第5章 着色基础 Shading Basics

5.3 实现着色模型 Implementing Shading Models

当然，这些着色和光照方程必须在代码中实现才能发挥作用。在本节中，我们将讨论设计和编写此类实现时的一些关键考虑因素，我们还将介绍一个简单的实现示例。

5.3.1 计算频率 Frequency of Evaluation

（译注：frequency of evaluation应是特殊名词，暂未查到正式翻译，联系上下文后，发现译为“计算频率”更易理解。）

当设计一个着色实现时，需要根据它们的计算频率对计算量进行划分。首先，确定给定计算的结果在整个绘制调用中是否始终是常量，在这种情况下，计算一般可由应用程序（通常是在CPU上）执行，即便GPU计算着色器可以用于特别昂贵的计算，计算结果通过统一着色器输入（uniform shader inputs）传递给图形API。

即使在这个类别中，可能也有一个广泛的计算频率范围，从“一次”开始（starting from "once ever."），最简单的一个例子是着色方程中可能存在的常量子表达式，这可应用于任意很少更改参数（如硬件配置和安装选项）的计算中，这样的着色计算可能在着色器编译的时候就完成了，这种情况下甚至不需要设置一个统一着色器输入，可以在安装阶段或应用加载时，用一个离线的预计算Pass执行计算。

另一种情况是，着色计算的结果在应用程序执行时不断变化，但是变化很慢，以至于不需要在每一帧都进行更新，例如虚拟游戏世界内基于时间的光照参数，如果计算花费昂贵，将其平摊到多个帧是很划算的。

其他情况有逐帧执行计算，例如连接视图矩阵和透视矩阵，或逐模型执行，例如更新依赖于位置的模型光照参数，或逐绘制调用执行，例如更新模型中各材质的参数，通过计算频率，我们将统一着色器输入分组，这样有助于提高应用程序的效率，且也可通过使更新常量最小来提升GPU性能[1165]。

如果着色计算的结果在一次绘制调用中改变，它就不能通过统一着色器输入传递给着色器，相反，它必须通过第3章中描述的一个可编程着色器阶段来计算，需要的话可通过不同的着色器输入传递到其他阶段，理论上，着色计算能在任意可编程阶段上执行，其中每个都对应着不同的计算频率：

* 顶点着色器（Vertex shader）——逐曲面细分前的顶点计算。
* 外壳着色器（Hull shader）**——**逐**表面补丁计算。**
* **域着色器（**Domain shader**）——**逐**曲面细分后的顶点计算。**
* **几何着色器（**Geometry shader**）——**逐**图元计算。**
* **像素着色器（**Pixel shader**）——**逐**像素计算。**

**图片包含 形状

描述已自动生成**

**图 5.9. 公式5.19所示着色模型的逐像素和逐顶点的计算结果的比较，显示了三个不同顶点密度的模型。左侧一列显示逐像素计算的结果，中间一列显示逐顶点计算的结果，右边呈现了每个模型的线框渲染以展示顶点的密度。（来自计算机图形学档案的中国龙网格模型[1172]，原模型来自斯坦福3D扫描存储库）**

**实际上，大部分着色计算是逐像素执行的，尽管这些通常是在像素着色器中实现的，但是计算着色器的实现正变得越来越普遍，一些相关例子会在第20章中讨论，其他阶段主要用于几何操作，例如变换和变形。为了理解为什么是这种情况，我们会对比逐顶点和逐像素着色计算的结果，在旧版的文本里，它们有时会被称作Gouraud着色（Gouraud shading）[578]和Phong着色（Phong Shading）[1414]，尽管这些术语如今已不常使用，对比中使用的着色模型在某些方面与公式5.1中的较为相似，但它经过修改，可以与多个光源一起使用，完整的模型将在稍后详细介绍示例实现时给出。**

**图5.9展示了不同顶点密度的模型的逐像素和逐顶点着色结果，对于龙这个顶点密度极高的网格模型，逐顶点与逐像素之间的区别很小，但在茶壶上，顶点着色计算会导致视觉错误，如棱角分明的高光，且在两个三角形平面上，顶点着色的版本显然是不正确的，造成这些错误的原因是着色方程的部分，尤其是高光部分，在网格表面有着非线性变化的值，这使得它们不适合用于顶点着色器，其计算结果会在递交给像素着色器之前在三角形上进行线性插值。**

**原则上来说，可以仅在像素着色器中计算着色模型的镜面高光部分（specular highlight），而在顶点着色器中计算其余部分，这可能不会导致视觉伪像（visual artifacts），并且理论上将节省一些计算，但在实践中，这种混合实现通常不是最佳的，着色模型的线性变化部分往往是计算成本最低的，且以这种方式拆分着色计算往往会增加相当多的开销，比如重复的计算和额外的可变输入，从而导致弊大于利。**

**正如我们之前提到的，在大部分实现中，顶点着色器负责非着色操作，例如几何变换和变形，生成的几何表面属性，转换到合适的坐标系统中，并被顶点着色器写入，在三角面上进行线性插值，然后作为可变着色器输入传入像素着色器，这些属性通常包括表面的位置，表面法线，以及可选的表面切线向量（如果需要法线贴图的话）。**

**注意，即使顶点着色器总是生成单位长度表面法线，插值也是能改变其长度的，见图5.10左侧。基于这个原因，法线需要在像素着色器中重新归一化（缩放至长度为1），但是顶点着色器生成的法线的长度仍然很重要，如果法线长度在顶点间明显不同的话，比如顶点混合的副作用会使插值倾斜，此情况可见图5.10右侧。由于这两个副作用，实现通常会在插值之前与之后，即在顶点着色器和像素着色器中，都进行插值向量的归一化。**

图示

描述已自动生成

**图 5.10. 在左侧，我们看到跨越表面的单位法线的线性插值将导致插值后的向量长度小于1。在右侧，我们看到线性插值的法线有着明显不同的长度，这导致了插值后的方向朝着两个法线中较长者的倾斜。**

**与表面法线不同，指向特殊位置的向量，例如视图向量（view vector）和精确光的光线向量（light vector），通常是不进行插值的，取而代之的是，在像素着色器中插值后的表面位置将被用来计算这些向量。除了在任何情况下都需要在像素着色器中执行的归一化操作外，每个向量都会用到向量减法运算，这是很快的。如果因为某些原因，需要对这些向量进行插值的话，不要事先对它们进行归一化，这会导致错误的结果，如图5.11所示。**

图示

描述已自动生成

**图 5.11. 两个光线向量间的插值。在左侧，在插值前将它们归一化将导致插值后方向不正确。在右侧，对未归一化向量插值，得到了正确的结果。**

**之前我们提到顶点着色器变换表面几何体到“合适的坐标系”，通过统一变量传递到像素着色器的相机与光源位置，通常被应用程序变换到相同的坐标系，这样可以最大程度减少像素着色器将所有的着色模型向量带入相同的坐标空间的工作，但究竟哪个坐标系是“合适”的呢？可能的答案包括全局世界空间以及相机的局部坐标系，或者更罕见的，当前渲染模型的局部坐标系。通常是基于系统性的考虑，例如性能，灵活性和简单性，为整个渲染系统做出选择，举个例子，如果渲染的场景预计包含大量的光源，那么可能选择世界空间以避免变换光源位置，但相机空间可能更好，这样可以更好地优化与视图向量相关的像素着色器操作，并且可能提高精确度。（第16.6节）**

**虽然大部分着色器实现，包括我们将要讨论的案例实现，都遵循上述一般概述，但是总有例外，比如，一些应用程序出于美术风格的原因选择了基于逐图元着色计算的多面外观，这种风格常被称为平面着色（flat shading），图5.12中展示了两个例子。**

电视萤幕画面

低可信度描述已自动生成

人在房间里

低可信度描述已自动生成

**图 5.12. 两款使用平面着色作为风格选择的游戏：肯塔基0号路（Kentucky Route Zero，上图）与癌症似龙（That Dragon, Cancer，下图）**

**原则上，可以在几何着色器中执行平面着色，但是近年来相关的实现通常使用顶点着色器，这是通过将每个图元的属性与其第一个顶点相关联并禁用顶点数值插值来完成的，禁用插值（可以对每个顶点值单独操作）将导致第一个顶点的值传递到图元中的所有像素。**

5.3.2 实现案例 Implementation Example

现在我们会展示一个着色模型实现的案例，正如之前所提，我们正在实现的着色模型与来自公式5.1的扩展的Gooch模型是相似的，但经过了修改以使其能够支持多个光源，它可以被描述为：

|  |  |
| --- | --- |
|  | (5.19) |

通过以下中间计算：

|  |  |
| --- | --- |
|  | (5.20) |

此公式适合公式5.6中的多光源结构，为方便起见，在此重复：

|  |  |
| --- | --- |
|  |  |

此例中，照亮项（lit）和非照亮项（unlit）为：

（译注：照亮表示被着色器中设置的光源照亮的部分）

|  |  |
| --- | --- |
|  | (5.21) |

通过调整冷色的非照亮项贡献值，可使结果看起来更像原始方程。

在大部分典型的渲染应用程序中，材质属性的变化值例如会被存储在顶点数据中，或者更普遍的，存储在纹理中（第六章）。然而为了让这个案例的实现保持简单，我们会假设在整个模型中是一个常数。

该实现方案会使用着色器的动态分支功能去循环处理所有的光源，尽管这种直接的方法可以很好地处理还算简单的场景，但是它并不合适庞大、具有复杂几何体，且拥有许多光源的场景，有效处理大量光源的渲染技术将会在第20章详细介绍。为了简单起见，我们只支持一种类型的光源：点光源，虽然这个实现相当简单，但是它遵循了我们之前提到的最佳实践。

（译注：“最佳实践”应该是指前文中“但在实践中，这种混合实现通常不是最佳的，着色模型的线性变化部分往往是计算成本最低的，且以这种方式拆分着色计算往往会增加相当多的开销，比如重复的计算和额外的可变输入，从而导致弊大于利。”）

着色模型并不是单独实现的，而是在一个更大的渲染框架上下文（context）中实现，该案例在一个简单的WebGL 2应用程序中实现，修改自Tarek Sherif [1623]的“Phong-shaded Cube”WebGL 2案例，但是相同的原则也适用于其他更复杂的框架。

我们将讨论一些GLSL着色器代码和JavaScript WebGL调用的应用程序的示例，目的并不是讲述WebGL API的细节，而是展示一般性的实现原理。我们将以“由内到外”的顺序来讲解实现过程，首先是像素着色器，然后是顶点着色器，最后是应用程序端的图形API调用。

着色器源文件应包含着色器输入与输出的定义，这样的着色器代码才是正确的，正如先前在第3.3节讨论的那样。使用GLSL术语，着色器输入会分为两类，其中之一是统一变量输入集，它们有着应用程序所设置的值，并且在一个绘制调用中保持不变，第二种类型由可变的输入（varying inputs）组成，它们的值可以在着色器调用（像素或顶点）之间改变。这里我们可以看到像素着色器的可变输入的定义，在GLSL中用in来标记，同时还有其输出：

in vec3 vPos ;

in vec3 vNormal ;

out vec4 outColor ;

像素着色器有单独的输出，表示最终的着色颜色，像素着色器的输入与顶点着色器的输出相匹配，其在被输入到像素着色器之前在三角形上进行插值。像素着色器有两个不同的输入：表面位置与表面法线，两者都在应用程序的世界空间坐标系内。统一变量输入的数量非常多，为了简洁，我们仅展示这两个的定义，且这两者都是与光源相关的：

struct Light {

vec4 position ;

vec4 color ;

};

uniform LightUBlock {

Light uLights [MAXLIGHTS];

};

uniform uint uLightCount ;

由于这些是点光源，因此每个光源的定义都包含位置和颜色，它们被定义为vec4而不是vec3，以符合GLSL std140数据布局标准的限制，尽管在这种情况下，std140布局可能导致一些空间的浪费，但是它简化了确保CPU与GPU之间的数据布局一致的任务，这也是我们为什么在该例中使用它的原因。Light结构体的数组是在一个已命名的统一代码块（uniform block）内定义的，这是GLSL的功能，用于将一组统一变量绑定到缓冲区对象，以加快数据传输，数组长度被定义为应用程序允许的在单个绘制调用中光源的最大数量，稍后我们将看到，在编译着色器之前，应用程序将着色器源码中的MAXLIGHTS字符串替换为正确的值(在本例中是10)。统一整数（uniform integer）uLightCount是在绘制调用中的实际的活动光源数。

接下来，我们来看一下像素着色器的代码：

vec3 lit( vec3 l, vec3 n, vec3 v) {

vec3 r\_l = reflect (-l, n);

float s = clamp (100.0 \* dot(r\_l , v) - 97.0 , 0.0 , 1.0) ;

vec3 highlightColor = vec3 (2 ,2 ,2);

return mix( uWarmColor , highlightColor , s);

}

void main () {

vec3 n = normalize ( vNormal );

vec3 v = normalize ( uEyePosition .xyz - vPos );

outColor = vec4 ( uFUnlit , 1.0) ;

for ( uint i = 0u; i < uLightCount ; i++) {

vec3 l = normalize ( uLights [i]. position .xyz - vPos );

float NdL = clamp (dot(n, l), 0.0 , 1.0) ;

outColor .rgb += NdL \* uLights [i]. color .rgb \* lit(l,n,v);

}

}

我们有一个照亮项的函数定义，它被main()函数调用，总的来说，这是公式5.20与公式5.21的GLSL直接实现，需要注意，和是作为统一变量被传入，因为这些值在整个绘制调用中是恒定不变的，因此应用程序可以计算这些值（译注：而不是着色器），从而节省一些GPU的周期。

该像素着色器使用了一些的内置的GLSL函数，reflect()函数反射一个向量（此例中为光线向量），反射位于第二个向量（此例中为表面法线）定义的平面上，由于我们想要光线向量和反射向量都指向远离表面的位置，我们需要在将前者传递给reflect()之前对其取反。clamp()函数有三个输入值，其中的两个输入值定义了第三个值被约束的范围，一个特别的例子是约束范围在0和1之间（这时就相当于HLSL的saturate()函数），该特例速度很快，在大部分GPU上没什么消耗，这也是我们使用它的原因，虽然我们只需要约束值到0，因为我们知道它不会超过1。mix()函数也有三个输入值，并基于第三个值（一个0到1之间的混合参数）在其中两个值（此例中，指的是暖色与高光色）之间进行线性插值，在HLSL中这个函数称为lerp()，意思是“线性插值”（linear interpolation）。最后是normalize()函数，它会将向量除以其长度，将其长度缩放为1。

现在让我们看一下顶点着色器，我们不会展示它的任何统一变量定义，因为我们已经在像素着色器里看过一些例子了，但是可变输入和输出的定义还是值得查看一下的：

layout ( location =0) in vec4 position ;

layout ( location =1) in vec4 normal ;

out vec3 vPos ;

out vec3 vNormal ;

注意如前所提，顶点着色器的输出对应像素着色器的可变输入，顶点着色器的输入中包含指定了数据在顶点数组中如何布局的指令，顶点着色器的代码如下：

void main () {

vec4 worldPosition = uModel \* position ;

vPos = worldPosition .xyz;

vNormal = ( uModel \* normal ).xyz;

gl\_Position = viewProj \* worldPosition ;

}

这些是顶点着色器的常规操作，着色器变换表面位置与法线到世界空间，并将它们传入像素着色器以便在着色过程中使用。最终，表面位置被变换到裁剪空间并且传入gl\_Position，一个被光栅器使用的特殊系统定义变量，gl Position变量是所有顶点着色器必需的输出。

注意，法线向量在顶点着色器中并没有归一化，它们不需要被归一化是因为它们在原始网格模型中的长度为1，并且应用程序没有执行任何可能会非均匀地改变它们长度的操作，例如顶点混合或非均匀缩放。模型矩阵可以有一个统一的缩放因子，但那会按比例地改变所有法线的长度，因此不会导致图5.10右侧中显示的问题。

应用程序为了各种不同的渲染和着色器设置而使用WebGL API，每个可编程着色器阶段都被单独地设置，并且它们都被绑定到程序对象上。以下是像素着色器的设置代码：

var fSource = document . getElementById (" fragment "). text . trim ();

var maxLights = 10;

fSource = fSource . replace (/ MAXLIGHTS /g, maxLights . toString ());

var fragmentShader = gl. createShader (gl. FRAGMENT\_SHADER );

gl. shaderSource ( fragmentShader , fSource );

gl. compileShader ( fragmentShader );

请注意提到的“片元着色器”，是WebGL（以及它所基于的OpenGL）中使用的术语，正如我们之前在本书中提到的，虽然“像素着色器”在某些方面描述不够精确，但它是更普遍的称呼方式，所以我们将继续在本书中使用这个称呼。这段代码也是将MAXLIGHTS字符串替换成合适数值的地方，大部分渲染框架执行类似的预编译着色器操作。

这里还有很多应用程序端代码用于设置统一变量，初始化顶点数组，清除，绘制等，你可以在程序[1623]中查看这些代码，并且许多API指南对此进行了说明。我们在此的目标是通过它们自身的编程环境，了解着色器是怎样被当作单独的处理器，因此我们在此结束本小节。

5.3.3 材质系统 Material Systems

正如我们的简单案例里，渲染框架很少只实现单个着色器，通常来说，需要一个专用的系统来处理大量的材质，着色模型，以及应用程序所使用的着色器。

正如之前章节所述，着色器是用于GPU的可编程着色阶段之一的一个程序，因此，着色器是低级的图形API资源，并且不是美术人员会直接接触的，与之相反，材质是面向美术人员封装的表面的视觉表现，材质有时也描述非视觉部分，比如碰撞属性，我们不会继续深入此话题，因为它已经超出了本书的范围。

虽然材质通过着色器实现，但不是简单的一对一对应关系，在不同的渲染情况下，相同的材质可能使用不同的着色器，一个着色器也可能被多种材质共用。最普遍的情况是材质参数化，在它的最简单形式中，材质参数化需要两种类型的材质实体：材质模板（material templates）与材质实例（material instances），每个材质模板描述一类材质，并且有一组参数，可以根据参数类型去分配数值，颜色，或者纹理的值，每个材质实例对应着一个材质模板与所有参数的一组特定值。一些渲染框架例如虚幻引擎[1802]允许更复杂的、分层的结构，其中材质模板派生自多层次的其他模板。

参数可以在运行时被解析，通过统一变量输入传递到着色器程序，或者也可以在编译时，通过在着色器编译前替换值来解析参数。一个常见的编译时参数类型是布尔开关量，用来控制激活给定材质的特征，这可由美术人员通过材质UI的复选框去设置，或者由材质系统在程序上设置，例如对于远处的物体，当它们的视觉效果特征可以忽略不计时，可减少着色器消耗。

尽管材质参数可以与着色模型参数一对一匹配，但我们不是总会遇到这种情况，一个材质可能会修改一个给定着色模型参数的值，例如表面颜色，可修改为一个常量，或者可以将多个材质参数以及插值的顶点或纹理值作为输入，通过一系列复杂的操作来计算着色模型参数。在某些情况下，表面位置，表面方向，甚至时间等参数也可能是计算的因素，基于表面位置和方向的着色在地形材质中尤为常见。举个例子，高度与表面法线可以被用来控制积雪特效，做法是在高处的水平面和接近水平面的表面以白色表面颜色做混合。基于时间的着色通常用于动画材质，例如闪烁的霓虹灯标志。

