部感知能力较强，因此整体栅栏可以通过多个特征图的组合来呈现（为简洁起见，此处仅列出两个特征图）。比较结果表明：不含 MP 的模型呈现出不连续的栅栏分割结果。而由于感受野有限和前景（人物）干扰，该模型未能完全整合与栅栏相关的多个特征。相比之下，由于 MP 的超感受野及其整合相似频率模式的能力，包含 MP 的模型能产生连续的栅栏分割结果。

示例 (b) 展示了小物体被大背景吞噬的情况。图像右上角的交通标志在建筑物背景下很容易被忽略。在特征融合之前，1/32 尺度的 Dec_5 特征包含被激活的大面积墙壁，而 1/16 尺度的 Enc_4 特征则包含被激活的交通标志。然而，在遍历所有 1/16 融合结果后发现，没有 MP 的模型由于墙的语义更强而吞噬了该位置的交通标志，而有 MP 的模型却能激活对应的交通标志。这表明 MP 具有在融合过程中选择特定语义的能力。

示例 (c) 说明了相似背景下的语义混淆问题。原始图像左侧方框内的电线杆与

背景墙的纹理相似，很难将其与背景墙区分开来。与示例（b）类似，融合前的两个特征图分别明确包含了电线杆和背景墙的激活特征。然而，由于语义混淆，不使用MP的模型，其融合特征无法定位框出的电线杆，最终预测结果也是如此。相反，有MP的模型，其融合特征和模型的最终预测都能识别出电线杆。这说明MP能够在频域中构建上下文参照，并避免语义混淆。

简而言之，MP 通过自适应幅度感知学习实现了特定频率感知。MP 有助于模型构建长程上下文依赖关系，减轻语义吞噬和混淆问题，从而提高语义准确性和模型泛化能力。

2. 基于相位修正的定位原型优化的实验分析

由表5.2可知，PA也大大提高了准确度，参数量和计算量与MP相同，与baseline相比分别下降了14.9%和33.1%。此外，图5.9通过三个PA消融研究的典型案例显式说明了PA对定位的修正效果，见图中分割结果和误差图的圈定区域。

示例（a）主要考虑了小物体的精细轮廓。小物体的语义特征通常由深层提取，其精确分割极其困难。从示例（a）中对比的预测结果可以看出，使用PA进行的

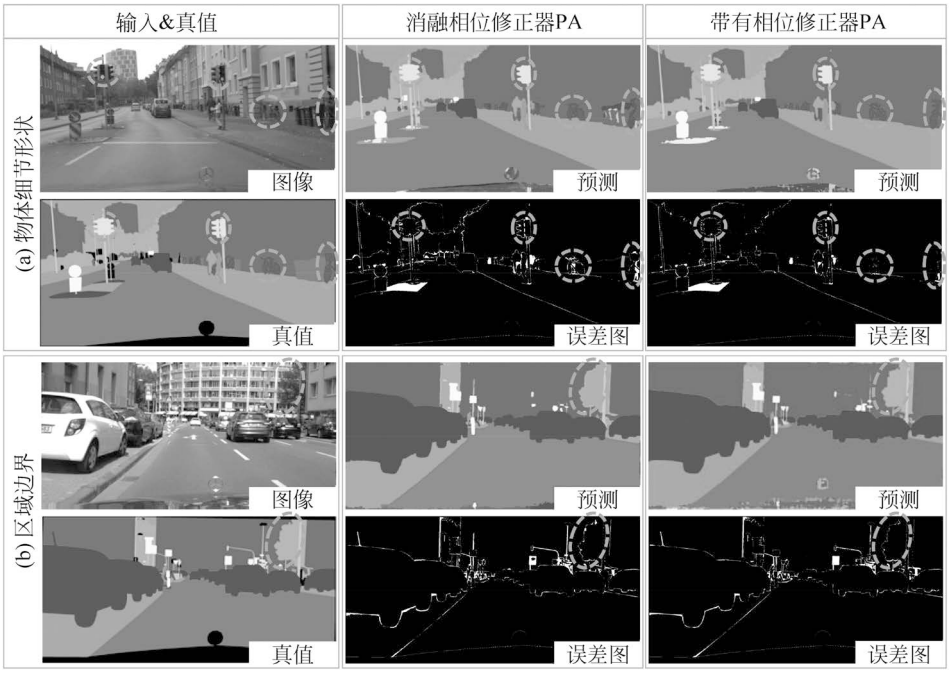

图 5.9　基于相位修正的定位原型优化方法消融研究的典型案例

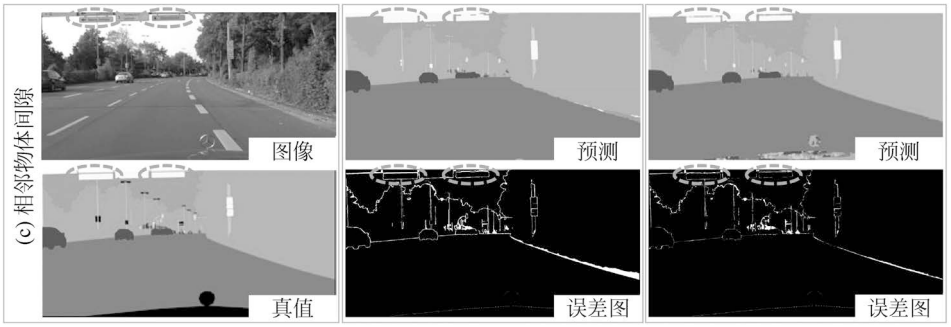

图 5.9　(续)

分割明显更精细，这种效果在误差图中更加明显，误差线更细，甚至没有误差。

示例（a）主要考虑了小物体的精细轮廓。小物体的语义特征通常由深层提取，其精确分割极其困难。从示例（a）中对比的预测结果可以看出，使用PA进行的分割明显更精细，这种效果在误差图中更加明显，误差线更细，甚至没有误差。

此外，示例（b）检查了形状复杂多变的区域，如树木区域。由于复杂区域面积较大，其特征经常出现在多个层级的特征中。而分辨率重建过程中的上采样操作往往会导致边界模糊。这与没有PA的模型表现一致。相反，PA能够驱动模型保持并预测精细的边界。

最后，示例（c）观察了包含不连续区域的物体的定位，如原始图像中被圈出的交通标志，它们具有同一个语义，但中间存在一条割裂的细缝。对比分割结果可以发现，在没有PA的模型中，原始细缝几乎不可见，而在使用PA的模型中，细缝被保留了下来。而使用PA的误差图显示的结果相对较弱。由此可见，PA能够提供更精确的定位，以缓解上采样等引起的语义定位不准问题。

总之，PA对偏离规范定位特征非常敏感，能够改善小物体细节、区域复杂边缘和具有内部缝隙的物体定位，可促进细粒度分辨率重建，提高模型性能。

3．相位敏感性约束的实验分析

在表5.2中，PSL对模型的准确率提升非常显著，尤其是与BCE相比（73.3% vs. 72.6% (0.7%↑) & 78.6% vs. 77.3% (1.3%↑)）。图5.10和图5.11给出了两个PSL消融研究的典型案例，以证明PSL的有效性。

图5.10中的对比结果体现了PSL在保护语义多样性方面的有效性。如图5.10中圈出的交通灯和交通标志，它们在Enc_5特征中都被激活了。然而，随着解码器使用BCE辅助损失来逐渐恢复分辨率，该案例中选择的特征激活程度逐渐降低，直

至在Dec_3阶段完全消失。造成这种情况的原因是，在BCE约束条件下，每一层都需要拟合相同的目标语义，从而导致每一层的语义特征之间出现了权衡，从而牺牲了一些小对象的语义。对比之下，PSL仅对每个对象的定位信息进行优化，而非直接针对语义。因此，在使用PSL的优化过程中，被圈起来的小对象总是会被激活，它们的位置也会不断被优化。