材质系统的最重要任务之一是将多种着色器函数划分为单独的元素，并控制这些元素的组合方式，在许多情况下，这种组合是很有用的，包括以下几种情况：

* 将表面着色与几何处理组合在一起，例如刚体变换，顶点混合，变形，曲面细分，实例化，以及裁剪，这些功能都是各不相同的：表面着色器依赖于材质，几何处理依赖于模型网格，所以，分开编写它们并让材质系统根据需求组合它们是很方便的。
* 将表面着色与一些组合操作例如像素丢弃（discard）与混合（blending）组合在一起，这与移动端GPU是尤为相关的，在移动端GPU的像素着色器里，混合是一种普遍执行的操作，通常我们希望独立于用于表面着色用的材质来选择这些操作。
* 将用来计算着色模型参数的操作与着色模型自身的计算组合在一起，这种方式允许编写一次着色模型的实现，并以与多种计算着色模型参数的方法组合一起的方式重用。
* 将独立可选的材质特征相互之间，以及与选择逻辑，剩余的着色器组合在一起，这种方式使得可以分别编写每个特征的实现。
* 将着色模型和用光源计算法（light source evaluation）其参数的计算组合在一起：在每个光源的着色点计算和的值，有些技术例如延迟渲（将在第20章讨论）改变了这个组合的结构，在支持该技术的渲染框架里，这种方式将增加额外的杂度。

如果图形API提供这种类型的着色器代码模块作为核心功能，将会很方便，可惜的是，不像CPU代码，GPU着色器不允许对代码片段进行后编译链接，每个着色器阶段的程序都作为一个单元被编译。着色器阶段之间的分离确实提供了一些有限的的模块化，这在某种程度上适合我们列表上的第一项：将表面着色（通常在像素着色器上执行）与几何处理（通常在其他着色阶段执行）组合在一起，但这个匹配并不完美，因为每个着色器也会执行其他操作，并且其他类型的组合仍然需要被处理。由于存在这些限制，材质系统能实现所有组合类型的唯一办法在源代码层级，主要包括如并列和替换的字符串操作，通常通过C风格的预处理指令例如#include，#if和#define去执行。

早期的渲染系统有着相对较小数量的着色器变体（shader variants，译注：这里提到的着色器变体很突兀，搜索了一下并结合上下文，其表示一个shader特定的变种。包含预编译宏等方式以控制不同分支的shader源代码，通过使用不同关键字编译，生成的不同shader就是变体），并且通常是手动去编写每个变体，这也有一些好处的，例如可以在充分了解最终的着色器程序的基础上去优化每个变体，然而这种手动编写的方法随着着色器变体数量的增加而变得不切实际。当我们将所有不同部分和选项都纳入考虑时，可能的不同着色器变体数量是巨大的，这就是为什么模块化和可组合性如此关键。

当设计一个系统用来处理着色器变体时，第一个需要解决的问题是，不同选项间的选择是否是通过动态分支（dynamic branching）在运行时执行，或者是在编译时通过条件预处理（conditional preprocessing）执行。在较老的硬件上，动态分支通常是不可能的或者是速度极慢的，所以运行时的选择是不可行的，所以所有变体都在编译阶段被处理，包括不同光源类型计数的所有可能的组合[1193]。

不过，如今的GPU能够很好地处理动态分支，尤其是在一个绘制调用中分支对所有像素有相同行为的情况下。现在许多功能性变体，例如光源的数量，都在运行时被处理，然而，为一个着色器添加大量的功能变体将产生一个不同的消耗：寄存器计数（register count）的增加和占用率的相应降低，进而导致性能下降，详情可见第18.4.5节。所以编译时变体仍然是有价值的，它能避免包含那些从不执行的复杂逻辑。

举个例子，让我们想象一个支持三种不同类型光源的应用程序，其中两种光源类型很简单：点光源与方向光，第三种类型是通用的聚光灯，它支持表格式照明模式（tabulated illumination patterns）以及其他复杂的功能，这需要大量的着色器代码去实现，但是这个通用的聚光灯使用率相对较小，此应用程序中只有不到5%的光源是这种类型。在过去，一个单独的着色器变体会为每个可能的三种光源的计数组合去编译，以避免动态分支，尽管如今已不再需要这种方式，但是编译两个单独的变体仍然是有好处的，一个变体适用于通用聚光灯数量大于等于1时，另一个变体适用于此类聚光灯数量正好为0时。由于第二个变体的代码更简单，它（更常被使用）可能有着更低的寄存器占用率，并且因此有着更高的性能表现。

现代材质系统同时使用了运行时着色器变体和编译时着色器变体，即使完整的负载已经不会仅在编译时处理，但总体的复杂度与变体的数量仍然保持增长，所以还是需要编译大量的着色器变体。举个例子，在游戏《命运：被夺走的国王》（Destiny: The Taken King）的某些场景中，单帧内使用了超过9000个编译期着色器变体[1750]，可能的变体数量还可以变得更为巨大，例如Unity渲染系统有着接近1000亿可能的变体，只有确确实实被使用的变体才会被编译，但是着色器编译系统必须被重新设计以处理大量可能的变体[1439]。

材质系统设计者使用不同的策略去解决这些设计目标，虽然这些策略有时候表现为互斥的系统体系结构[342]，但这些战略能够被——也通常是——结合到相同的系统中，这些策略包含以下内容：

* 代码重用——在共享的文件中实现函数，使用#include预处理指令去访问那些任意着色器的所需函数。
* 减法——使用一个通常被称为超级着色器（üershader或supershader）[1170,1784]的着色器，它聚集了一大批功能，使用编译时预处理器与动态分支的组合去移除无用的部分并在互斥的备选方案之间切换。
* 加法——将各种功能定义为具有输入输出连接器的节点，并将它们组合到一起，这与代码重用策略相似，但是更结构化。节点的组合可以通过文本[342]或一个可视图形编辑器来完成，后者旨在使非工程师（例如技术美术）更容易编写新的材质模板[1750，1802]，通常来说，可视化图形编写只能访问到着色器的一部分，例如在虚幻引擎的可视化图形编辑器中，只能作用到着色模型输入相关的计算，如图5.13所示。

电脑主机

中度可信度描述已自动生成

**图 5.13. 虚幻引擎材质编辑器，注意在节点图右侧的高的节点，这个节点的输入连接器对应着渲染引擎使用的多种着色输入，包括所有的着色模型参数。（材质示例由Epic Games提供）**

* 基于模板——定义了一个接口，不同的实现只要符合该接口，就可以接入该接口，这比加法策略更加正式，因此通常被用于更大的功能块中。基于接口的一个普遍案例是着色模型参数计算与着色模型自身计算的分离，虚幻引擎[1802]有着不同的“材质域”，包括计算着色模型参数的表面域，以及计算给定光源用来调整的标量的光照函数域，一个与此相似的“表面着色器”结构也存在于Unity中[1437]。注意延迟着色技术（将在第20章讨论）强制采用了一个类似的结构，采用G缓冲区用作接口。

更具体的案例，书籍《WebGL Insights》（现在免费）中的一些章节讨论了各种引擎是怎样控制它们的渲染管线。除了组合外，现代材质系统还有一些重要的设计注意事项，例如如何以最少的着色器代码重复去支持多平台，这包括函数性的变体，以解决平台，着色语言和API之间的性能和功能差异。游戏《命运》的着色器系统（The Destiny shader system）[1750]是对这种问题的最具代表性的解决方案，它采用了一个专有的预处理器层从而使着色器能够由一个自定义的着色器语言书写，这允许我们编写与平台无关的材质，然后自动翻译成不同的着色语言与具体实现，虚幻引擎[1802]与Unity[1436]有着相似的系统。

材质系统也需确保具有好的性能表现，除了着色变体的专门编译外，材质系统还能执行一些其他常见的优化。《命运》着色器系统以及虚幻引擎会自动检测那些在一次绘制调用中保持恒定的计算（例如在之前实现案例中提到的暖色与冷色计算），并且将它们移到着色器之外。另一个例子是在《命运》中使用的作用域系统（scoping system），它用来区分以不同频率更新的常数（例如每帧一次，每个光源一次，每个物体一次），并在合适的时间更新每组常数以减少API的开销。

正如我们所见过的，实现一个着色方程的关键是决定哪个部分可被简化，各种表达式计算的频率，以及用户能够如何修改和控制其表现，渲染管线的最终输出是颜色和混合值。剩余的部分是反走样，透明度，与图像显示的部分详细介绍了这些值将怎样合并与修改以用于显示。

5.4 走样与反走样 Aliasing and Antialiasing

**（译注：aliasing可译为走样，锯齿，混叠，失真）**

**想象一个巨大的黑色三角形缓缓地穿过一个白色背景，当屏幕网格单元格（screen grid cell）被三角形覆盖，代表这个单元格的像素值强度应该平滑地下降，然而通常发生在所有类型的基础渲染器中的情况是，当网格单元的中心被覆盖的那一刻，像素颜色立即从白色变为黑色，标准GPU的渲染也不例外，见图5.14的最左列。**

**三角形在像素里的显示是要么存在，要么不存在，线的绘制也有类似的问题，边缘有着锯齿状的外观，因此这种视觉伪像被称作“锯齿”（the jaggies），当物体运动时则被称作“爬虫”（crawlies），此问题更正式的称呼为“走样”（aliasing），同时我们称旨在避免这个问题的相关技术为“反走样”（antialiasing）。**

**关于采样理论与数字滤波的话题已经足够另外写一本书了[559，1447,1729]，因为这是渲染的关键领域，我们会在这里阐述采样和滤波的基础理论，接下来我们会专注于当前在实时渲染中我们能做一些什么以减轻走样伪像（aliasing artifacts）。**

图片包含 游戏机, 画, 钟表

描述已自动生成

**图 5.14. 上排图像显示了三个不同反走样级别的三角形、线、点，下排图像是上排图像放大后的样子。最左列每个像素仅用一个采样，这意味着没有使用反走样技术，中间列的图像以每像素四个采样（以网格模式）的方式渲染，最右列则是使用每像素八个采样（在4×4 的棋盘格中，一半的正方形被采样）。**

5.4.1 采样与滤波理论 Sampling and Filtering Theory

**渲染图像的处理本身便是一个采样任务，之所以这样是因为图像的生成就是三维场景采样的处理过程，其目的是获取图像中每个像素（一个离散的像素数组）的颜色值。为了使用纹理映射（第6章），纹素（texels）必须能被重采样以在各种条件下获得好的结果，为了在动画文件中生成一系列图像，动画文件通常以均匀的时间间隔采样。本节将介绍采样，重建，以及滤波的话题，为了简单起见，大部分材质将以一维呈现，这些概念也可自然地扩展到二维，因此能够在处理二维图像时使用。**

**（注：纹素（英语：Texel，即textureelement或texture pixel的合成字）是纹理元素的简称，它是计算机图形纹理空间中的基本单元，如同图像是由像素排列而成，纹理是由纹素排列表示的。）**

**图5.15展示了一个连续的信号是怎样以均匀的间隔被采样的，这称之为离散化（discretized），采样处理的目标是数字化地呈现信息，这样做是为了减少信息量。然后采样的信号需要被重建（reconstructed）以恢复原始信号，这是通过对采样信号进行滤波（filtering）来完成的。**

图表, 条形图, 直方图

描述已自动生成

**图 5.15. 一个连续信号（左图）被采样（中图），接下来原始信号通过重建被恢复（右图）。**

**无论何时进行采样，都可能出现走样，这是我们不想造成的伪像，我们需要与走样进行战斗以生成令人满意的图像。西方老年人见过的一个关于走样的经典案例，是电影摄像机拍摄的一个旋转的马车车轮，由于车轮辐条移动得比摄像机记录图像的速度快得多，车轮看起来像是在向后或向前缓慢旋转，或者甚至有可能看起来根本没有转动，如图5.16所示，之所以出现这种现象，是因为车轮的图像是以一系列时间步长被记录的，这被称作时间走样（temporal aliasing）。**

形状

中度可信度描述已自动生成

**图 5.16. 最上面一行显示一个纺车的车轮(原始信号)。在第二行采样不足，使它似乎在相反的方向移动，这是一个由于采样率太低导致走样的例子。在第三行中，采样率正好是每转两个样本，我们不能确定轮子旋转的方向，这就是奈奎斯特极限（Nyquist limit）。在第四行中，采样率高于每转两个样本，我们突然可以看到轮子在正确的方向上旋转。**

**在计算机图形中走样的普遍案例有光栅化后的线与三角形边缘的“锯齿”，被称为“萤火虫”（fireflies）的闪烁的高光，以及缩小具有方格模式的纹理时发生的走样（见6.2.2节）。**

**当一个信号被以过慢的频率采样时，就会出现走样，采样后的信号看起来是一个频率比原始信号低的信号，如图5.17所示。为了使一个信号被合适地采样（换句话说，这样就能够从样本中重建原始信号），采样频率必须大于被采样信号最大频率的两倍，这通常被称作采样定理（sampling theorem），并且该采样频率被称为奈奎斯特率（Nyquist rate）[1447]或奈奎斯特极限（Nyquist limit），以一位在1928年发现此频率的瑞典科学家哈里·奈奎斯特（Harry Nyquist）(1889–1976)命名，奈奎斯特极限也在图5.16进行了描述。该定理使用术语“最大频率”的这一事实暗示着信号应该受到频带限制（band-limited），而这意味着任何频率都不能超过特定限制，换句话说，信号相对于相邻样本间的间隔应该足够平滑。**

图表, 折线图

描述已自动生成

**图 5.17. 蓝色实线为原始信号，红色圆圈表示间距均匀的采样点，绿色虚线为重构信号。上图显示了过低的样本率，导致重构信号看上去频率更低，即原始信号的混叠信号。下图显示了采样率正好是原始信号频率的两倍，重构出来的信号是一条水平线。可以证明，如果采样率稍微提高一点，完美的重建是可能的。**

**当一个三维场景以点样本渲染时，正常情况下是不会有频带限制的，但三角形的边缘，阴影边界，以及其他现象会产生不连续变化的信号，因此会产生无限的频率[252]，同时无论样本被打包得多紧密，物体仍然能足够小，以至于它们根本不能被采样，因此当我们使用点样本渲染场景时，完全避免走样的问题是不可能的，而我们几乎总是使用点采样。不过有时候我们可以知道信号在何时有频带限制，其中一个例子是当纹理被应用到表面的时候，这时计算纹理采样频率与像素采样率之比是可能的，如果该频率低于奈奎斯特极限，那就不用做特殊的操作以对纹理进行合适的采样，如果频率过高的话，就要使用各种算法对纹理进行频带限制（第6.2.2节）。**

**重建 Reconstruction**

**我们现在来讨论给定一个频带受限的采样信号，原始信号如何从采样信号中重建，为做到这点，我们必须用到一个滤波器，三个普遍使用的滤波器如图5.18所示。需要注意的是，滤波器的面积应该总是为1，否则重建的信号可能会扩大或收缩。**

图表, 折线图

描述已自动生成

**图 5.18. 左上为box滤波器（box filter），右上为tent滤波器（tent filter），底部为sinc滤波器（sinc filter）（这里已经clamped在x轴上）**

电脑屏幕的照片

中度可信度描述已自动生成

**图 5.19. 采样后的信号（左侧）使用box滤波器进行重建，这是通过以下步骤完成的：首先在每个采样点上放置box滤波器，并且在y方向上将其缩放，使滤波器的高度与采样点相同，之后求出的和就是重建后的信号（右侧）。**

**在图5.19中，相邻的box滤波器被用于重建一个采样信号，这是所使用的最糟糕的滤波器，因为产生的结果信号是一非连续的阶梯状，然而由于它很简单，所以仍然经常在计算机图形学中使用。如图所示，box滤波器被放置在每个采样点上，之后会被缩放，这样滤波器最上方的点就可以与样本上的点重合，所有这些缩放与平移后的box函数之和就是右侧所示的重建后的信号。**

**box滤波器可以替换成任意其他滤波器，在图5.20中，tent滤波器，也被称作三角形滤波器，被用来重建采样后的信号，注意这个滤波器在相邻采样点之间实施了线性插值，所以它比box滤波器要更好，因为重建的信号现在是连续的。**

卡通人物

中度可信度描述已自动生成

**图 5.20. 采样后的信号（左侧）使用tent滤波器进行重建，重建后的信号如右侧所示。**

**然而使用tent滤波器重建的信号的平滑程度并不好，在采样点有突然的斜率改变，这与tent滤波器并非是完美的重建滤波器这一事实有关。为了得到完美的重建，必须使用理想的低通滤波器，其中信号的频率分量是正弦波：，f是该分量的频率，考虑到这一情况，低通滤波器将去除频率高于滤波器定义的某个频率的所有频率分量，直观的，低通滤波器移除了信号的尖锐特征，即滤波器对信号做了模糊处理。理想的低通滤波器是sinc滤波器（见图5.18底部）：**

|  |  |
| --- | --- |
|  | (5.22) |

**傅里叶分析[1447]的理论解释了为什么sinc滤波器是理想的低通滤波器，简单来说，理由如下：理想的低通滤波器是频率域的box滤波器，当它与信号相乘时，移除了所有高于滤波器宽度的频率，将box滤波器从频域转到空间域就得到了sinc函数，与此同时，乘法操作被转换为了卷积（convolution）函数，卷积是我们在本节中一直使用，但却没有实际描述过的术语。**

图表, 图示, 雷达图

描述已自动生成

**图 5.21. 此处，sinc滤波器被用来重建信号，sinc滤波器是理想的低通滤波器。**

**使用sinc滤波器重建信号能得到更平滑的结果，如图5.21所示。采样过程在信号中引入了高频部分（突变），而低频滤波器的任务就是移除这些高频部分，事实上，sinc滤波器消除了所有频率大于1/2采样率的正弦波，sinc函数如公式5.22所示，当采样频率是1.0时（即采样信号的最大频率必须小于1/2），它是完美的重建滤波器。更一般地，假设采样频率是****，也就是说，相邻样本的间隔为，对这种情况来说，****是完美的重建滤波器，它消除了所有高于的频率，这在重采样中是很有用的（下一节），然而sinc滤波器的宽度是无限的，并且在某些区域是负值，所以它在实践中很少使用。**

**一方面在低品质的box与tent滤波器之间存在有用的中间区域，另一方面是不切实际的sinc滤波器，大部分广泛使用的滤波器函数[1214,1289,1413,1793]处于这些极端情况之间，所有这些滤波器函数都有一点类似sinc函数，但是对它们影响的像素数量有所限制。最接近sinc函数的滤波器在它们的部分域上有负值，对于应用程序而言，负滤波器值是不需要也不实用的，我们通常使用有着非负瓣（no negative lobes）的滤波器[1402]（通常被称作高斯滤波器，因为它们源于或类似于高斯曲线），第12.1节更详细地讨论了滤波器函数以及它们的使用。**