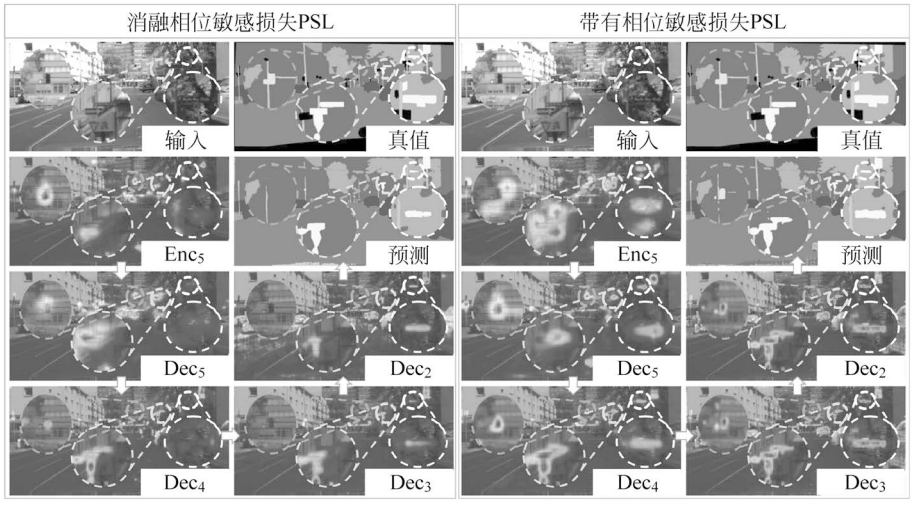

图 5.10　相位敏感性损失函数消融研究中PSL促进语义多样性的典型案例

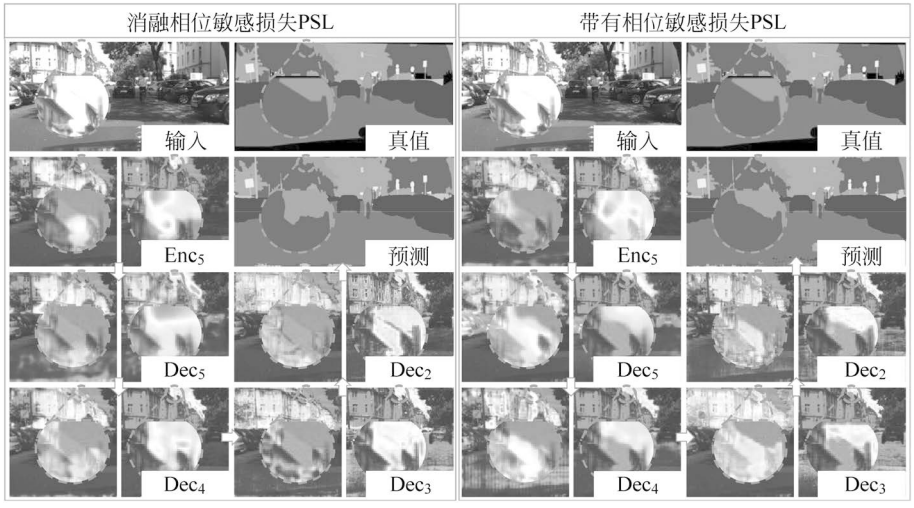

图 5.11　相位敏感性损失函数消融研究中PSL优化语义定位的典型案例

此外，图5.11中的对比结果说明了PSL对语义定位的调节作用。在原始图像中的圈出区域，建筑物的顶部几乎与天空融为一体，难以分辨。可以看出，在编码器特征图中，该区域并没有被激活为建筑物，而是被错误地分类为天空。而未使用PSL的模型输出的激活特征和最终的分割结果中，误分类始终存在。对比之下，使用了PSL的模型，当特征被输入解码器后，建筑物-天空的分割便逐渐趋于正确。该对比结果说明PSL能够通过调节相位来优化物体或区域的定位。

为进一步揭示PSL在保留特征多样性方面的优越性，本实验遍历了Cityscapes[107] 验证集，并计算了从Dec_5到Dec_2的每组特征图中，“人”这一类别的相关矩阵，如图5.12所示。图（a）～（d）分别绘制了$Dec_{5\sim2}$各阶段特征的相关性矩阵，其中，左下三角代表使用了PSL的模型，而右上三角展示了未使用PSL的

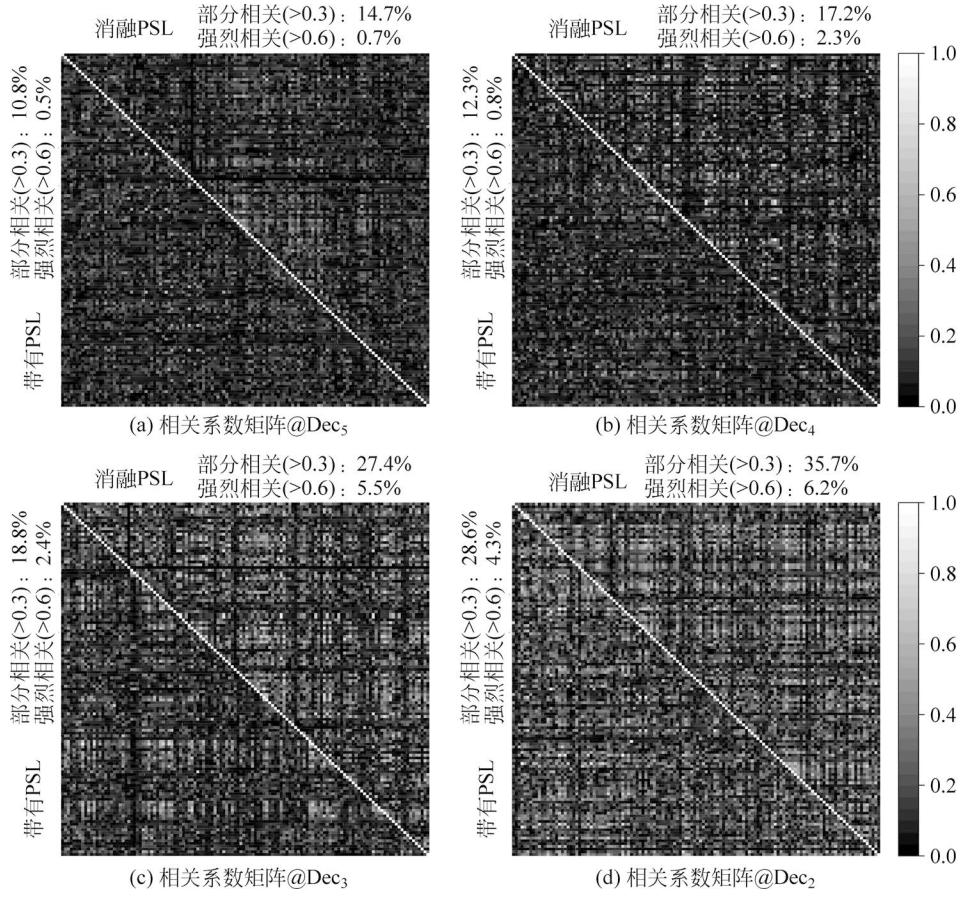

图 5.12　解码器各阶段特征相关性矩阵对比分析示意图

模型（替换为BCE Loss）。颜色越接近白色，相关性越高，越接近黑色，相关性越低。直观地，使用PSL的每个解码器阶段的特征相关性略低于使用标准BCE loss的阶段，即在每个可视化结果中，左下部分比右上部分更接近黑色。

为了定量描述特征冗余度，本实验统计了每个相关矩阵中部分相关（相关度>0.3）和强烈相关（相关度>0.6）的比例，如图5.12和表5.3所示。可以看出，PSL降低了每个阶段的特征相关性，尤其是对于中等尺度阶段。这可能是由于Dec_5的语义最强、分辨率最低，因此受辅助约束的影响相对较小。另外，Dec_2与最终输出相邻，其表示与分割结果高度相关，因此有无PSL在特征相关性上的差异较小。这一结果充分证明了PSL在保持特征多样性方面的优越性。

表 5.3 特征相关性定量分析表

相关程度	Dec_5		Dec_4		Dec_3		Dec_2	
	w/o	w	w/o	w	w/o	w	w/o	w
部分相关(%)	14.7	10.8	17.2	12.3	27.4	18.8	35.7	28.6
强烈相关(%)	0.7	0.5	2.3	0.8	5.5	2.4	6.2	4.3