**在使用任意滤波器之后，便得到了一个连续的信号，然而在计算机图形学中我们不能直接显示一个连续的信号，但是我们可以重采样连续信号使其变为另一个大小，即放大或缩小信号，这个话题将在接下来讨论。**

**重采样 Resampling**

**重采样被用来放大或缩小一个被采样信号，假设原采样点位于整数坐标系内（0, 1, 2, ...），即样本间的间隔是单位整数，更进一步，假设在重采样后，我们想要新的采样点以样本的间隔a均匀地放置，如果a > 1，产生缩小（下采样），对于a < 1，产生放大（上采样）。**

**放大是两种情况中较为简单的一个，让我们从放大开始。假设采样信号如上一节所示那样被重建，直观地，因为信号现在已经被完美重建并且是连续的，所需要的便是以我们期望的间隔去重采样重建后的信号，该过程如图5.22所示。**

图表, 图示

描述已自动生成

**图 5.22. 左边是采样信号和重建信号，****右边是重建信号以两倍的采样频率进行重采样，即发生了放大（译注：频率放大，间隔缩小）。**

**然而当缩小时，这个技术不起作用，原始信号的频率对采样率来说过高，以至于无法避免走样，不过已经证明了使用的滤波器可被用来从采样信号中创建连续信号[1447,1661]，之后便可以用期望的间隔进行重采样，如图5.23所示。换句话说，通过使用作为滤波器，低通滤波器的宽度增加了，以至于更多的信号高频率内容被移除了，正如图中所示，滤波器（sinc函数专属的）宽度被翻倍以减少重采样率到原采样率的一半，将此与数字图像联系起来，这类似于先模糊它(去除高频)，然后以较低的分辨率对图像进行重采样。**

图表, 图示

描述已自动生成

**图 5.23. 左边是采样信号和重建信号，右边滤波器宽度放大为原来的两倍，以使样本之间的间隔变为原来的两倍，即发生了缩小。**

**以采样和滤波理论为框架，现在讨论在实时渲染中用于减少走样的各种算法。**

5.4.2 基于屏幕空间的反走样 Screen-Based Antialiasing

如果采样与滤波的效果不好，三角形的边缘会产生明显的伪像，阴影边缘，高光，以及颜色迅速变化的其他现象都可能导致类似的问题，在本小节中讨论的算法会帮助前述的这些案例提升渲染品质，它们的共同点是，它们都是基于屏幕空间的，即它们只在管线输出的采样样本上进行操作。没有最好的反采样技术，每一种技术都有不同的质量优势，如捕捉敏锐的细节或者其他现象的能力，运动时的表现，内存消耗，GPU硬件要求，以及速度等等。

在如图5.14所示的黑色三角形案例中，存在一个问题是采样率较低，在每个像素的网格单元中心进行单个采样，因此对于该网格单元，最多被了解的是它是否被三角形所覆盖（译注：0 or all）。通过在每个屏幕网格单元使用更多的采样并以一些方式将它们混合，就能计算出更好的像素颜色，如图5.24中所示。

形状

描述已自动生成

**图 5.24. 左侧，以像素中心采样一次的方式渲染一个红色三角形，因为三角形并没有覆盖样本，像素是白色，即使像素的大部分已经被红色三角形覆盖。右侧，对每个像素采样四次，可以看到其中两个采样点被红色三角形所覆盖，因此像素为粉红色。**

基于屏幕的反走样方案的一般策略是使用一个针对屏幕的采样模式，并对这些样本进行加权与求和，得到像素的颜色，p：

|  |  |
| --- | --- |
|  | (5.23) |

式中是用于单个像素的采样数，函数是一个采样颜色，是权重，范围是[0,1]，样本对整个像素的颜色有所贡献，样本的位置根据其在序列中的顺序来确定，如1, …, n，并且可选函数也是使用像素位置(x,y)的整数部分，换句话说，每个样本在屏幕网格的采样位置都是不同的，并且每个像素可以选用不同的采样模式。在实时渲染系统（以及大多数其他渲染系统）中，样本通常是点样本，所以，函数c可以被认为是两个函数，首先函数f(i,n)检索所需样本在屏幕上的位置浮点数()，然后对屏幕上的该位置进行采样，即检索该精确点处的颜色，选择采样方案，并且配置渲染管线以计算特定子像素位置的采样，这通常是基于逐帧（或逐应用）设置的。

在反走样中的另一个变量是，每个样本的权重，这些权重的和为1，大部分用于实时渲染系统的方法都对它们的样本给出了统一权重，即。图形硬件的默认模式，像素中心的单个采样，是上述反走样方程的最简单情况，只有一个项，该项的权重为1，并且采样函数f总是返回被采样像素的中心。

每个像素计算超过一个完整采样以上的反走样算法，被称作超级采样（supersampling）方法（或过采样方法（oversampling)））。简单地从概念上说，全场景反走样（full-scene antialiasing, FSAA），又名“超级采样反走样”（supersampling antialiasing, SSAA），以更高的分辨率渲染场景，然后对相邻的样本进行滤波以得到图像。举个例子，假设我们需要一张1280×1024像素的图像，如果你在屏幕外渲染一个2560×2048像素的图像，然后对每2×2的像素区域取平均值，我们需要的图像就会以每像素采样四次的方式生成，并使用box滤波器进行滤波，需要注意的是，这相当于图5.25中的2×2网格采样，此方法比较消耗性能，因为所有的子采样必须被完整地着色与填充，其中每个样本都具有z缓冲区的深度信息。FSAA的主要优点在于简单，此外，这种方法的低质量版本只在一个屏幕轴向上以两倍的速率采样，因此被称为1×2或2×1超级采样，通常为了简便，使用二次幂分辨率和box滤波器。英伟达（NVIDIA）的动态超分辨率（dynamic super resolution）功能是一个更加复杂的超级采样形式，其中以更高的分辨率渲染场景，并且使用13个采样的高斯滤波器（13-sample Gaussian filter）去生成显示的图像[1848]。

图片包含 信件

描述已自动生成

**图 5.25. 一些像素采样方案的对比，按照单个像素采样数从少到多排列，Quincunx共享角落的样本以及对中心样本进行加权，以使其值达到像素最终颜色的一半。2×2旋转网格比2×2直形网格在几乎水平的边缘上会捕获更多的灰度级。类似地，8rooks模式为此类线条捕获的灰度级别比4×4网格更多，尽管其使用的样本更少。**

有一个与超级采样相关的采样方法基于累积缓冲区（accumulation buffer）[637,1115]，该方法不使用一个大的离屏缓冲区，而是使用一个与最终期望图像具有相同分辨率的缓冲区，但是每个颜色槽（channel）使用更多的字节位，为得到一个场景的2×2采样，生成了四幅图像，视图根据需要在屏幕x轴或y轴上移动半个像素，每个生成的图像都是基于网格单元内的不同采样位置，每帧必须重新渲染场景几次，并将结果复制到屏幕上，这种额外花费使该算法在实时渲染系统中成本很高，当性能不是关键时，这种方法对生成高质量的图像来说是很有用的，因为每个像素可以使用任何数量的样本，并且可以放置在任何地方[1679]。它曾直接被OpenGL API支持，但是在3.0版本中被弃用，在现代GPU中，累积缓冲区这个概念可通过在输出缓冲区使用高精度的颜色格式，从而在像素着色器中实现。

当物体边缘、镜面高光和锐利阴影等现象引起突变的颜色变化时，需要额外的采样样本，这样阴影通常能够变得更软，高光也可以变得更平滑以避免走样，可以增加特定类型对象的大小，例如电线，以保证它们在长度上每个位置至少覆盖一个像素[1384]。物体边缘的走样仍然是一个主要的采样问题，在渲染中当物体边缘被检测以及它们的影响被考虑在内时，可以使用分析方法（analytical methods），但是这些方法通常更为昂贵，并且与简单地进行更多的采样相比，健壮性更低。不过，GPU的功能例如保守光栅化和光栅化顺序视图开启了新的可能性[327]。

像超级采样与累积缓冲区等技术，它们通过生成完全由单独计算的阴影和深度指定的样本来工作，由于每个样本都必须通过像素着色器，因此总体收益相对较低，性能消耗也较高。

多重采样反走样（Multisampling antialiasing, MSAA）通过每像素一次的表面着色计算，并在样本间共享计算结果，从而降低了高额的计算成本。Pixels may have, say,（译注：不会译）每个片元有四个(x,y)的样本位置，每个都有它们自己的颜色与z深度值，但是对于应用到像素的每个对象的片元，像素着色器只进行一次计算。如果所有的MSAA位置样本都被片元覆盖，那么着色样本就会在像素的中心被计算，相反如果片元覆盖较少的位置样本，则着色样本的位置可以移动，以更好地表示所覆盖的位置，这么做可以避免例如纹理边缘的着色采样，这种位置调整方法被称作质心采样（centroid sampling）或质心插值（centroid interpolation），如果开启此功能，GPU将自动完成。质心采样能避免出现三角形外的问题（off-triangle problems），但会导致导数计算返回不正确的值[530, 1041]，见图5.26。

形状

描述已自动生成

**图 5.26. 中间图，一个像素中有两个物体重叠，红色物体覆盖了三个样本，蓝色物体只有一个，像素着色器计算该位置以绿色显示。因为红色三角形覆盖了像素的中心，这个位置被用于着色器计算，用于蓝色物体的像素着色器在此样本位置进行计算。对于MSAA来说，所有四个位置中存储了单独的颜色与深度值，在右侧展示了EQAA的2f4x模式（the 2f4x mode for EQAA），四个样本现在有四个ID值，这些ID索引了一张存储起来的表，表内有两种颜色和深度的信息。**

MSAA比纯粹的超级采样方案快，因为片元只进行一次着色，它致力于以高频率对片元的像素覆盖区域进行采样，以及共享计算出的着色数据。通过进一步分离采样和覆盖范围（coverage），可以节省更多的内存，这反过来又可以使反走样的速度更快——使用的内存越少，渲染速度就越快。英伟达（NVIDIA）在2006年推出了覆盖采样反走样（coverage sampling antialiasing, CSAA），AMD随后推出了增强质量反走样（enhanced quality antialiasing, EQAA），这些技术的工作原理是，以更高的采样率存储仅被覆盖的片段。例如，EQAA的“2f4x”模式存储了两个颜色与深度值，在四个样本位置之间共享，颜色与深度值信息不再储存在特定的位置里，而是储存在一张表中，四个样本每个只需要一位（bit）空间用来指定两个存储值中的哪个与其位置相关联，见图5.26。覆盖样本明确规定了每个片元对最终像素颜色的贡献，如果储存的颜色数量超出限制，一个储存的颜色就会被移除并且对应的样本会被标记为未知，这些样本对最终颜色不产生贡献[382,383]，对大多数场景来说，相对较少的像素会包含三个或更多的在着色上完全不同的可见不透明片元，所以这个方案在实践中表现良好[1405]。然而对于最高品质来说，游戏极限竞速：地平线2（Forza Horizon 2）运行时会使用4倍MSAA，即便EQAA有更好的性能优势[1002]。

一旦所有的几何体被渲染到一个多重采样缓冲区，将会执行一个解析（resolve）操作，这段程序会将总体样本颜色进行平均以决定像素的颜色，值得注意的是，当使用具有高动态范围的颜色值进行多重采样时，可能会出现一个问题，在这种情况下，为了避免伪像，在进行解析操作前，你通常需要对颜色值进行色调映射[1375]，这个开销可能很昂贵，所以可以使用更简单的色调映射函数的近似函数或者其他方法[862,1405]。

默认的情况下，MSAA通过box滤波器进行解析，2007年，ATI推出了自定义滤波器反走样（custom filter antialiasing, CFAA）[1625]，它能够使用更狭窄或更宽的tent滤波器并且稍微拓展到其他像素格，这种模式已经被EQAA支持所取代。在现代GPU上，像素或者计算着色器能够访问MSAA的样本并且使用任何我们所期望的重建滤波器，包括从像素周围样本中得到的样本，一个更宽的滤波器能减少走样，虽然丢失锐利的细节，Pettineo[1402, 1405]发现立方体的smoothstep以及有着2或3像素宽度的B样条滤波器总体上得出了最好的结果（译注：Smoothstep is a scalar interpolation function commonly used in computer graphics Smoothstep at Microsoft Developer NetworkGLSL Language Specification, Version 1.40 and video game engines. 维基百科）。当然还有性能消耗，因为即使使用自定义着色器模拟默认的box滤波器解析也会花费很长的时间，而一个更宽的滤波器核心意味着增加了样本的访问成本。

英伟达（NVIDIA）内建的TXAA支持与之类似，在比单个像素更大的区域上使用了更好的重建滤波器，以获得更好的结果，它和更新的多帧反走样（multiframe antialiasing, MFAA）方案都使用了时间性反走样（temporal antialiasing, TAA），这是一类通用技术，它可以使用之前帧的结果用来改进图像，在某种程度上，这种技术之所以成为可能，是因为程序员可以设置每帧MSAA采样模式[1406]，这种技术可以解决例如旋转的马车车轮等反走样问题，并且能够改进边缘渲染质量。

想象通过生成一系列图像来“手动”执行采样模式，其中每次渲染使用不同的位置进行采样，这种偏移是通过在投影矩阵上附加一个微小的平移来完成的[1938]，生成和取平均的图像越多，结果就越好，这种使用多个偏移图像的概念被用于时间性反走样算法。可能使用MSAA或其他方法生成单个图像，然后与之前的图像做混合，通常只有2~4帧被使用[382, 836, 1405]，较旧的图像被赋予的权重可能呈指数减小[862]，尽管如果观察者和场景不移动，这可能会导致帧闪烁（shimmering），所以通常情况下，最后一帧和当前帧的权重相等。由于每帧的样本位于不同的子像素位置，这些样本的加权和比单个帧能更好地估计边缘的覆盖范围，因此，使用最新的两帧平均的系统可以给出更好的结果，每一帧不需要额外的样本，这是这种方法如此吸引人的原因，甚至可以使用时间性采样来生成放大到显示器分辨率的低分辨率图像。此外，需要很多样本才能得到好结果的照明方法或其他技术，可以改为每帧使用更少的样本，因为结果将在数帧中混合。

在不增加额外采样成本的情况下为静态场景提供反走样功能时，这种类型的算法在使用时间性反走样时会遇到一些问题，如果没有对帧进行均等的加权，则静态场景中的对象可能会出现闪烁，快速移动物体或快速的摄像机移动会导致重影（ghosting），即由于先前帧的影响而在物体后方留下轨迹。重影的一个解决方案是只在缓慢移动的物体上执行这种反走样[1110]，另一个重要的方法是使用重投影（reprojection）（第12.2节）来更好地关联先前和当前帧的对象，在这种方案中，对象生成运动向量，并将其存储在单独的“速度缓冲区”（velocity buffer）中（第12.5节），这些向量用于将前一帧与当前帧相关联，即从当前像素位置减去该向量，以找到该对象表面位置前一帧的颜色像素，在当前帧中不可能成为表面一部分的样本被丢弃[1912]。由于时间性反走样不需要额外的样本，所以额外的工作相对较少，因此近年来人们对这种算法产生了浓厚的兴趣，并且该算法也得到了广泛的应用，其中一些关注是因为延迟着色技术（第20.1节）与MSAA等其他多采样支持不兼容[1486]。时间性反走样的实现方法各不相同，并且根据应用程序的内容和目标，已经开发了避免伪像和提高质量的一系列技术[836, 1154, 1405, 1533, 1938]。以Wihlidal的演讲[1885]为例，它展示了如何将EQAA，时间性反走样和应用于棋盘采样模式（checkerboard sampling pattern）的各种滤波技术结合起来，以保持画面质量，同时减少像素着色器的调用次数。Iglesias-Guitian等[796]总结了以前的工作，并提出了他们的方案，使用像素的历史信息并预测，从而最小化滤波伪像（filtering artifacts），Patney等人[1357]扩展了Karis和Lottes在虚幻4引擎的实现中[862]用于虚拟现实应用程序的TAA工作，增加了可变大小的采样以及对眼睛运动的补偿（第21.3.2节）。

**采样模式 Sampling Patterns**

有效的采样模式是减少走样、时间及其他消耗的关键要素。Naiman[1257]表明，人类最容易受到水平和垂直边缘附近走样的干扰，斜面接近45度的边缘是第二大的干扰。旋转网格超级采样（Rotated grid supersampling, RGSS）使用旋转的正方形模式来在像素内提供更多垂直和水平分辨率，图5.25显示了此模式的一个例子。

RGSS模式是一种拉丁超立方体（Latin hypercube）或N-rooks采样的形式，其中n个样本放置在n×n的网格中，每行和每列一个样本[1626]，使用RGSS时，这四个样本分别位于4×4的子像素网格中的单独行和列中，与常规2×2的采样模式相比，这种模式特别适合捕获接近水平和垂直的边缘，后者的边缘可能覆盖偶数个样本，因此有效程度较低。

N-rooks是创建良好采样模式的开始，但这还不够，比如样本可能都位于子像素网格的对角线上，所以对于几乎平行于该对角线的边缘，结果会很差，见图5.27。为了获得更好的采样，我们要避免将两个样本靠得很近，我们还希望分布均匀，将样本均匀分布在整个区域，为了形成这样的模式，我们会将如拉丁超立方体采样的分层采样（stratified sampling）技术，与其他例如抖动（jittering）、霍尔顿序列（Halton sequences）和泊松磁盘采样（Poisson disk sampling）的方法相结合[1413，1758]。

游戏机里面的人物

低可信度描述已自动生成

**图 5.27. N-rooks采样，左侧是一个符合规则的N-rooks模式，但是它在捕捉沿对角线的三角形边缘上表现较差。**

实际中，GPU制造商通常将这种采样模式硬连接到他们的硬件中，用于多重采样反走样，图5.28显示了在实践中使用的一些MSAA模式。对于时间性抗锯齿来说，覆盖模式是程序员想要的任何模式，因为样本位置可以在帧与帧之间变化，例如，Karis[862]发现基本的霍尔顿序列比GPU提供的任何MSAA模式效果都要好，霍尔顿序列在空间中产生看似随机但差异很小的样本，也就是说，它们很好地分布在空间上，没有聚集在一起[1413, 1938]。

图表, 散点图

描述已自动生成

**图 5.28. 用于AMD和NVIDIA图形加速器的MSAA采样模式，绿色的方块是阴影样本的位置，红色的方块是计算和保存的位置样本，从左到右：2×，4×，6×（AMD），和8×（NVIDIA）采样。（由D3D FSAA查看器生成）**

虽然子像素网格模式可以更好地近似每个三角形如何覆盖网格单元，但它并不理想，场景可以由屏幕上任意小的物体组成，这意味着没有采样率可以完美地捕获它们，如果这些微小的物体或特征形成图案，则以恒定间隔进行采样可能会导致莫尔条纹（Moiré fringes）和其他干涉图案，超级采样中使用的网格模式特别容易产生走样。