5.6.3 与当前先进方法的性能对比

1. 在数据集Cityscapes上的性能对比

表5.4展示了MPLSeg与现有先进方法在Cityscapes[107]测试集上的对比结果。为便于比较，本节实验尽量比较了使用相同骨干网络的方法。对于大尺寸模型，为实现公平比较，本章方法如其他方法一样，网络是在训练-验证集上训练的。

表 5.4 MPLSeg与现有先进方法在Cityscapes测试集上的对比

模型大小	方法	骨干网络	mIoU
小模型	BiSeNetV1(ECCV'18)[28]	ResNet-18	74.7
	UperNet(ECCV'18)[146]	ResNet-18	75.4
	ShelfNet(ICCV'19)[116]	ResNet-18	74.8
	MSFNet(BMVC'20)[147]	ResNet-18	77.1
	SwiftNet(CVPR'21)[114]	ResNet-18	76.4
	MSFNet(TIM'21)[125]	ResNet-18	77.1
	MPLSeg(本章方法)	ResNet-18	78.1
	MPLSeg(本章方法)	Swin-T	79.4
	MPLSeg(本章方法)	ConvNeXt-T	**79.5**

续表

模型大小	方　法	骨干网络	mIoU
大模型	PSPNet(CVPR'17)[35]	ResNet-101	78.4
	PSANet(ECCV'18)[148]	ResNet-101	79.7
	CCNet(ICCV'19)[61]	ResNet-101	81.4
	DANet(CVPR'19)[99]	ResNet-101	81.5
	CPNet(CVPR'20)[149]	ResNet-101	81.3
	OCRNet(ECCV'20)[150]	ResNet-101	81.8
	SFNet(ECCV'20)[124]	ResNet-101	81.8
	SPNet(CVPR'20)[151]	ResNet-101	82.0
	GFFNet(AAAI'20)[140]	ResNet-101	82.3
	OCNet(IJCV'21)[152]	ResNet-101	80.1
	ContrastiveSeg(ICCV'21)[153]	ResNet-101	79.2
	MaskFormer(NeurIPS'21)[154]	ResNet-101	80.3
	DeepLabV3+MCIBI(ICCV'21)[155]	ResNet-101	82.0
	DeepLabV3+MCIBI++(TPAMI'22)[156]	ResNet-101	82.2
	MPLSeg(本章方法)	ResNet-101	82.6
	MPLSeg(本章方法)	Swin-L	83.1
	MPLSeg(本章方法)	ConvNeXt-L	**83.3**

在小模型上，本章方法大幅提升了现有方法的准确率，尤其是基于ResNet-18骨干模型取得了78.1%的mIoU，甚至接近了PSPNet[35]在ResNet-101下的结果（MPLSeg-ResNet18-78.1% vs. PSPNet-ResNet101-78.4%）。MPLSeg在高效卷积网络（ConvNeXt-Tiny 79.5%）和Transformer模型（Swin transformer-tiny 79.4%）上的表现也同样令人欣慰。

基于ResNet-101骨干模型时，尽管现有方法已经取得了接近极限的效果，但本章方法依然对各模型的准确率有所提升。注意到TPAMI'22的方法MCIBI++是一个表现极好的基准网络, 在类似的训练和测试条件下，本章方法取得了0.4%的提升及最先进的分割性能。这充分说明了本章方法的优越性。

2. 在数据集ADE20K上的性能对比

表5.5显示了MPLSeg与现有先进方法在ADE20K[109]验证集上的比较结果。当采用 ResNet-18 骨干网络时，MPLSeg 的 mIoU 达到了 40.9%，比之前的 UperNet+ConvNeXt 高出 2.1% mIoU。在以 ResNet-101 为骨干的方法中，本章方法

也表现出了很强的竞争力。值得一提的是，使用 Swin-Transformer 或 ConvNeXt 的 MPLSeg 优于其原始论文中报告的最佳结果（54.0% vs. 52.1% (1.9%↑)，54.5% vs. 53.2% (1.3%↑)）。这对于细粒度语义分割任务提供了新的思路和方法。