一种解决方案是使用随机采样（stochastic sampling），这样可以提供更加随机的模式，如图5.28所示的模式肯定可以使用。想象一下，远处有一个拥有漂亮梳齿的梳子，每个像素覆盖几根梳齿，当采样模式与梳齿频率异相时，规则模式就会产生严重的伪像，不那么有序的采样模式可以分解这些模式（译注：意译。原文Having a less ordered sampling pattern can break up these patterns.）。随机化倾向于用噪声代替重复的走样效果，因为人类视觉系统对此更宽容[1413]，这对结构较少的模式是有益的，但当像素间重复时，它仍会出现走样。一种解决方法是在每个像素上使用不同的采样模式，或者随时更改每个采样位置，在交错采样（Interleaved sampling）和索引采样（index sampling）方法中，一组像素的每个像素都有不同的采样模式（译注：原文中这一句是乱码，Interleaved samplingindexsampling!interleaved,），过去几十年中偶尔会在硬件中提供支持。例如，ATI的SMOOTHVISION允许每个像素最多有16个样本和最多16个用户定义的采样模式，这些模式可以混合在一个重复的模式中(例如，在一个4-4像素贴图中)。Molnar[1234]以及Keller和Heidrich[880]发现，对于每个像素使用相同的模式时，使用交错式随机采样可以最大程度地减少走样伪像。

值得注意的是其他一些GPU支持的算法，NVIDIA的较早的Quincunx方法[365]是一种使样本影响一个以上像素的实时反走样方案，“Quincunx”是指五个对象的排列，四个在正方形上，第五个在中心，例如在骰子（six-sided die）上五个点那一面的图案。Quincunx多重采样反走样使用此模式，将四个外部采样置于像素的角落，见图5.25，每个角落的采样值被分发给它相邻的四个像素，与其他大多数实时方案对每个样本进行平均加权不同，中心样本赋予1/2的权重，每个角落样本赋予1/8的权重，由于这种共享，每个像素平均只需要两个样本，其结果比两个样本的FSAA方法要好得多[1678]。这种模式近似于二维tent滤波器，如上一节所述，它优于box滤波器。

通过每像素使用单个样本的方法，也可以将Quincunx采样应用于时间性反走样上[836, 1677]，其中每帧在每个轴上都比之前的帧偏移半个像素，偏移方向在帧与帧之间交替，前一帧提供像素角落的点样本，并且使用双线性插值来快速计算每个像素的贡献值，将结果与当前帧取平均值。每个帧的权重相等意味着静态视图没有闪烁的伪像，对齐移动中物体的问题仍然存在，但是该方案本身很容易编写代码，并且在每帧每像素只使用一个样本的情况下会产生更好的效果。

当在单帧中使用时，Quincunx通过在像素边界共享样本而仅具有两样本的低成本消耗，RGSS模式更擅长捕捉更多的接近水平和垂直边缘的渐变。FLIPQUAD模式最初是为移动图形设备开发的，结合了这两种理想的功能[22]，它的优点是成本仅为每像素两样本，并且质量类似于RGSS（每像素四样本），这种采样模式如图5.29所示。另外，Hasselgren等人[677]探索了其他一些利用样本共享方式的廉价采样模式。

图表, 图示

描述已自动生成

**图 5.29. 左侧显示了RGSS采样模式，每像素花费四个样本，通过将这些位置移到像素边缘，可以跨边缘进行样本共享，但是要做到这样，必须使每个其他像素具有镜像的采样模式，如右图所示，所得的样本模式称为FLIPQUAD，每个像素花费两个样本。**

与Quincunx一样，双样本的FLIPQUAD模式也可以与时间性反走样一起使用，并分布在两个帧上。Drobot[382，383，1154]解决了在他关于混合重建反走样（hybrid reconstruction antialiasing, HRAA）的研究工作中哪种双样本模式最好的问题，他探索了用于时间性反走样的不同采样模式，并发现FLIPQUAD模式是所测试的五种模式中最好的。棋盘格模式也被用于时间性反走样，El Mansouri[415]讨论了使用两个样本MSAA创建棋盘渲染，以减少着色器成本，同时解决走样问题。Jimenez[836]使用SMAA，时间性反走样技术以及多种其他技术来提供一种解决方案，其中反走样质量可以根据渲染引擎负载来改变。Carpentier和Ishiyama[231]在边缘采样，将采样网格旋转了45°，他们将此时间性反走样方案与FXAA（稍后讨论）结合在一起，以在更高分辨率的显示器上进行有效的渲染。

**形态学方法 Morphological Methods**

走样通常是由边缘引起的，例如由几何形状，尖锐阴影或明亮高光形成的边缘，走样具有与边缘相关的结构，可以加以利用来提供更好的反走样结果。2009年，Reshetov[1483]沿着这些思路提出了一种算法，称其为形态学反走样（morphological antialiasing, MLAA），其中“形态”是指“与结构或形状有关”。早在1983年，Bloomenthal[170]就在这一领域做了较早的工作[830]，Reshetov的论文重新激活对了多重采样替代方法的研究，强调寻找和重建边缘[1486]。

这种反走样形式是在后处理（post-process）中执行的，也就是说以通常的方式进行渲染，然后将结果反馈到生成反走样结果的过程中去。自2009年以来，已经开发出了多种技术，那些依赖于其他缓冲区（例如深度和法线）的分析方法可以提供更好的结果，例如子像素重建反走样（subpixel reconstruction antialiasing，SRAA）[43, 829]，但仅适用于对几何边缘进行反走样。诸如几何缓冲区反走样（geometry buffer antialiasing, GBAA）和距离边缘反走样（distance-to-edge antialiasing, DEAA）之类的分析方法，会使渲染器计算有关三角形边缘位于何处的附加信息，例如边缘距像素中心的距离有多少[829]。

最通用的方案只需要颜色缓冲区，这意味着它们还可以从阴影，高光或之前应用的各种后处理技术，如轮廓边缘渲染（silhouette edge rendering）中改善边缘（见15.2.3节），例如方向局部反走样（directionally localized antialiasing, DLAA）[52, 829]是基于以下观察结果，接近垂直的边缘应水平模糊，同样接近水平的边缘也应与其相邻像素垂直模糊。

边缘检测的更复杂形式尝试寻找可能包含任意角度的边缘的像素并确定其覆盖范围，检查潜在边缘周围的邻域，目标是尽可能地重建原始边缘所在的位置，然后可以使用边缘对像素的效果来融合相邻像素的颜色，有关过程的概念视图，请参见图5.30。

卡通人物

中度可信度描述已自动生成

**图 5.30. 形态学反走样，左侧是走样图像，我们的目的是确定形成边缘的可能方向。中间表示，该算法通过检查相邻像素来记录其为边缘的可能性，给定样本后，显示了两个可能的边缘位置。右侧表示，利用最佳猜测边缘将相邻颜色按估计覆盖率成比例的混合到中心像素中。图像中的每个像素重复会这个过程。**

Iourcha等人[798]通过检查像素中的MSAA样本来计算更好的结果，从而改善了边缘查找。请注意，边缘预测和融合可以比基于样本的算法提供更高的精度，例如，一种使用每像素四样本的技术只能为对象的边缘提供以下五个混合级别：无样本覆盖，一个样本覆盖，两个样本，三个样本和四个样本。估计的边缘位置可以有更多的位置，因此提供了更好的结果。

基于图像的算法中有几种可能会误入歧途：首先，如果两个对象之间的色差低于算法的阈值，则可能无法检测到边缘；具有三个或更多不同表面重叠的像素很难进行转换；在具有高对比度或高频率元素的表面，颜色在像素间快速变化会导致算法错过边缘；特别地，当对文本应用形态学反走样时，其显示质量通常会受到影响；对象的角落部分可能是一个挑战，有些算法可能会给它们圆润的外观，假设边缘是直的，曲线也会受到其不好的影响；单个像素变化可能会导致边缘重建方式发生很大变化，从而在帧与帧之间产生明显的伪像，改善此问题的一种方法是使用MSAA覆盖遮挡（MSAA coverage masks，译注：mask在上下文中多次提到，Unity官方文档中对其有如下解释：mask不是可见的UI控件，而是一种修改控件子元素外观的方法，mask将子元素限制（即“遮挡”）为父元素的形状。其使用GPU模板缓冲区实现）来改进边缘确定性。

形态学反走样方案仅使用所提供的信息，举个例子，宽度小于像素的物体，如电线或绳索，只要它没有覆盖一个像素的中心位置，就将在屏幕上出现间隙，在这种情况下，采集更多的样本可以提高质量，但仅基于图像的反走样是不行的。此外，执行时间会根据展示的内容而变化，例如，一片草地的景色所需的反走样时间是天空的三倍[231]。

综上所述，基于图像的方法可以为较小的内存和处理成本提供反走样支持，因此它们被用于许多应用程序中，仅颜色的版本还与渲染管线分离，使其更容易修改或禁用，并且甚至可以公开为GPU的驱动程序选项。两种最流行的算法是快速近似反走样（fast approximate antialiasing, FXAA）[1079, 1080, 1084]和子像素形态反走样（subpixel morphological antialiasing, SMAA）[828, 830, 834]，部分原因是它们都为各种设备提供了可靠的（以及免费的）源代码实现，两种算法都仅使用颜色的输入，SMAA具有能够访问MSAA样本的优势，每个算法都有自己的各种可配置选项，以便在速度和质量之间进行权衡，每帧消耗通常在1-2毫秒的范围内，主要因为这是视频游戏所愿意花费的时间，最后，两种算法都可以应用时间性反走样[1812]，Jimenez[836]提出了一种改进的SMAA实现，比FXAA更快，并描述了一种时间性反走样方案。最后，我们向读者推荐Reshetov和Jimenez[1486]对形态学技术及其在视频游戏中的使用的广泛评论。

5.5 透明度，Alpha值，与合成 Transparency, Alpha, and Compositing

**半透明物体允许光通过的方法有许多，对于渲染算法而言，这些方法大致可以分为基于光线的效果及基于观察的效果，基于光线的效果指那些物体使光线衰弱或转移，导致场景中的其他物体被不同地照亮和渲染的效果，基于观察的效果指的是半透明物体本身如何被渲染的效果。**

**在本节中，我们将讨论基于视线的透明度的最简单形式，其中半透明的物体作为它后面的物体颜色的衰减器。更详细的基于视线和光线的效果，例如毛玻璃（frosted glass），光的弯曲（折射），由于透明物体的厚度导致的光线衰减，以及由于视角导致的反射率和透射率变化等，将在后面的章节中讨论。**

**一种制造透明假象的方法叫做屏幕门透明（screen-door transparency）[1244]，其思路是用像素对齐的棋盘格填充模式渲染透明三角形，也就是说三角形的像素每隔一个（every other）被渲染，从而使其后面的对象部分可见，通常情况下屏幕上的像素距离很近，以至于棋盘图案本身是不可见的。这种方法的一个主要缺点是，通常只能在屏幕的一个区域上令人信服地渲染仅一个透明对象，举个例子，如果透明的红色对象和透明的绿色对象在蓝色对象之上渲染，则三种颜色中只有两种可以出现在棋盘格图案上。此外，50%的棋盘格效果是很有限的，其他更大的像素遮掩可以用来给出其他百分比，但那往往产生可使人察觉（是透明假象）的模式。**

**之前说过，这种技术的一个优势是它比较简单，透明对象可在任意时间，以任何顺序渲染，且不需特殊的硬件支持，通过使所有对象在它们覆盖的像素处不透明，透明度问题就解决了。同样的想法也用于裁剪纹理的边缘反走样，但在子像素级别，使用了一种称为通道印射（alpha to coverage）的特性（第6.6节）。**

**由Enderton等人介绍[423]，随机透明（stochastic transparency）的方法使用子像素屏幕门遮挡与随机采样相结合而成，一个合理但充满噪声的图像，可以通过使用随机点画模式（random stipple patterns）表示片元通道印射而生成，见图5.31。为了让结果看起来合理，每个像素需要大量的样本，以及为所有子像素样本准备相当数量的内存。此方法吸引人的是不需要混合，和反走样，透明度，以及任何其他由单一机制产生的部分覆盖像素的现象。**

图示

描述已自动生成

**图 5.31. 随机透明，产生的噪声显示在放大区域中。图片来自NVIDIA SDK11 [1301]样本，由NVIDIA公司提供。）**

**大多数透明度算法会将透明对象的颜色与其后面对象的颜色混合在一起，为此，需要阿尔法混合（alpha blending）的概念[199, 387, 1429]。当一个对象在屏幕上渲染，每个像素都关联有一个RGB颜色和一个z缓冲区深度，也可以为对象覆盖的每个像素定义，另一个称为alpha(****α)的组成部分，alpha是一个值，描述给定像素的对象片段的不透明度和覆盖率。alpha为1.0意味着物体是不透明的，并且完全覆盖了对应的像素区域，为0.0表示像素完全不被遮挡，即片段是完全透明的。**

**像素的alpha可以表示不透明度或覆盖率，或同时表示两者，这取决于环境。例如，肥皂泡的边缘可能会覆盖像素的四分之三，即0.75，并且可能几乎是透明的，从而使十分之九的光线直达眼睛，所以它是十分之一不透明的，即0.1，所以它的alpha值是0.75×0.1=0.075。但是，如果我们使用MSAA或类似的反走样方案，覆盖率将通过样本自身而被考虑在内，四分之三的样本将受到肥皂泡的影响。因此在每个这样的样本中，我们将使用0.1不透明度作为alpha值。**

5.5.1 混合顺序 Blending Order

**为了使对象看起来透明，它得以小于1.0的alpha值渲染到现有场景的顶部，对象覆盖的每个像素都将从像素着色器收到一个结果****RGBα(也叫RGBA)，通常使用over操作将此片段的值与原始像素颜色混合，如下所示：**

|  |  |
| --- | --- |
|  | (5.24) |

**其中是透明对象的颜色（称作来源），是对象的alpha值，是混合前的像素颜色（称作目标），是将透明对象over在现有场景上而产生的最终颜色。在渲染管线传入和的情况下，像素的原始颜色被结果所取代。如果传入的RGBα实际上是不透明的（=1.0），则该公式简化为用对象的颜色完全替换像素的颜色。**

**示例：混合。红色的半透明对象被渲染到蓝色背景上，假设对象在某个像素处的RGB着色为（0.9**, **0.2**, **0.1），背景为（0.1**, **0.1**, **0.9），并且对象的不透明度设置为0.6，然后将这两种颜色混合**

|  |  |
| --- | --- |
|  |  |

**得到颜色为（0.58**, **0.16**, **0.42）。**

**over操作符为被渲染的对象提供半透明外观，这样做的透明度是有效的，在某种意义上，只要可以透过物体看到它背后的物体，我们就认为它是透明的[754]。使用over模拟现实世界中薄纱织物的效果时，织物后面的物体被部分遮蔽，因为织物的线是不透明的。在实际中，宽松织物的alpha覆盖率随（观察）角度变化而变化[386]，我们这里的重点是alpha模拟了有多少材质覆盖像素。**

图片包含 游戏机, 房间, 桌子

描述已自动生成

**图 5.32. 红色薄纱正方形的织物与红色的塑料滤波器，具有不同的透明效果。注意，它们的阴影也不同。（照片由Morgan McGuire提供。）**

**over操作符模拟其他透明物体时没什么说服力，尤其是模拟通过有色玻璃或塑料观察时，在现实世界中，放置在蓝色物体前面的红色滤镜通常会使蓝色物体看起来很暗，因为该物体反射的可以穿过红色滤镜光线很少，见图5.32。当使用over进行混合时，结果是红色的一部分和蓝色部分相加在一起，更好的方法应该是将这两种颜色相乘，也可以添加对透明对象本身的任意反射，第14.5.1节和第14.5.2节中讨论了这种类型的物理透射率。**

**在基本的混合阶段运算符中，over是通常用于透明效果的运算符[199，1429]，另一种有用的操作是加法混合（additive blending），即将像素值简单地求和，如下所示，**

|  |  |
| --- | --- |
|  | (5.25) |

**这种混合模式可以很好地用于发光效果，如闪电或火花，这些效果不会使对象后的像素衰减，只会使它们变得更亮[1813]。然而此模式的透明度看起来不正确，因为不透明的表面似乎没有被滤波[1192]，对于几种分层的半透明表面，例如烟或火焰之类，加法混合具有使结果的颜色更饱和的效果[1273]。**

**为了正确渲染透明对象，我们需要在不透明对象之后绘制它们，这通过首先关闭混合以渲染所有不透明对象，然后开启over以渲染透明对象。从理论上讲，我们总是可以让over开启，因为不透明的alpha为1.0会给出源颜色并隐藏目标颜色，但是这样做成本更高，没什么收益。**

**z缓冲区的限制是每个像素只能存储一个对象，如果多个透明对象重叠在同一像素，仅z缓冲区无法存储且在之后无法解决所有可见对象的影响。当使用over时，任意给定像素处的透明表面通常都需要以从后向前的顺序进行渲染，不这样做的话会给出错误的知觉暗示，一种实现这种排序的方法是，按单个对象的形心沿视线方向的距离对其进行排序，这种粗略的分类可以很好地工作，但是在各种情况下都有许多问题。首先这里的顺序只是一个近似值，因此被归类为较远的物体可能在被认为较近的物体前面；互相穿透的对象无法在每个网格的基础上对所有视角进行解析，除非把每个网格都拆成单独的部件，相关案例，请参见图5.33的左图；甚至具有凹面的单个网格，也会在屏幕上自重叠的视线方向上出现排序问题。**

图示

描述已自动生成

**图 5.33. 左侧，使用z缓冲区以透明方式渲染模型，以任意顺序渲染网格会产生严重的错误。右侧，深度剥离（depth peeling）可提供正确的外观，但要消耗额外的pass数。（图片由NVIDIA Corporation提供。）**

**尽管如此，由于其简单性与速度，以及它不需要额外的内存或特殊GPU支持，我们仍然经常使用这种对透明度进行粗糙排序的方法。如果应用这种方法，通常最好在执行透明度时关闭z深度替换功能，也就是说，z缓冲区仍被正常测试，但保留下来的表面不会改变已存储的z深度，最接近的不透明表面的深度保持不变。这样，所有透明物体至少会以某种形式出现，而不是在照相机旋转导致的排序顺序更改时突然出现或消失。其他技术也有助于改善外观表现，比如你可以把每个透明网格画两遍，首先渲染背面，然后渲染正面[1192**, **1255]。**

**over方程也可被修改，使得从前向后混合能得到相同的结果，这种混合模式称为under操作符：**

|  |  |
| --- | --- |
|  | (5.26) |

**注意，under要求目标保持alpha值，而over则不需要，换句话说，目标——在它之下混合了更近的透明表面——并不是不透明的，因此需要有alpha值，under的公式和over相似，但是交换了源和目标。另外注意，用于计算alpha值的公式与顺序无关，因为源alpha值和目标alpha值可以交换，结果都是相同的最终alpha值。**

**alpha的方程来自于将片元的alpha视为覆盖率，Porter和Duff[1429]注意到，由于我们不知道每个片元覆盖区域的形状，因此我们假设每个片元都按其alpha值比例去覆盖其他片元。例如如果，则以某种方式将像素分为两个区域，其中0.7被源片元覆盖，而另外0.3不是。在没有任何其他条件的情况下，假设目标片元覆盖率，将会被源片元按比例重叠（译注：整个源片元，而不是只有0.7的那部分），该公式具有几何解释，如图5.34所示。**

图示

描述已自动生成

**图 5.34. 一个像素和两个片元s和d，通过将两个片元沿不同轴线对齐，每个片元会按一定比例覆盖另一个，也就是说，它们是不相关的。被两个片元覆盖的面积等于under输出的alpha值，这相当于把两个区域相加，然后减去它们重叠的区域。**

5.5.2 与顺序无关的透明度 Order-Independent Transparency

under操作符用于将所有透明物体绘制到单独的颜色缓冲区中，然后使用over将这个颜色缓冲区合并到在不透明的场景之上。under操作符的另一种用途是执行称为深度剥离（depth peeling）[449, 1115]的与顺序无关的透明度（order-independent transparency，OIT）算法，顺序无关意味着应用程序不需要执行排序，深度剥离背后的思想是使用两个z缓冲区和多个pass。首先，渲染一个pass，以使所有表面的z深度（包括透明表面）都位于第一个z缓冲区中，在第二个pass中，将会渲染所有透明物体，如果一个物体的z深度与第一个z缓冲区中的值匹配，则我们知道这是最近的透明物体，然后将它的**RGBα保存到单独的颜色缓冲中。我们还通过保存超出第一个z深度并且最接近的透明对象（如果有的话）的z深度，来“剥离”该层。这个z深度是第二近的透明物体的距离，接下来的一系列pass继续使用under剥离和添加透明层，经过一定数量的pass渲染之后停下来，然后将透明图像混合在不透明图像之上，见图5.35。**

画里面的卡通人物

中度可信度描述已自动生成

**图 5.35. 每个深度剥离pass都绘制一个透明层。左侧是第一遍pass，显示了直接可见的图层。中间图显示的第二层，在每个像素处展示了距离第二近的透明表面，在这种情况下为对象的背面。右侧的第三层是一组距离第三近的透明表面。最终结果可以在图14.33中找到。（图片由Louis Bavoil提供。）**

该方案的几种变体已经研发出来了，例如Thibieroz[1763]给出了一种从后向前计算的算法，其优点是能够立即混合透明值，这意味着不需要单独的alpha channel。深度剥离的一个问题是需要知道究竟多少pass足以捕获所有透明层，一种硬件上的解决方案是提供一个像素绘制计数器，它告知渲染过程中写入了多少像素，当pass未渲染任何像素时，渲染就完成了，使用under的好处是，最重要的透明层——眼睛首先看到的那些——会在早期渲染。每个透明表面总是增加它覆盖的像素的alpha值，如果像素的alpha值接近1.0，则混合的贡献会使像素几乎不透明，所以更远的物体会有一个可以忽略不计的效果[394]。当一个pass渲染的像素数低于某个最小值，或可以指定固定数量的pass时，可以缩短从前向后的剥离过程，然而这对于从后向前的剥离效果不佳，因为距离最近（并且通常是最重要）的层是最后绘制的，因此可能会因提前终止而丧失。

尽管深度剥离是有效的，但速度可能很慢，因为每一层剥离都是针对所有透明对象的一个独立渲染pass，Bavoil和Myers[118]提出了双重深度剥离（dual depth peeling），其中每个pass剥离和保存最近最远两个深度剥离层，从而将渲染pass的数量减少一半。Liu等人[1056]探索了一种桶排序方法（bucket sort method），该方法单个pass可捕获多达32层，这种方法的一个缺点是，它需要大量内存才能为所有层保持排序顺序，通过MSAA或类似方法进行反走样将极大增加成本。

将透明物体以交互速率正确地混合在一起的难点不是我们缺乏算法，而是如何将这些算法有效地映射到GPU。1984年，Carpenter提出了A缓冲区[230]，这是多重采样的另一种形式，在A缓冲区中，渲染的每个三角形都会为其完全或部分覆盖的每个屏幕网格创建一个覆盖遮挡（coverage mask），每个像素存储所有相关片元的列表（list），不透明的片元可以清除它们后面的片元，类似于z缓冲区，对于透明表面，所有片元都被存储，一旦所有列表形成，就可以通过遍历片元和解析每个样本来产生最终结果。

在GPU上创建片元的链表的想法通过DirectX 11[611, 1765]中公开的新功能成为可能，使用的特性包括无序访问视图(unordered access views, UAVs)和原子操作（atomic operations），描述于第3.8节。基于MSAA的反走样可以通过访问覆盖遮挡和计算每个样本的像素着色器的方式来实现，该算法通过对每个透明表面进行光栅化，并将生成的片元插入长数组从而实现，连同颜色和深度一起，生成一个单独的指针结构，用于将像素的每个片元与之前存储的片元相链接，然后执行单独的pass，在此渲染屏幕填充四边形，以便在每个像素处计算像素着色器，着色器通过追踪链接来检索每个像素处的所有透明片元，检索到的每个片元都与先前的片元依次排序，然后将排序后的链表从后向前混合，从而得出最终的像素颜色，由于混合由像素着色器执行，如果需要，可以为每个像素指定不同的混合模式。随着GPU和API的不断发展，性能表现也不断通过减少原子运算符的使用消耗来提高[914]。

A缓冲区有只分配每个像素所需的片段的优点，GPU上的链表实现也是如此，从某种意义上讲，这也可能是一种缺点，因为在开始渲染帧之前所需的存储量是未知的，具有头发、烟雾或其他有可能具有许多重叠的透明表面的物体，会产生巨量的片元。Andersson[46]指出，对于复杂的游戏场景，最多可以重叠50个物体（例如树叶）的透明网格和最多200个半透明粒子。

GPU通常有预先分配好的内存资源，例如缓冲区和数组，链表方法也不例外，用户需要决定需要多少内存，而内存耗尽会导致明显的伪像。Salvi和Vaidyanathan[1532]提出了一种解决此问题的方法，多层alpha混合（multi-layer alpha blending），使用英特尔推出的称为像素同步的GPU特性，见图5.36。该功能提供可编程混合模式，且开销比原子操作要少，他们的方法重新定义了存储和混合，因此如果内存耗尽，它会适当的降低性能，粗略的排序顺序有利于他们的方案。DirectX 11.3引入了光栅化程序顺序视图（rasterizer order views）（第3.8节），这是缓冲区的一种类型，它允许在支持该功能的任何GPU上实现这个透明方法[327，328]。移动设备也具有类似的技术，称为图块本地存储（tile local storage），这允许它们实现多层alpha混合[153]，然而这种机制具有性能成本，所以这种类型的算法的消耗可能很大[1931]。

墙上挂着一幅画

低可信度描述已自动生成

**图 5.36. 在左上方，执行传统的从后到前的alpha混合，由于排序顺序不正确，导致渲染错误。在右上方，A缓冲区用于提供完美的非交互结果。左下方展示了具有多层alpha混合的渲染。右下方显示了A缓冲区和多层alpha混合之间图像的差异，乘以4以增加可见度[1532]。（图片由英特尔公司的Marco Salvi和Karthik Vaidyanathan提供。）**

这种方法（多层alpha混合）建立在Bavoil等人[115]提出的k缓冲区的概念上，其中保存了前几层可见图层并尽可能地进行了排序，而更深的图层则被丢弃并尽可能地进行了合并。Maule等人[1142]使用k缓冲区，和加权平均（weighted averaging）来计算这些更远的深层图层，加权和[1202]与加权平均[118]透明技术都是顺序无关的，且都是单pass，并且几乎可以在任意GPU上运行，但它们的问题在于没有考虑对象的顺序，例如使用alpha表示覆盖率，在淡蓝色薄纱围巾上的淡红色薄纱围巾呈现出紫罗兰色，而不是正确地看到带有一点蓝色的红色围巾。虽然对于接近不透明的对象得到的结果很差，但这类算法对于可视化是有用的，并且对高度透明的表面和粒子可以很好地工作，见图5.37。

图表, 气泡图

描述已自动生成

**图 5.37. 随着不透明度的增加，对象的顺序变得越来越重要。（图片由Dunn提供[394]。）**

在加权和透明度中，公式为

|  |  |
| --- | --- |
|  | (5.27) |

其中是透明表面的数量，和表示透明值的集合，是场景中不透明部分的颜色，渲染透明表面时将两个和累加并分别存储，并在处理透明的pass结尾处，在每个像素上计算该公式。该方法的问题在于，第一个求和是饱和的，即会得到大于（1.0, 1.0, 1.0）的颜色值，并且因为alpha的总和可能超过1.0，所以背景颜色可能会产生负面影响。

通常我们首选加权平均公式，因为它避免了这些问题：

|  |  |
| --- | --- |
|  | (5.28) |

第一行表示透明渲染过程中在两个缓冲区独自生成的结果，每个对有贡献的表面都获得一个由其alpha提供的影响权重，几乎不透明的表面提供了更多的颜色，几乎透明的表面则几乎没有影响，通过将除以，我们得到了加权的平均透明度颜色，值是所有alpha值的平均值，值是对个透明表面应用此平均alpha次数次后，目标（不透明场景）的可见性估计值，最后一行实际上是over运算符，其中代表源的alpha。

加权平均值的一个限制是对于相同的alpha，它均等地混合了所有颜色，而不论顺序如何。McGuire和Bavoil[1176, 1180]引入了加权混合（weighted blended）的顺序无关的透明方案，以提供更具说服力的结果，在他们的模拟中，到表面的距离也会影响权重，越近的表面会受到更大的影响，同时比起对alpha进行平均，是通过将项相乘在一起，并用1减去它，从而获得一组表面的真实alpha覆盖率。这种方法在视觉上产生更令人信服的结果，如图5.38所示。

图示

中度可信度描述已自动生成

**图 5.38. 查看同一引擎模型的两个不同的摄像机位置，均使用加权混合的顺序无关的透明方案进行渲染。按距离加权有助于弄清哪些表面更靠近观察者[1185]。（图片由Morgan McGuire提供。）**

该方案的一个缺点是，在一个大环境中，彼此靠近的对象的距离权重几乎是相等的，使结果与加权平均值几乎没有区别。另外，随着相机到透明物体距离的改变，深度权重实际上可能会发生变化，但是这种变化是逐渐发生的。

McGuire和Mara[1181, 1185]扩展了该方法，使其包含了合理的颜色传输效果（transmission color effect），如前所述，本节中讨论的所有透明度算法都将各种颜色混合在一起而不是对其进行滤波，从而模仿了像素覆盖率。为了得到颜色滤波器的效果，不透明场景由像素着色器读取，每个透明表面将它在场景中覆盖的像素乘以自身的颜色，并将结果保存到第三个缓冲区，这个缓冲区，其中不透明的对象现在被半透明的对象染色，在接下来解析透明缓冲区时，会用来代替不透明的场景。该方法之所以有效是因为，与透明度造成覆盖不同，颜色的传输是顺序无关的。

还有其他算法用到了这里所介绍的技术中的部分内容，例如，Wyman[1931]通过内存需求，插入和合并方法对之前的这些工作进行了分类，无论是用到的alpha值或几何覆盖率，还是如何处理丢弃的片元。通过寻找先前研究中的漏洞，他展示了两种发现的新方法，他的随机分层alpha混合方法使用了k缓冲区，加权平均值和随机透明度，另一种算法是Salvi和Vaidyanathan方法（多层alpha）的一种变体，其中使用覆盖遮挡代替alpha值。

尽管我们给出了种类繁多的透明相关内容、渲染方法和GPU功能，但是并没有完美的解决方案来渲染透明对象。我们推荐感兴趣的读者阅读Wyman的论文[1931]和Maule等人的关于交互式透明性算法的更详细的研究[1141]，McGuire的演讲[1182]提供了该领域的更广阔视野，它融会贯通了其他相关现象，例如体积照明，颜色传输和折射等，这些在本书后面会有更深入的讨论。

5.5.3 预乘Alpha与合成 Premultiplied Alphas and Compositing

**over操作符还用于将照片或对象的合成渲染混合到一起，该过程称之为合成（compositing）[199, 1662]，在这种情况下，每个像素的alpha值将与对象的RGB颜色值一起存储，由alpha channel形成的图像有时称为无光粗糙层（matte），它显示了对象的轮廓形状，相关示例参见图6.27。这个RGBα图像可以用来混合其他类似的内容或与背景做对比。**

**使用合成RGBα数据的一种方法是使用预乘alpha（premultiplied alphas，也称为关联的alpha），即在被使用之前将RGB值乘以alpha值，这使合成over公式的效率更高：**

|  |  |
| --- | --- |
|  | (5.29) |

**其中**表示**预乘的源通道，代替公式5.25中的**，预乘alpha也可以在不更改混合状态的情况下使用over和叠加混合，因为现在在混合过程中添加了源颜色[394]，请注意使用预乘的**RGBα**值，RGB分量通常不大于alpha值，尽管可以将它们创建为特别明亮的半透明值。

**合成图像的渲染很自然地与预乘alpha相吻合，默认情况下，在黑色背景上渲染的反走样不透明对象会提供预乘值，假设白色（1, 1, 1）三角形沿其边缘覆盖了某些像素的40％，使用（极精确的）反走样功能，这些像素值将设置成值为0.4的灰色，即我们将为此像素保存颜色（0.4, 0.4, 0.4），如果存储的是alpha值，那么也将为0.4，因为这是三角形覆盖的区域。最终****RGBα值为（0.4, 0.4, 0.4, 0.4），这是一个预乘值。**

**图像存储的另一种方式是使用未相乘的alpha（unmultiplied alphas），它也被称为未关联的alpha（unassociated alphas），甚至一个奇怪的术语——非预乘的alpha（nonpremultiplied alphas），未相乘的alpha就是它的字面意思：RGB值不乘以alpha值。对于白色三角形的示例，未相乘的颜色为（1, 1, 1****, 0.4），这种表示形式的优点是可以存储三角形的原始颜色，但是在显示之前，始终需要将该颜色乘以存储的alpha值。最好在执行滤波和混合操作时使用预乘数据，因为使用未相乘的alpha不能正确执行线性插值之类的操作[108, 164]，会产生诸如围绕对象边缘的黑色条纹之类的伪像[295, 648]，第6.6节的末尾有更多讨论。预乘alpha也可进行更清晰的理论处理[1662]。**

**对于图像处理应用程序来说，未关联的alpha值有利于遮挡（mask）照片而不影响基础图像的原始数据，而且未关联的alpha意味着可以使用颜色channel的全部精度范围，也就是说，必须小心地将未相乘的RGBα值正确地转换到用于计算机图形计算的线性空间，或转换回来。比如没有浏览器能正确地完成这一点，它们也不可能这样做，因为不正确的行为现在是满足预期的[649]，支持alpha的图像文件格式包括PNG（仅非关联的alpha），OpenEXR（仅关联）和TIFF（支持两种alpha类型）。**

**一个与alpha channel相关的概念是色度键控（chroma-keying）[199]，这是视频制作中的一个术语，表示演员是在绿色或蓝色屏幕上拍摄并与背景融合在一起的，在电影工业中，此过程称为绿幕（green-screening）或蓝幕（blue-screening）。这里的思路是，将特定的色调（用于胶卷）或精确值（用于计算机图形）指定为透明，当它被检测到时就会显示背景，这允许图像只使用RGB颜色来给出一个轮廓形状，无需存储任何alpha。该方案的一个缺点是，对象在任意像素处要么完全不透明，要么完全透明，即alpha实际仅为1.0或0.0，比如GIF格式允许一种颜色被指定为透明。**

5.6 显示编码 Display Encoding

**当我们计算照明效果，纹理或其他操作时，假定使用的值是线性（linear）的，非正式情况下，这意味着加法和乘法按预期执行，但是为了避免各种视觉伪像，显示缓冲区和纹理使用非线性编码，这是我们必须要考虑的。关于该问题，简短而粗略的答案如下：取着色器输出颜色范围为[0, 1]并以1/2.2作为它的幂，进行所谓的伽马校正（gamma correction），同时对传入的纹理和颜色执行相反的操作。大多数情况下，你可以告诉GPU让它为你执行这些操作，本节将快速总结这么做的方法和原因。**

**我们从阴极射线管（cathode-ray tube, CRT）开始，在数字成像的早期，CRT显示器是标准设备。这些设备在输入电压和显示辐射率之间表现出指数关系（power law），当施加于像素的能量级别增加时，发出的辐射率不会呈现线性增长，而是（令人惊讶地）以大于1的幂比例增长，例如假设指数为2，**比起被设置为1的像素，**一个设置为50%的像素会发出四分之一的光，即**。虽然LCD和其他显示技术具有与CRT不同的固有色调响应曲线（intrinsic tone response curves），但它们是通过转换电路制造的，这使它们与CRT的响应类似。

**该幂函数几乎与人类版本的亮度灵敏度相反[1431]，这种幸运巧合的结果是，编码在感知上大致是一致，也就是说，在可显示范围内，一对编码值N和N+1之间的感知差异大致是恒定的，通过测量阈值反衬（threshold contrast），我们可以在很大范围的条件下检测到大约1%的亮度差异。当颜色存储在精度有限的显示缓冲区时，这种近乎最佳的值分布最小化了条带伪像（banding artifacts），（第23.6节），同样的优点也适用于纹理，它们通常使用相同的编码。**

**显示传递函数（display transfer function），描述了显示缓冲区中的数字值与从显示器发出的辐射等级之间的关系，因为这个理由，它也被称为电光传递函数（electrical optical transfer function, EOTF）。显示传输功能是硬件的一部分，并且对于计算机显示器，电视和电影放映机有着不同的标准，还有一个标准传递函数，用于过程的另一端：图像和视频捕获设备，该函数被称为光电传递函数（optical electric transfer function, OETF）[672]。**

**在编码用于显示的线性颜色值时，我们的目标是消除显示传递函数的影响，使得无论我们计算出什么值，都会发出相应的辐射级别，比如如果计算值翻倍，则希望输出亮度翻倍，为了保持这种关系，我们使用了显示传递函数的逆函数来抵消它的非线性影响，这个使显示器响应曲线无效的过程也称为伽马校正，我们很快就会知道为什么要这样做。在解码纹理值时，我们需要应用显示传递函数来生成一个用于着色的线性值，图5.39显示了解码和编码在显示过程中的使用。**

图示

描述已自动生成

**图 5.39. 在左侧，一个PNG颜色纹理被GPU着色器访问，并且它的非线性编码值被转换（蓝色）到线性值，在着色和色调映射（tone mapping）之后（8.2.2节），最终计算值被编码（绿色）并且存储在帧缓冲区中，该值与显示传递函数决定了发射的辐射量（红色）。绿色和红色功能的组合抵消了，因此发出的辐射与线性计算值成正比。**

**个人计算机显示器的标准传递函数由被称为sRGB的颜色空间规范来定义，当从纹理中读取值或将值写入颜色缓冲区时，大多数控制GPU的API可被设置为自动应用正确的sRGB转换[491]。如6.2.2节所述，mipmap生成时也将考虑sRGB编码，通过首先转换为线性值，然后执行插值，纹理值之间的双线性插值将正确工作。alpha混合是通过将存储的值解码回线性值，混合新值，然后对结果进行编码来正确完成的。**

**在渲染的最后阶段应用这种转换很重要，此时值被写入帧缓冲区以进行显示，如果在显示编码之后应用后处理，这些结果将在非线性值上计算，这通常是不正确的，并将经常导致伪像。显示编码可以被认为是一种压缩形式，它可以最好地保留值的感知效果[491]，认知它的一个好用法是我们用线性值来执行物理计算，当我们想要显示结果或访问可显示的图像(如颜色纹理)时，我们需要使用适当的编码或解码转换，将数据移到或移自其显示编码的形式。**