表 5.5 MPLSeg 与现有先进方法在 ADE20K 验证集上的对比

模型大小	方法	骨干网络	mIoU
小模型	UperNet(ECCV'18)[146]	ResNet-18	38.8
	Swin Transformer(ICCV'21)[145]	Swin-T	44.5
	MaskFormer(NeuraIPS'21)[154]	Swin-T	46.7
	ConvNeXt(CVPR'22)[144]	ConvNeXt-Tiny	46.1
	MPLSeg(本章方法)	ResNet-18	40.9
	MPLSeg(本章方法)	Swin-T	46.7
	MPLSeg(本章方法)	ConvNeXt-T	**47.3**
大模型	PSPNet(CVPR'17)[35]	ResNet-101	43.3
	PSANet(ECCV'18)[148]	ResNet-101	43.8
	UperNet(ECCV'18)[146]	ResNet-101	42.9
	EncNet(CVPR'18)[157]	ResNet-101	44.7
	CCNet(ICCV'19)[61]	ResNet-101	45.8
	CPNet(CVPR'20)[149]	ResNet-101	46.3
	OCRNet(ECCV'20)[150]	ResNet-101	45.3
	GFFNet(AAAI'20)[140]	ResNet-101	45.3
	OCNet(IJCV'21)[152]	ResNet-101	45.5
	MaskFormer(NeuraIPS'21)[154]	ResNet-101	45.5
	DeepLabV3+MCIBI(ICCV'21)[155]	ResNet-101	47.2
	UperNet+MCIBI++(TPAMI'22)[156]	ResNet-101	47.9
	Swin Transformer(ICCV'21)[145]	Swin-L	52.1
	MaskFormer(NeuraIPS'21)[154]	Swin-L	54.1
	ConvNeXt(CVPR'22)[144]	ConvNeXt-L	53.2
	MPLSeg(本章方法)	ResNet-101	47.9
	MPLSeg(本章方法)	Swin-L	54.0
	MPLSeg(本章方法)	ConvNeXt-L	**54.5**

3. 在数据集 COCO-Stuff 上的性能对比

MPLSeg 与现有先进方法在 COCO-Stuff 164K[108] 验证集上的对比结果见表5.6。无论是在 ResNet 系列、ConvNeXt 系列还是在 Transformer 模型（Swin transformer）

上，本章方法都取得了优异结果，进一步证明了MPLSeg的有效性。

表 5.6 MPLSeg与现有先进方法在COCO-Stuff 164K验证集上的对比

模型大小	方法	骨干网络	mIoU
小模型	BiSeNetV1(ECCV'18)[28]	ResNet-18	28.6
	MPLSeg(本章方法)	ResNet-18	32.2
	MPLSeg(本章方法)	Swin-T	40.0
	MPLSeg(本章方法)	ConvNeXt-T	40.3
大模型	SVCNet(CVPR'19)[158]	ResNet-101	39.6
	DANet(CVPR'19)[99]	ResNet-101	39.7
	OCRNet(ECCV'20)[150]	ResNet-101	39.5
	SpyGR(CVPR'20)[159]	ResNet-101	39.9
	MaskFormer(NeuraIPS'21)[154]	ResNet-101	39.3
	DeepLabV3+MCIBI(ICCV'21)[155]	ResNet-101	41.5
	UperNet+MCIBI++(TPAMI'22)[156]	ResNet-101	41.8
	MPLSeg(本章方法)	ResNet-101	43.6
	MPLSeg(本章方法)	Swin-L	46.5
	MPLSeg(本章方法)	ConvNeXt-L	**46.8**

5.7 本章小结

本章提出的MPLSeg是一种新颖的分割架构，它致力于增强网络的分辨率细节重建能力和泛化能力。通过揭示图像幅值和相位在语义和定位方面的对称反向固有特性，MPLSeg的核心组件AFM利用幅度感知器MP和相位修正器PA促进模型保持对突出频率组合和非规范定位特征的敏感性。此外，辅助相位约束PSL强调了纯相位监督在原定位优化中的有效性。细致的消融研究突出了MPLSeg的核心价值在于，针对定位敏感的视觉任务进行语义定位解耦建模和分析具有普适性和优越性。大量的实验证明了MPLSeg的优越性，其在公共数据集上实现了最先进的性能。本章工作还阐述了以往架构在原型定位表征建模方面的一些局限性。该工作进一步证明了将谱建模方法引入神经架构工程领域的潜力，促进了对视觉模型内在机制的深入研究。