**如果你确实需要手动应用sRGB，这里有标准转换公式或一些简化版本可供使用，实际上显示是由每个颜色通道的位数控制的，例如对于消费者级别的显示器，该位数为8，并提供一组在[0, 255]范围内的级别，这里我们将显示编码的级别表示为[0.0, 1.0]范围，而忽略位数，线性值也在[0.0, 1.0]范围内，表示浮点数，我们用x表示这些线性值，用y表示存储在帧缓冲区中的非线性编码值。要将线性值转换为sRGB非线性编码值，我们应用sRGB显示传递函数的逆函数：**

|  |  |
| --- | --- |
|  | (5.30) |

**其中代表线性RGB三元组通道，该公式适用于每个通道，并且这三个生成的值会用来显示。如果你手动应用转换功能，请当心，错误的一个来源是使用编码颜色而不是其线性形式，另一个来源是解码或编码颜色两次。**

**这两个变换表达式的底部是一个简单的乘法，这是因为数字硬件需要使变换完全可逆[1431]。上面的表达式涉及到一个幂，几乎应用于整个[0.0, 1.0]范围的输入值x。考虑到偏移量和比例，此函数非常近似于一个更简单的公式[491]：**

|  |  |
| --- | --- |
|  | (5.31) |

**其中**，这个希腊字母（读音伽马）是名称“伽马校正”的来源。

**就像必须对计算值进行编码以便显示一样，静态或视频摄像机捕捉到的图像必须转换成线性值才能用于计算。你在显示器或电视上看到的任何颜色都有一些显示编码的RGB三元组，可以从屏幕捕获或颜色选择器中获取它们，这些值是以PNG，JPEG和GIF等文件格式存储的，这些是可以直接发送到帧缓冲区并在屏幕上显示而不用进行转换的格式，换句话说，你在屏幕上看到的所有内容都是被定义为显示编码的数据，在着色计算中使用这些颜色之前，我们必须将这种编码形式转换回线性值，从显示编码到线性值所需的sRGB转换是**

|  |  |
| --- | --- |
|  | (5.32) |

**其中代表标准化的显示道通值，即存储在图像或帧缓冲区中的，表示为[0.0, 1.0]范围内的值，这个解码函数与我们之前的sRGB公式相反，这意味着如果纹理被着色器访问且不更改的情况下输出，它将与处理之前有相同的显示，一如预期。解码函数与显示传递函数相同，因为存储在纹理中的值已经被编码以正确显示，相比转换为给出线性响应显示，我们转换为给出线性值。**

**更简单的伽马显示传递函数是公式5.31的逆函数：**

|  |  |
| --- | --- |
|  | (5.33) |

**有时你会看到一个更简单的转换对，特别是在手机和浏览器应用上[1666]：**

|  |  |
| --- | --- |
|  | (5.34) |

**也就是说，取线性值的平方根进行转换显示，然后仅用这个值乘以它自己求逆函数，虽然这是一个粗略的近似，但这种转换也比完全忽略这个问题要好。**

徽标

描述已自动生成

**图 5.40. 两个重叠的聚光灯照亮一个平面，左侧图像中，在将亮度值0.6和0.4相加后不执行伽马校正，有效地对非线性值执行加法操作，从而导致错误，注意左侧的光看起来比右侧的光要明亮得多，且重叠部分的亮度表现出不真实的明亮。在右图中，相加后的值将进行伽马校正，灯光本身会成比例地变亮，且它们在重叠的地方正确组合。**

**如果我们不注意伽马，较低的线性值将在屏幕上显示得太暗，一个相关的错误是，如果没有伽马校正，一些颜色的色调会偏移，假设我们的**，我们想从显示的像素中发出与计算值成线性比例的辐射率，意味着我们必须将线性值的幂设为（1/2.2），线性值0.1给出0.351，0.2给出0.481，而0.5给出0.730。如果未编码，按原样使用这些值将导致显示器发出比所需更少的辐射率，请注意，任何转换都始终不会更改0.0和1.0。在使用伽马校正之前，场景建模人员通常会人为地增加深色表面的颜色，并在逆变换中折叠。

忽略伽马校正的另一个问题是，对物理线性辐射值正确的阴影计算在非线性值上执行，图5.40给出了一个实例。

图片包含 游戏机, 家具, 桌子

描述已自动生成

**图 5.41. 在左侧，黑色（显示为灰色）背景上的白色三角形边缘覆盖了四个像素，并显示了真实的区域覆盖率。如果不执行伽马校正，则中间色调的变暗会导致边缘的感知失真，如右图所示。**

忽略伽马校正也会影响反走样边缘的质量，例如，假设一个三角形的边缘覆盖了四个屏幕网格单元（图5.41），三角形的标准化辐射率为1（白色），背景为0（黑色），从左到右，单元格的覆盖率为、、、，因此如果我们使用box滤波器，我们希望将像素的归一化线性辐射度表示为0.125、0.375、0.625和0.875，正确的方法是对线性值执行反走样操作，然后将编码函数应用于四个结果值。如果不这么做，像素表示的辐射将太暗，造成如图右侧所示的边缘的感知变形，这种伪像被称为扭绳（roping），因为边缘看起来有点像扭曲的绳子[167, 1265]，图5.42展示了这种效果。

图片包含 游戏机, 梳子, 猫, 体育

描述已自动生成

**图 5.42. 在左侧，一组反走样线条经过了伽玛校正；在中间，这组线条被部分校正；在右侧，没有线条被校正。（图片由Scott R. Nelson提供。）**

sRGB标准创建于1996年，已经成为大多数电脑显示器的标准，不过从那时起，显示技术不断发展，显示器更亮，且可以显示更广泛的颜色范围已经开发出来，颜色显示和亮度将在8.1.3节中讨论，用于高动态范围显示器的显示编码在第8.2.1节中介绍，Hart的文章[672]是关于高级显示的详细信息的一个特别详尽的来源。

进一步阅读和资源 Further Reading and Resources

Pharr等[1413]更深入地讨论了采样模式和反走样，Teschner的课程笔记[1758]展示了各种采样模式的生成方法，Drobot[382，383]融会贯通了前面对实时反走样的研究，解释了各种技术的属性和性能表现，我们可以在SIGGRAPH相关课程[829]的注释中找到有关各种形态学反走样方法的信息，Reshetov和Jimenez[1486]提供了游戏中使用的形态学和相关时间性反走样工作的最新回顾。

对于透明度研究，我们再次向感兴趣的读者推荐McGuire的陈述[1182]和Wyman的工作[1931]，Blinn的文章“像素是什么？”(What Is a Pixel?)[169]在讨论不同定义的同时，提供了对计算机图形学多个领域的绝佳浏览，Blinn的书籍“脏像素”(Dirty Pixels)和“符号，符号，符号”（Notation, Notation, Notation）[166，168]包含一些介绍性文章，涉及滤波和反走样，alpha，合成和伽马校正，Jimenez的演讲[836]给出了用于反走样的最先进技术的详细信息。

Gritz和d'Eon[607]对伽马校正问题有很好的总结，Poynton的书[1431]对各类传媒中的伽马校正以及其他与颜色相关的主题提供了详尽的介绍，Selan的白皮书[1602]是更新的资料，它解释了显示编码及其在电影行业中的用途，以及许多其他相关信息。

第6章 纹理化 Texturing

“所需要的是使渲染图像看起来正确。”

——Jim Blinn

表面的纹理就是外观和感觉——只要想想油画的质感就明白了，在计算机图形学中，纹理化是获取一个表面，然后使用某些图片、函数或其他数据源在每个位置修改其外观的过程。举个例子，对于砖墙显示，一般是将砖墙的彩色图片应用在两个三角形组成的矩形上，而不是精确的显示它的几何形状，当矩形被查看时，彩色图片会显示在矩形所在的位置，除非观察者很靠近墙壁，否则缺少的几何细节是不会引起注意的。

然而，某些带纹理的砖墙可能会因为缺乏几何形状而无法令人信服，例如，如果砂浆应该是哑光的（matte），而砖块是光滑的（glossy），则观察者会注意到两种材料的粗糙度相同。为了产生更令人信服的体验，第二个图片纹理可以被应用到表面上，根据表面上的位置，这个纹理可以更改墙壁的粗糙度，而不是更改表面的颜色。现在，砖块和砂浆从图片纹理获取颜色，而从新纹理获取粗糙度。

现在观察者可能会看到所有砖块都是光滑的，而砂浆不是，每个砖块表面看起来都非常平坦，这看起来是不对的，因为砖的表面通常会有点不规则，通过应用凹凸贴图（bump mapping），砖块的着色法线就可以被改变，因此渲染时，它们看起来并不十分平滑，这种类型的纹理会摆动矩形原始表面法线的方向，以用于计算光照。

以低视角观察时，这种凹凸不平的错觉可能会被拆穿，砖块应突出在砂浆上方，以遮挡住它，即使从正面观察，砖块也应该将阴影投射到砂浆上，视差贴图（Parallax mapping）使用一张纹理让平面在渲染时似乎变形了，而视差遮蔽贴图（parallax occlusion mapping）将射线投射到高度场纹理上以提高真实感，位移贴图（Displacement mapping）通过修改形成模型的三角形高度来真正位移曲面，图6.1显示了带有颜色纹理和凹凸贴图的示例。

这些是可以通过纹理解决的问题类型的示例，使用了越来越复杂的算法。在本章中将详细介绍纹理化技术，首先介绍纹理化过程的一般框架；接下来，我们聚焦于使用图片对表面进行纹理化处理，因为这是实时工作中最流行的纹理化形式；接下来简要讨论了程序纹理，然后说明了使纹理影响表面的一些常用方法。

红色的鱼

低可信度描述已自动生成

**图 6.1. 纹理化，将颜色和凹凸贴图应用于此鱼以增加其视觉细节水平。（图片由Elinor Quittner提供。）**

6.1 纹理化管线 The Texturing Pipeline

纹理化是一种对表面材质的变化有效地建模与完善的技术，认知纹理化的一种方法是去思考单个着色像素在此过程中会发生什么。如上一章所述，我们通过考虑材质和灯光的颜色以及其他因素来计算着色，如果透明物体存在，其也能影响采样，我们通过修改着色方程中使用的值来进行纹理化，这些值的更改方式通常基于表面上的位置。因此，对于砖墙的示例，我们会基于表面位置，用砖墙图片中的相应颜色替换表面上任意点的颜色，图片纹理中的像素通常称为纹理像素（texels，简称纹素），以区别于屏幕上的像素，粗糙度纹理会修改粗糙度值，而凹凸纹理会更改着色法线的方向，因此它们每个都会改变着色方程的结果。

纹理化过程可以通过概括的纹理管线来描述，稍后我们将介绍许多术语，但请打起精神：管线的每个部分都将被详细介绍。

空间中的位置是纹理化过程的起点，该位置可以位于世界空间中，但通常位于模型的参照系中，因此随着模型的移动，纹理也随之移动，用Kershaw的术语[884]来描述，空间上此点已应用了投影函数（projector function），以便获得一组被称为纹理坐标（texture coordinates）的数值，这些数值将用于访问纹理，此过程称为映射（mapping），这引出了短语纹理映射（texture mapping），有时纹理图片本身就被称为纹理贴图（texture map），尽管严格来说这并非正确。

在这些新值可能被用于访问纹理之前，可以使用一个或多个匹配函数（corresponder function）将纹理坐标转换到纹理空间，这些纹理空间的位置用于从纹理中获取值，例如，它们可能是用于检索图像纹理中像素的数组索引，然后，这些取回的值可能会通过值转换函数（value transform function）再次进行转换，最后，这些新值将用于修改表面的某些属性，比如材质或着色法线。图6.2详细展示了这个应用单个纹理的过程，管线复杂的原因是每个步骤都为用户提供了有用的控制权，应该注意的是，并非所有步骤都需要始终保持激活。

图示

描述已自动生成

**图 6.2. 单个纹理的纹理管线概括。**

使用该管线，当具有砖墙纹理的三角形在其表面上生成一个样本时，具体发生的事情如图（见图6.3）。找到对象的局部参照系中的位置，假设是，然后将投影函数应用到此位置，正如世界地图是将三维对象投影到二维中一样，此处的投影函数通常将向量变为二维向量，此例中使用的投影函数等效于正交投影（第2.3.1节），其作用类似于幻灯片投影仪，将砖墙图片照在三角形的表面上。回到墙这边，其表面上的一个点可以被转换为范围从0到1之间的一对值，假设这里得到的值为，这些纹理坐标将用于查找图像中位于此处的颜色，假设我们的砖墙纹理分辨率为256×256，因此匹配函数会将分别乘以256，得到，舍弃小数点，在砖墙图片中找到的对应像素位置为，且它的颜色为，纹理颜色位于sRGB颜色空间中，因此如果要在着色方程式中使用该颜色，需要将其转换到线性空间，得到（第5.6节）。

图示

描述已自动生成

**图 6.3. 砖墙的管线。**

6.1.1 投影函数 The Projector Function

纹理处理的第一步是获取表面的位置并将其投影到纹理坐标空间中，通常是二维空间，建模包（Modeling packages）通常允许美术师定义每个顶点的坐标，这些可以从投影函数或网格展开算法（mesh unwrapping algorithms）来初始化，美工可以用与编辑顶点位置相同的方法来编辑坐标，投影函数通常通过将空间中的三维点转换为纹理坐标来工作，建模程序中常用的投影函数包括球形投影，圆柱形投影和平面投影[141, 884, 970]。

图片包含 游戏机

描述已自动生成

**图 6.4. 不同的纹理投影，从左到右显示了球形，圆柱形，平面和自然（u, v）投影，底部行表示每个投影应用于单个对象（其中没有自然投影）。**

其他的输入也可用于投影函数，比如可以使用表面法线来选择将六个平面投影方向中的哪个用于该表面，面部相接处的接缝会出现纹理匹配问题，Geiss[521, 522]讨论了一种将它们混合的技术，Tarini等人[1740]描述了多立方体贴图（polycube maps），其中一个模型被映射到一组立方体投影，用不同体积的空间映射到不同的立方体。

其他投影函数则根本不是投影，但它们是表面创建和细分的固有部分，例如参数化曲面具有一组自然的值集作为它们定义的一部分，见图6.4。纹理坐标也可以由各种不同的参数生成，例如视图方向，表面温度或其他任何你能想到的参数，投影函数的目标是生成纹理坐标，推导出位置的函数只是其中一种方法。

非交互式渲染器通常将这些投影函数称为渲染过程本身的一部分，单个投影函数可能就足以满足整个模型的需要，但美术人员通常必须使用工具来划分模型并分别应用各种投影函数[1345]，参见图6.5。

图示

描述已自动生成

**图 6.5. 各种纹理投影是如何被应用在单个模型上，盒式映射（Box mapping）由六个平面映射组成，每个盒子面都有一个映射。（图片由Tito Pagán提供。）**

在实时渲染中，通常在建模阶段应用投影函数，并将投影结果存储在顶点上，然而并非总是如此，有时在顶点或像素着色器中应用投影函数也是有好处的，这么做可以提高精度，并有助于启用各种效果，包括动画（第6.4节）。一些渲染方法，比如环境映射（environment mapping）（第10.4节），具有自己的专有的逐像素进行计算的投影函数。

球形投影（图6.4的左侧）将点投射到以某个点为中心的假想球体上，此投影与Blinn和Newell的环境映射方案（第10.4.1节）中使用的投影相同，因此公式10.30描述了此函数，这种投影方法也遭遇了本节所述的相同的顶点插值问题。

圆柱形投影计算的u纹理坐标与球形投影相同，而v纹理坐标作为沿圆柱轴的距离计算，此投影对于具有自然轴的对象，例如旋转表面（surfaces of revolution），是很有用的，此外当表面几乎垂直于圆柱体的轴时，会发生变形。

平面投影就像X射线一样，沿着一个方向进行平行投影，并将纹理应用于所有表面，它使用正交投影（第4.7.1节），举个例子，这种类型的投影可用于贴花（decals）（第20.2节）。

由于投影方向边缘的曲面存在严重的变形，因此美术人员通常需要手动将模型分解为近似平面的部分。有一些工具可通过展开网格，或是创建一组接近最佳的平面投影来帮助最大程度地减少变形，也可以通过其他方式帮助此过程。我们的目标是让每个多边形在纹理区域中占有更公平的份额，同时也保持尽可能多的网格连接，连接性很重要，因为采样伪像（sampling artifacts）可以沿着纹理的各个单独部分接触的边缘出现，具有良好展开效果的网格也可以简化美术人员的工作[970, 1345]，16.2.1节讨论了纹理变形如何对渲染产生不利影响，图6.6展示了用于创建图6.5中的雕像的工作空间，这种展开过程是网格参数化（mesh parameterization）这一较大研究领域的一个方面，有兴趣的读者可以参考Hormann等人的SIGGRAPH课程笔记[774]。

卡通人物

中度可信度描述已自动生成

**图 6.6. 雕像模型的几个较小的纹理，保存在两个较大的纹理中，右图显示了三角形网格如何展开并显示在纹理上以帮助其创建。（图片由Tito Pagán提供。）**

纹理坐标空间并不总是二维平面，有时它是三维体积，在这种情况下，纹理坐标表示为三元向量，其中w为沿投影方向的深度，其他系统最多使用四个坐标，通常指定为[885]，q用作齐次坐标中的第四个值，它的作用类似于电影或幻灯片的投影仪，投影纹理的大小随距离而增加。举个例子，对于将装饰性聚光图案（称为gobo）投影到舞台或其他表面上来说，它是很有用的[1597]。

纹理坐标空间的另一种重要类型是方向型，其空间中的每个点都可以通过输入方向访问，可视化这种空间的一种方法是将其作为单位球面上的点，每个点的法线表示用于访问该位置的纹理的方向，使用方向参数化的最常见纹理类型是立方体贴图（cube map）（第6.2.4节）。

另外还值得注意的是，一维纹理图像和函数也有其用途，例如，在地形模型上，可以通过其高度确定颜色，比如低地是绿色，山峰是白色，线条也可以进行纹理化，其中一种用法是将雨水渲染为一组带有半透明图像的长线，此外，这种纹理对于从一个值转换为另一值，即作为查找表，也是有用的。

由于可以将多个纹理应用于一个表面，因此可能需要定义多组纹理坐标，如果是应用坐标值的话，思路是相同的：这些纹理坐标在表面上插值并用于检索纹理值，然而在插值之前，这些纹理坐标会由匹配函数转换。

6.1.2 匹配函数 The Corresponder Function

匹配函数将纹理坐标转换为纹理空间位置，它们提供了将纹理应用于表面的灵活性，匹配函数的一个应用案例就是使用API选择现有纹理的一部分进行显示，只有该子图像才会被后续操作流程使用。

匹配函数的另一种类型是矩阵变换，可以将其应用于顶点或像素着色器，这样就可以在表面上平移，旋转，缩放，剪切或投影纹理，正如第4.1.5节所述。转换的顺序很重要，但令人惊讶的是，纹理转换的顺序必须与预期的顺序相反，这是因为纹理变换实际上影响确定被查看图像位置的空间，图像本身并不是要转换的对象，定义图像位置的空间被更改。

还有另一类匹配函数控制应用图像的方式，我们知道图像将出现在处于范围内的表面上，但是超出此范围会发生什么？匹配函数确定其具体行为。在OpenGL中，这种类型的匹配函数被称为“包装模式”（wrapping mode）；在DirectX中，它被称为“纹理寻址模式”（texture addressing mode），这种类型的常见匹配函数有：

* 包裹（wrap）（DirectX），重复（repeat）（OpenGL），或图块（tile）——图像在整个表面上重复，算法上将丢弃纹理坐标的整数部分。此函数对于使材质的图像重复并覆盖表面很有用，并且通常是默认设置。
* 镜像（mirror）——图像在整个表面上重复，但在其他每个重复图像上都被镜像（翻转），例如，图像通常从0到1正常出现，然后在1和2之间反转，然后在2和3之间正常，然后反转，依此类推。这可以让纹理边缘带有一定连续性。
* 夹取（clamp）（DirectX）或夹取到边缘（clamp to edge）（OpenGL）——超出[0, 1]范围的值都将被clamp到该范围，这会导致图像纹理的边缘不断重复，此函数对于在纹理的边缘附近采用双线性插值时，避免意外地从纹理的对边进行采样[885]很有用。
* 边框（border）（DirectX）或夹取到边框（clamp to border）（OpenGL）——[0, 1]以外的纹理坐标由单独定义的边框颜色渲染，例如，由于纹理的边缘会与边框颜色平滑融合，这种函数可以很好地将贴花渲染到单色表面上。

图片包含 游戏机

描述已自动生成

**图 6.7. 图像纹理的包裹、镜像、夹取、边框函数的效果。**

见图6.7，可以为每个纹理轴分配不同的这些匹配函数，例如纹理可以沿u轴重复同时clamp在v轴上，在DirectX中，还有一个单次镜像模式（mirror once mode），该模式沿着纹理坐标的零值镜像一次纹理，然后clamp，这对于对称贴花来说很有用。

重复平铺纹理（Repeated tiling of a texture）是向场景添加更多视觉细节的便宜方法，但是，这种技术在重复大约三遍纹理后通常看起来并不可信，因为眼睛会挑选出重复图案。避免此类周期性问题（periodicity problem）的常见解决方案是将纹理值与另一个非平铺纹理组合，这种方法可以被大大地扩展，如在Andersson[40]描述的商业地形绘制系统中所描述的那样。在该系统中，会基于地形类型，高度，坡度和其他因素组合多个纹理，纹理图像还会与场景中放置几何模型的位置相关联，例如灌木和岩石。

避免周期性的另一种选择是使用着色器程序来实现专门的匹配函数，随机地重新组合纹理图案或图块，王图块（Wang tiles）是这种方法的一个例子，一个王图块集是一小组具有匹配边缘的正方形图块，在纹理化过程中[1860]会随机选择图块，Lefebvre和Neyret[1016]使用有依赖关系的纹理读取和制表，以避图案重复，实现了相似类型的匹配函数。

应用的最后一个匹配函数边框（border）是隐式的，它是从图像的大小得出的，一个纹理通常在u和v的范围内应用，如砖墙案例所示，通过在该范围内将纹理坐标乘以图像的分辨率，可以获取像素位置。能够在范围内指定值的优点在于，可以交换具有不同分辨率的图像纹理，而不必修改存储在模型顶点处的值。

6.1.3 纹理值 Texture Values

在使用匹配函数生成纹理空间坐标之后，会使用坐标去获取纹理值。对于图像纹理，这是通过访问纹理，并从图像中检索纹素信息来完成的，此过程将在6.2节中广泛处理，图像纹理构成了实时渲染中绝大多数使用的纹理，但程序化函数（procedural functions）也是可以使用的，在程序纹理化的情况下，从纹理空间位置获取纹理值的过程不涉及存储器查找，而是函数的计算，程序纹理化将在6.3节中进一步描述。

最直接的纹理值是用于替换或修改表面颜色的RGB三元组，类似的可以返回单个灰度值，如5.5节所述，另一种可返回的数据是**RGB****α**，**α（**alpha）值通常是颜色的不透明度，决定了颜色可能影响像素的程度，也就是说，任何其他值都可以被存储，比如表面粗糙度。还有许多其他类型的数据可以存储在图像纹理中，我们将在详细讨论凹凸映射时看到这一点（第6.7节）。

从纹理返回的值可以在使用前进行转换，这些转换可以在着色器程序中执行，一个常见的例子是将数据从无符号范围（0.0到1.0）重映射到有符号范围（-1.0到1.0），用于将着色法线存储在颜色纹理中。

6.2 图像纹理化 Image Texturing

图像纹理化中，一个二维图像会有效地粘贴到一个或多个三角形的表面上，我们已经完成了计算纹理空间位置的过程，现在，我们将解决在给定位置信息后，从图像纹理获取纹理值的相关问题和算法。在本章的其余部分中，图像纹理将简称为纹理，另外，当我们提到像素单元（pixel’s cell）时，我们指的是围绕像素的屏幕网格单元，如第5.4.1节所述，像素实际上是显示的颜色值，它可能（且应该，为了更好的质量）受到与其关联的网格单元外部的样本的影响。

在本节中，我们特别关注快速采样和纹理图像的滤波方法，第5.4.2节讨论了反走样的问题，特别是在渲染对象的边缘时，纹理也可能存在采样问题，但它们发生在被渲染的三角形内部。

像素着色器通过将纹理坐标值传递给诸如texture2D之类的调用来访问纹理，这些值在纹理坐标中，由匹配函数映射到范围，GPU负责将该值转换为纹素坐标，不同API中的纹理坐标系之间有两个主要区别，在DirectX中，纹理的左上角为，右下角为，这与多种存储数据的图像类型相匹配，顶行是文件中的第一行。在OpenGL中，纹素位于左下角，是DirectX的y轴翻转。像素具有整数坐标，但是我们经常想要访问像素之间的位置，并在它们之间进行混合，这引出了一个问题：像素中心的浮点坐标是什么，Heckbert[692]讨论了为何有两种可能的系统：截断（truncating）和四舍五入（rounding）。DirectX 9将每个中心定义为——使用四舍五入，这个系统有些令人困惑，因为作为左上角像素的左上角，DirectX的原点的值为。DirectX 10则改变为OpenGL的系统，纹理元素的中心具有小数值——即截断，或者更准确地讲，是向下取整（flooring），小数部分被丢弃，向下取整是一个更自然的系统，可以很好地映射到语言，例如像素为u坐标定义了5.0到6.0的范围，为v定义了9.0到10.0的范围。

这里一个值得解释的术语是依赖纹理读取（dependent texture read），它有两个定义，第一种特别适用于移动设备，当通过texture2D或类似方法访问纹理时，当像素着色器计算纹理坐标而不是使用从顶点着色器传入的未修改纹理坐标时，就会产生依赖纹理读取[66]。请注意，这意味着对传入的纹理坐标的任何更改，甚至包括交换u和v值之类的简单操作。那些不支持OpenGL ES 3.0的较早的移动GPU，着色器没有依赖纹理读取时可以更有效地运行，因为可以预先获取纹理像素数据。该术语的另一个较老的定义对于早期的台式机GPU特别重要，当一个纹理的坐标取决于某些先前纹理的值的结果时，就会发生依赖纹理读取。例如，一种纹理可能会更改着色法线，进而改变用于访问立方体贴图的坐标，这种功能在早期的GPU上受到限制甚至不存在，如今这类读取可能会影响性能，具体取决于批处理中计算的像素数量以及其他因素。更多信息参见第23.8节。

GPU中使用的纹理图像大小通常为纹素，其中m和n为非负整数，这些被称为二次幂（power-of-two, POT）纹理，现代GPU可以处理任意大小的非二次幂（non-power-of-two, NPOT）纹理，它允许生成的图像被视为纹理，但是，某些较旧的移动GPU可能不支持NPOT纹理的mipmapping（第6.2.2节）。图形加速器对纹理大小有不同的上限，例如，DirectX 12最多允许个纹理像素。

假设我们有一个256×256纹素的纹理，我们想把它用作一个正方形的纹理，只要投影在屏幕上的正方形与纹理的大小大致相同，正方形上的纹理看起来就几乎与原始图像相同，但是，如果投影的正方形覆盖的像素是原始图像的十倍（称为放大，magnification），或者投影的正方形只覆盖屏幕的一小部分（缩小，minification），会发生什么呢？答案是，这取决于你决定对这两个不同的情况使用哪种采样和滤波方法。

本章中讨论的图像采样和滤波方法应用于从每个纹理中读取的值，但是，期望的结果是防止在最终渲染的图像中出现走样，这在理论上需要对最终像素的颜色进行采样和滤波，这里的区别是滤波着色方程的输入，还是滤波它的输出，只要输入和输出是线性相关的（对于颜色之类的输入来说就是如此），那滤波单个纹理值等价于最终颜色的滤波，然而许多存储在纹理中的着色器输入值，如表面法线和粗糙度值，与输出有非线性关系。标准的纹理滤波方法可能不能很好地处理这些纹理，导致走样，用于滤波这些纹理的改进方法将在9.13节中讨论。

6.2.1 放大 Magnification

在图6.8中，将大小为48×48纹理像素的纹理贴到一个正方形上，相对于纹理大小而言，该正方形看起来相当紧密（也就是很糊），因此基础图形系统必须放大该纹理。放大最常见的滤波技术是最近邻（实际滤波器称为box滤波器，参见第5.4.1节，nearest neighbor）和双线性插值（bilinear interpolation），还有三次卷积（cubic convolution），其使用4×4或5×5纹素阵列的加权和，这使得放大质量更高，尽管对三次卷积（也称为双三次插值，bicubic interpolation）的原生硬件支持目前并不普遍，但它可以在着色器程序中执行。

在图6.8的左侧，我们使用的是最近邻法，这种放大技术的一个特点是，单个的纹理可能会变得明显，这种效果被称为像素化，因为该方法在放大时取用离每个像素中心最近的纹理像素值，从而导致块状的外观，虽然这种方法的质量有时很差，但它只需要每个像素提取一个纹理像素值。

人不同表情的照片

描述已自动生成

**图 6.8.48×48图像放大率为320×320像素，左边：最近邻滤波，其中每像素选择最近的纹理像素。中间：使用四个最近的纹理像素的加权平均值进行双线性滤波。右边：使用5×5最近的纹理像素的加权平均值进行三次滤波。**

在中间的图像中，使用了双线性插值（有时称为线性插值），对于每个像素，这种滤波方法找到四个相邻的纹素，并在二维上进行线性插值，以找到像素的混合值，其结果更加模糊，使用最近邻法产生的锯齿也消失了很多。这里做个实验，尝试看向左图的同时眯着眼睛，能看到更多的脸部，这和低通滤波的效果差不多。

回到前述的砖块纹理示例：在不丢弃小数的情况下，我们得到了。我们在这里使用OpenGL的左下原点纹理像素坐标系，因为它与标准笛卡尔坐标系相匹配。我们的目标是在四个最接近的纹素之间插值，使用它们的纹素中心定义一个纹素大小的坐标系统，参见图6.9，为了找到四个最近的像素，我们从我们的样本位置减去像素中心分数，得到。丢弃小数部分，四个最接近的像素范围从到，在我们的示例中，小数部分是样本相对于由四个纹理像素中心形成的坐标系的位置，我们将此位置表示为。

图示, 工程绘图

描述已自动生成

**图 6.9. 双线性插值，涉及的四个纹理像素由左侧的四个正方形表示，纹理像素中心为蓝色，右边是由四个纹理像素的中心形成的坐标系。**

将纹理访问函数（texture access function）定义为，其中和为整数，并返回纹素的颜色，任何位置的双线性插值颜色的计算可分为两步。首先，底部的纹素和，将被水平插值（使用），对顶部的两个纹素和也做同样的操作。对底部纹素，我们得到（图6.9中的底部绿色圆圈），对顶部纹素，我们得到（上方的绿色圆圈），然后将这两个值垂直插值（使用），因此在处的双线性插值颜色**b**为

|  |  |
| --- | --- |
|  | (6.1) |

直观地，接近我们样本位置的纹理像素将对最终值产生更大的影响，这确实是我们在公式中看到的。在处的右上纹理像素具有对的影响，请注意对称性：右上角的影响等同于由左下角和采样点组成的矩形的面积。回到我们的示例，这意味着从该纹理像素检索的值将乘以0.42×0.74，即0.3108，从该纹理像素的顺时针方向，其他乘数为0.42×0.26，0.58×0.26和0.58×0.74，这四个权重之和为1.0。

一种解决放大倍率模糊的常见方法是使用细节纹理（detail textures），这些纹理代表了精细的表面细节，从手机上的划痕到地形上的灌木丛，这样的细节作为单独的纹理以不同的比例覆盖在放大的纹理上，细节纹理的高频重复图案与低频放大的纹理相结合，具有类似于使用单个高分辨率纹理的视觉效果。

双线性插值在两个方向上进行线性插值，但是，线性插值并不是我们想要的，假设纹理由棋盘图案中的黑白像素组成，使用双线性插值会得出整个纹理上变化的灰度样本，通过重新映射，例如，所有低于0.4的灰度都是黑色，所有高于0.6的灰度都是白色，并且中间的灰度都被拉伸以填充间隙，同时还使纹理像素之间进行一些混合，这样就可以让纹理看起来更像是棋盘，见图6.10。

图片包含 游戏机, 厨具, 过滤网

描述已自动生成

**图 6.10. 最近邻，双线性插值，以及通过重新映射介于两者之间的情况，使用相同的2×2棋盘纹理。注意由于纹理和图像网格不完全匹配，最近邻采样给出的正方形大小略有不同。**

使用更高分辨率的纹理将具有相似的效果，例如，假设每个方格由4×4像素组成，而不是1×1。在每个方格的中心周围，插值的颜色将完全是黑色或白色。

在图6.8的右侧，采用了双三次滤波器，剩余的块状现象已被大大消除，需要注意的是，双三次滤波器比双线性滤波器更昂贵，但是，许多高阶滤波器可以表示为重复线性插值（repeated linear interpolations）[1518]（见第17.1.1节），因此可以通过多次查找来利用纹理单元中用于线性插值的GPU硬件。

如果认为双三次滤波器太昂贵，Quílez[1451]提出了一种简单技术，使用平滑曲线在一组2×2纹素之间进行插值，我们首先描述曲线，然后介绍技术，两条常用的曲线是平滑阶跃曲线（smoothstep curve）和五阶曲线（quintic curve）[1372]：

|  |  |
| --- | --- |
|  | (6.2) |

当您想要平滑地从一个值插入到另一个值时，这些方法非常有用，平滑阶跃曲线具有的性质，且它在0到1之间是平滑的，五阶曲线也有同样的性质，但同时，即曲线起点和终点的二阶导数也为0，两种曲线如图6.11所示。

一些文字和图案

低可信度描述已自动生成

**图 6.11. 平滑阶跃曲线（左）和五阶曲线（右）**

该技术首先通过将样本乘以纹理尺寸并加上0.5来计算（与公式6.1和图6.9中所示相同），保留整数部分以备后来使用，并将分数存储在和中，它们在范围内，然后将变换为，仍在范围内，最后减去0.5，再将整数部分加回来，然后将得到的坐标除以纹理宽度，对也做同样的操作。此时，新的纹理坐标与GPU提供的双线性插值查找一起使用，注意，这个方法将在每个纹素处达到稳定状态，这意味着如果纹素位于RGB空间的一个平面上，那么这种插值会得到一个平滑的，但仍然呈阶梯状的外观，这可能并不总是我们所希望的，见图6.12。

图片包含 游戏机, 物体, 衣架, 钟表

描述已自动生成

**图 6.12. 放大一维纹理的四种不同方法，橙色的圆圈表示纹理的中心以及纹理的值(高度)。从左到右：最近邻，线性，以及在每对相邻纹理之间使用五阶曲线，和使用三次插值**

6.2.2 缩小 Minification

当一个纹理被缩小时，多个纹素可能会覆盖一个像素的单元格，如图6.13所示。为使每个像素获得正确的颜色值，你应该整合影响像素的所有纹素的效果，但是，很难精确地确定一个特定像素附近的所有纹素的确切影响，实际上，要完美地实时做到这一点是不可能的。

围栏里的雪地上滑雪

中度可信度描述已自动生成

**图 6.13. 缩小：通过一排像素单元，看一个棋盘纹理的正方形，大致显示若干像素如何作用于每个像素。**

由于这种限制，在GPU上使用了几种不同的方法，一种方法是使用最近邻，它的工作原理与相应的放大滤波器完全相同，即选择在像素单元中心可见的纹素，该滤波器可能会导致严重的走样问题。在图6.14中，最上面的图使用了最近邻方法，朝向地平线时会出现伪像，因为我们只选择影响像素的众多纹素中的一个来表示表面，当表面相对于观察者移动时，这样的伪像甚至更明显，这是所谓的时间走样（temporal aliasing）的一种表现形式。

图片包含 户外, 建筑, 笔记本, 屋顶

描述已自动生成

**图 6.14. 上面的图像是用点采样（最近邻）渲染的，中间是mipmapping渲染的，底部是求和区域表（summed area tables）渲染的。**

另一种常用的滤波器是双线性插值，它的工作原理也与放大滤波器完全相同，在缩小方面，这种滤波器只比最近邻方法好一点点，它混合了4个纹素而不是只用一个，但是当一个像素受到超过4个纹素的影响时，滤波器会迅速失效并产生走样。

更好的解决方案是可能的，如5.4.1节所述，走样问题可以通过采样和滤波技术来解决。纹理的信号频率取决于它的纹理在屏幕上的间距，由于奈奎斯特极限，我们需要确保纹理的信号频率不大于样本频率的一半。例如，假设图像是由交替的黑白线组成的，相隔一个纹理，然后波长为两个像素宽（从黑线到黑线），因此频率为，为了在屏幕上正确地显示这个纹理，频率必须至少为，即至少每纹素一个像素，因此，对于一般的纹理，每个像素最多应该有一个纹素，以避免走样。

为了达到这一目的，像素的采样频率必须增加或纹理频率必须降低，前一章讨论的反走样方法提供了提高像素采样率的方法，然而，这只能有限地增加采样频率，为了更充分地解决这个问题，各种纹理缩小算法已经被开发出来。

所有纹理抗锯齿算法背后的基本思想是相同的：对纹理进行预处理并创建数据结构，以帮助计算一组纹理在一个像素上的效果的快速近似值。对于实时渲染，这些算法具有使用固定的时间和资源以执行的特点，用这种方式，每个像素取一个固定数量的样本，并结合起来计算（可能是巨量的）大量纹素的效果。

**多级渐远纹理 Mipmapping**

最流行的纹理反走样方法叫做多级渐远纹理（mipmapping）[1889]，目前生产的所有图形加速器上都以某种形式实现了它，“Mip”代表拉丁文multum in parvo，意思是“小地方有很多东西”——这是一个很好的名字，用来描述一个过程，在这个过程中，原始纹理被反复滤波成更小的图像。

当使用mipmapping缩小滤波器时，在进行实际渲染之前，使用一组较小版本的纹理来扩展原始纹理。纹理（在级别零）被下采样到原始区域的四分之一，其中每个新的纹素值通常计算为原始纹理中四个相邻纹素的平均值，新的一级纹理有时被称为原始纹理的子纹理（subtexture），递归地执行缩减，直到纹理的一个或两个维度等于一个纹素，这个过程如图6.15所示，作为一个整体的图像集通常被称为一个mipmap链（mipmap chain）。

在屋顶上

低可信度描述已自动生成

**图 6.15. mipmap是通过以原始图像（第0层）作为金字塔底部，将其中的每个2×2区域平均为下一个级别的纹素值而形成的，垂直轴是第三纹理坐标d。在该图中，d不是线性的，它用于衡量一个样本用来插值的两个纹理级别。**

形成高质量mipmap的两个重要元素是良好的滤波和伽马校正，形成mipmap级别的常用方法是将每组2×2纹素取平均值，得到mip纹素值，这时使用的滤波器是一个box滤波器，它可能是最坏的滤波器之一，这可能导致较差的质量，因为它不必要的模糊低频带来的影响，同时保留一些导致走样的高频[172]，最好使用高斯、Lanczos、Kaiser或类似的滤波器。有快速、免费的源代码可用于任务[172, 1592]，并且某些API在GPU身上支持更好的滤波，在纹理的边缘附近滤波时，必须注意纹理是重复的还是单一的副本。

对于在非线性空间中编码的纹理（如大多数颜色纹理），在滤波时忽略伽马校正将修改mipmap级别的感知亮度[173, 607]，使用未校正的mipmap时，当你离物体更远，物体整体看起来会更暗，对比度和细节也会受到影响，因此将这种纹理从sRGB转换为线性空间是很重要的（第5.6节），在该空间中执行所有的mipmap滤波，并将最终结果转换回sRGB空间中存储。大多数API都支持sRGB纹理，因此可以在线性空间中正确生成mipmap并将结果存储在sRGB中，当访问sRGB纹理时，它们的值首先被转换到线性空间，以便正确地执行放大和缩小。

正如前面提到的，一些纹理与最终的着色颜色有基本的非线性关系，这给一般的滤波带来了一个问题，但mipmap生成对这个问题特别敏感，因为要滤波数百或数千个像素。通常需要专门的mipmap生成方法来获得最佳结果，这些方法在第9.13节中详细介绍。

电脑萤幕画面

描述已自动生成

**图 6.16. 左边是一个正方形像素单元格及其纹理视图，右边是像素单元格在纹理本身上的投影。**

纹理化时访问这个结构的基本过程很简单，一个屏幕像素在纹理上围出来一个区域，当像素的区域投影到纹理上时（图6.16），它包含一个或多个纹理像素，严格来说使用像素的单元边界并不是正确的，但在这里使用它以简化演示，单元格外的纹素可以影响像素的颜色，参见第5.4.1节，我们的目标是大致确定纹理对像素的影响程度。有两种常用方法来计算d（OpenGL称之为λ，它也被称为纹理细节级别（texture level of detail）），一种方法是使用像素单元形成的四边形的长边来近似像素的覆盖范围[1889]，另一种方法是用，，和这四个微分的最大绝对值作为度量[901, 1411]，每个微分是纹理坐标相对于屏幕轴的变化量的度量，例如是一个像素的u纹理值沿x屏幕轴的变化量，参见Williams的原创文章[1889]，或Flavell[473]或Pharr[1411]的文章了解更多关于这些方程的内容，McCormack等人[1160]讨论了用最大绝对值法引入走样的问题，并提出了一个替代公式，Ewins等人[454]分析了几种质量相当的算法的硬件开销。

使用Shader Model 3.0或更高版本的像素着色器程序可以使用这些梯度值，因为它们是基于相邻像素值之间的差异，它们在受动态流控制影响的像素着色器部分中不可访问（第3.8节），对于在这样的部分中（例如，在循环中）执行的纹理读取，导数必须被提前计算。注意，因为顶点着色器不能访问梯度信息，梯度或细节级别需要在顶点着色器本身中计算，并在使用顶点纹理时提供给GPU。

计算坐标d的目的是确定沿mipmap的金字塔轴采样的位置，见图6.15，我们的目标是像素与纹理的比例至少为1:1，以达到奈奎斯特率，这里的重要原则是，随着像素单元包含更多的纹素以及d的增加，将访问一个更小，更模糊的纹理版本，三元组用于访问mipmap，值d与纹理级别类似，但d不是一个整数值，而是等级之间距离的小数值，高于d位置的纹理级别和低于d位置的纹理级别被采样，位置用于从这两个纹理级别中逐个检索双线性插值的样本，然后根据从每个纹理级别到d的距离，对得到的样本进行线性插值，这整个过程被称为三线性插值（trilinear interpolation），每像素执行一次。

d坐标上的一个用户控制参数是细节偏差级别（level of detail bias，简称LOD偏差），这是一个增加到d的值，因此它影响纹理的相对感知清晰度，如果我们向上移动金字塔到顶（增加d），则纹理看起来会很模糊。对于任意给定的纹理，良好的LOD偏差将随图像类型和使用方式而变化，例如一开始就有些模糊的图像可以使用负偏差，而滤波不良（走样）的合成图像用于纹理可以使用正偏差，可以为整个纹理或像素着色器中每个像素指定偏差，用户可以提供d坐标或用于计算d的导数，以提供更好的控制。

mipmapping的好处在于，它不是试图单独地将影响一个像素的所有纹素相加，而是访问和插值了预先组合的纹素集，这个过程需要花费固定的时间，无论缩小的量是多少。然而mipmapping有几个缺陷[473]，一个主要缺陷是过度模糊（overblurring），想象一个像素单元，它在u方向上覆盖了大量的纹素，而在v方向上只覆盖了少量的纹素，这种情况通常发生在观察者几乎侧着沿纹理表面看时。实际上，有可能需要沿着纹理的一个轴缩小和沿着另一个轴放大，访问mipmap的效果是纹理上的正方形区域被检索，而不能检索矩形区域，为了避免走样，我们选择纹理上像素单元近似覆盖的最大量，这导致检索的样本通常是相对模糊的，这种效果可以在图6.14的mipmap图中看到，右边的线条向远处移动时，显示出过度模糊。

**求和区域表 Summed-Area Table**

另一种避免过度模糊的方法是求和区域（summed-area table, SAT）[312]，要使用此方法，首先创建一个与纹理大小相同但包含更高精度的颜色存储的数组（例如，红色、绿色和蓝色分别为16位或更多），在这个数组中的每个位置，必须计算所有对应纹理的纹素的总和，并存储在由该位置和纹素（原点）组成的矩形中。纹理化过程中，像素单元格在纹理上的投影被一个矩形限制，然后访问求和区域表以确定该矩形的平均颜色，该颜色作为像素的纹理颜色传回，使用矩形的纹理坐标计算平均值的过程如图6.17所示，它通过公式6.3来实现：

|  |  |
| --- | --- |
|  | (6.3) |

在这里，x和y是矩形的纹素坐标，而是该纹素的面积总和值，这个方程的原理是取从右上角到原点的整个面积和，然后通过减去相邻角的贡献来减去区域A和B，区域C已经被减去了两次，所以它被添加回左下角，注意是C区域的右上角，即是边界框的左下角。

使用求和区域表的结果如图6.14所示，接近右边缘地平线的线更加清晰，但中间对角交叉的线仍然过于模糊，问题在于，当沿着纹理的对角线观察时，会生成一个大矩形，其中许多纹素不在被计算的像素附近，例如，想象一个长而细的矩形表示像素单元的反向投影，按对角线横贯整个纹理（如图6.17所示），整个纹理矩形的平均值将被返回，而不仅仅是像素单元内的平均值。

图示

描述已自动生成

**图 6.17. 像素单元反向投影到纹理上，并以矩形为界，矩形的四个角用于访问求和区域表。**

求和区域表是所谓各向异性滤波（anisotropic filtering）算法的一个例子[691]，这样的算法在非正方形区域上检索纹理像素值，不过SAT主要在水平和垂直方向上能够最有效地做到这一点。还需要注意的是，对于大小为16×16或更小的纹理，求和区域表至少需要两倍的内存，而对于更大的纹理则需要更高的精度。

求和区域表以合理的整体内存成本提供更高的质量，它可以在现代GPU上实现[585]，改进的滤波对高级渲染技术的质量至关重要，例如Hensley等人[718, 719]提供了一种有效的实现，展示了求和区域采样如何改善光滑反射。还可以通过SAT改进其他使用区域采样的算法，例如景深（depth of field）[585, 719]、阴影贴图（shadow maps）[988]和模糊反射（blurry reflections）[718]。

**无约束各向异性滤波 Unconstrained Anisotropic Filtering**

对于当前的图形硬件，进一步改进纹理滤波的最常用方法是重用现有的mipmap硬件，其基本思想是对像素单元反向投影，纹理上的这个四边形被采样几次，然后样本被组合。如上所述，每个mipmap样本都有一个位置和一个与之关联的正方形区域，该算法不是使用一个单一的mipmap样本来近似这个四边形的覆盖范围，而是使用几个正方形来覆盖四边形。可以用四边形中较短的边来确定d（不像在mipmapping中，其通常使用较长的边），这使得每个mipmap样本的平均面积更小（因此更不模糊），四边形的较长的边被用来创建一条平行于较长的边并穿过四边形中间的各向异性线（ine of anisotropy），当各向异性的量在1:1和2:1之间时，沿这条线取两个样本（见图6.18），当各向异性比率越高时，沿轴取的样本越多。

图示

描述已自动生成

**图 6.18. 各向异性滤波，像素单元的反向投影创建一个四边形，在较长的边之间形成了一条各向异性线。**

这种方案允许各向异性线向任何方向延申，因此没有求和区域表的限制，它也不需要比mipmap更多的纹理内存，因为它使用mipmap算法来进行采样，各向异性滤波的一个例子如图6.19所示。

在屋顶上

中度可信度描述已自动生成

**图 6.19. Mipmap与各向异性滤波对比，左侧完成了三线性mipmapping，右侧进行了16：1各向异性滤波。朝向水平线方向，各向异性滤波提供了更清晰的结果，走样最小。（图片来自three.js示例webgl materials texture anisotropy [218]。）**

这种沿轴采样的想法最早是由Schilling等人提出的，同时提出的还有他们的Texram动态存储设备[1564]，Barkans描述了该算法在Talisman系统中的应用[103]，McCormack等人[1161]提出了一个名为猫（Feline）的类似系统，Texram的最初方案是对沿着各向异性轴的样本（也称为探针，probes）给予相等的权重，Talisman给予轴两端相对的两个探针一半权值，Feline使用高斯滤波内核对探针集合进行加权。这些算法接近高质量的软件采样算法，如椭圆加权平均（Elliptical Weighted Average, EWA）滤波器，其将像素的影响区域转换为纹理上的椭圆，并通过滤波器内核对椭圆内的纹素进行加权[691]，Mavridis和Papaioannou提出了几种在GPU上使用着色器代码实现EWA滤波的方法[1143]。

6.2.3 体积纹理 Volume Textures

图像纹理的直接延伸是由（或值）访问的三维图像数据，例如，可以将医学成像数据生成为三维网格，通过在网格中移动多边形，可以查看这些数据的二维切片，一个相关的想法是用这种形式表示体积灯光，一个表面上一个点的亮度是通过找到它在这个体积内的位置值，结合光的方向来得到的。

大多数GPU支持体积纹理的mipmapping，因为在体积纹理的单个mipmap级别的滤波涉及到三线性插值，所以mipmap级别之间的滤波需要四线性插值（quadrilinear interpolation）。因为这涉及到来自16个纹素的结果的平均，可能会产生精度问题，这可以通过使用更高精度的体积纹理来解决，Sigg和Hadwiger[1638]讨论了这个以及其他与体积纹理相关的问题，并提供了执行滤波和其他操作的有效方法。

尽管体积纹理对存储的要求明显更高，滤波成本也更高，但它们确实有一些独特的优势，为三维网格寻找一个好的二维参数化的复杂过程可以被跳过，因为三维位置可以直接用作纹理坐标，这避免了二维参数化通常会出现的失真和接缝问题，体积纹理也可以用来表示诸如木材或大理石等材质的体积结构，一个具有这种纹理的模型似乎是用这种材料雕刻出来的。

使用体积纹理的表面纹理化是非常低效的，因为绝大多数的样本没有使用，Benson和Davis[133]以及DeBry等人[334]讨论了用松散八叉树（sparse octree）结构存储纹理数据，这种方案非常适合交互式三维绘画系统，因为在创建时，不需要分配明确的纹理坐标给表面，而且八叉树可以将纹理细节保持在所需的任何级别，Lefebvre等人[1017]讨论了在现代GPU上实现八叉树纹理的细节，Lefebvre和Hoppe[1018]讨论了一种将稀疏的体积数据打包到一个明显更小的纹理中的方法。

6.2.4 立方体贴图 Cube Maps

另一种类型的纹理是立方体纹理（cube texture）或立方体贴图（cube map），它有六个正方形纹理，每个纹理都与立方体的一个面相关联，一个立方体贴图是通过一个三分量的纹理坐标向量来访问的，它指定了从立方体中心向外的射线的方向，射线与立方体相交点的表达式如下所示。具有最大量级的纹理坐标选择相应的面（例如，向量选择−z面），剩下的两个坐标除以最大量级坐标的绝对值，即8.4，它们现在的范围是−1到1，之后简单地重新映射到以计算纹理坐标，例如坐标映射为，立方体映射用于表示方向函数的值，它们最常用于环境映射（第10.4.3节）。

6.2.5 纹理表示 Texture Representation

在处理应用程序中的多个纹理时，有几种方法可以提高性能，纹理压缩在第6.2.6节中描述，本节的重点是纹理图集（texture atlases）、纹理数组（texture arrays）和无绑定纹理（bindless textures），所有这些都是为了避免在渲染时改变纹理造成的成本，在第19.10.1和19.10.2节中，描述了纹理流和转码。

为了能够让GPU批处理尽可能多的工作，通常最好尽可能少地改变状态（第18.4.2节），为此，可以将多个图像放入一个较大的纹理中，称为纹理图集，如图6.20左侧所示，注意，子纹理的形状可以是任意的，如图6.6所示。Nöll和Stricker[1286]描述了子纹理放置图集的优化，还需要注意mipmap的生成和访问，因为mipmap的上层可能包含几个独立的、不相关的形状，Manson和Schaefer[1119]提出了一种通过考虑表面参数化来优化mipmap的创建的方法，可以产生更好的结果，Burley和Lacewell[213]提出了一种称为Ptex的系统，其中细分曲面中的每个四边形都有自己的小纹理，这样做的好处是避免了在一个网格上分配唯一的纹理坐标，并且在纹理图集中不连接的部分的接缝上没有伪像，为了能够跨四边形进行滤波，Ptex使用了邻接（adjacency）的数据结构，虽然最初的目标是产品渲染化，Hillesland[746]提出了打包的Ptex（packed Ptex），它将每个面的子纹理放入纹理图集，并使用相邻面的填充，以避免滤波时的不连续性，Yuksel[1955]提出了网格颜色纹理（mesh color textures），它改进了Ptex，Toth[1780]通过实现一种如果滤波器抽头（tap）超出的范围则丢弃它们的方法，从而为类似ptex的系统提供了高质量的跨面滤波。

图形用户界面

中度可信度描述已自动生成

**图 6.20. 左边：一个纹理图集，其中九个较小的图像被合成为一个单一的大纹理，右边：一种更现代的方法，将较小的图像设置为纹理数组，这是大多数API中的一个概念。**

使用图集的一个困难是包装（wrapping）/重复（repeat）和镜像模式（mirror modes），这将会不正确地影响子纹理，而不只影响整个纹理，另一个问题是在为图集生成mipmap时，一个子纹理可能会渗透到另一个子纹理中，不过这可以通过为每个子纹理分别生成mipmap层次结构来避免，之后将它们放入一个大的纹理图集中，并对子纹理使用二次方分辨率[1293]。

针对这些问题的一个更简单的解决方案是使用一个叫做纹理数组的API结构，它完全避免了mipmapping和重复模式的所有问题[452]，参见图6.20的右边部分，纹理数组中的所有子纹理都需要具有相同的维度、格式、mipmap层次结构和MSAA设置，与纹理图集一样，纹理数组的设置只需要完成一次，然后可以使用着色器中的索引访问任何数组元素，这比绑定每个子纹理要快5倍[452]。

API对无绑定纹理的支持[1407]也是有助于避免状态改变成本的一个特性。如果没有无绑定的纹理，纹理将使用API绑定到特定的纹理单元，一个问题是纹理单元数量的上限，这让程序员的工作变得复杂，驱动程序确保纹理驻留在GPU端。对于无绑定的纹理，没有纹理数量的上限，因为每个纹理都由一个64位指针（有时称为句柄（handle））关联到它的数据结构，这些句柄可以通过多种不同的方式访问，例如通过统一变量，通过可变的数据，来自其他纹理或来自着色器存储缓冲对象（SSBO）。应用程序需要确保纹理驻留在GPU端，无绑定纹理避免了驱动程序中任何类型的绑定成本，这使得渲染更快。

6.2.6 纹理压缩 Texture Compression

一个直接解决内存问题、带宽问题以及缓存问题的解决方案是固定速率纹理压缩（fixed-rate texture compression）[127]。通过让GPU动态解码压缩的纹理，纹理可以只需更少的纹理内存，从而增加有效缓存大小，至少同样重要的是，这样的纹理使用起来更有效，因为它们在访问时消耗更少的内存带宽，一个相关但不同的用例是添加压缩以提供更大的纹理（译注：意思是原理相同，都是压缩可以减少带宽，但目的不同，一个是为了减少消耗、一个为了增加质量）。例如一个非压缩纹理每个纹素使用3字节，分辨率为，将占用768kB，使用纹理压缩，以6:1的压缩比，分辨率的纹理只占用512kB。

在图像文件格式中使用了多种图像压缩方法，例如JPEG和PNG，但在硬件中实现解码这些是昂贵的（参见19.10.1节关于纹理转码的信息），S3开发了一个称为S3纹理压缩（S3 Texture Compression, S3TC）的方案[1524]，该方案被选为DirectX的标准并被称为DXTC，在DirectX 10中它被称为BC（Block Compression，块压缩），此外它是OpenGL中事实上的标准，因为几乎所有GPU都支持它，它的优点是可以创建固定大小的压缩图像，具有独立的编码片段，而且解码简单（因此也很快）。图像的每个压缩部分都可以独立处理，没有共享查找表或其他依赖项，这简化了解码。

DXTC/BC压缩方案有七个变体，它们具有一些共同的属性，编码是在4×4的纹素块上完成的，也称为铺砖（tile），每个块单独编码，编码是基于插值的，对于每个编码的数量，存储两个参考值（例如颜色），一个插值因子被保存到块中16个纹素的每一中，它沿着两个参考值之间的直线选择一个值，例如一种已存储颜色或从两种已存储颜色中插入的颜色。压缩来自于每个像素通过短索引值仅存储两种颜色。

表格

描述已自动生成

**表 6.1. 纹理压缩格式，所有这些压缩块均为4×4纹素，存储列显示每个块的字节数（B）和每个纹素的位数（bpt）。参考颜色的符号首先是通道，然后是每个通道的比特数，例如RGB565表示红蓝通道5位，绿色通道6位。**

具体的编码在七个变体之间是不同的，总结在表6.1中，注意“DXT”表示DirectX 9中的名称，“BC”表示DirectX 10及以上的名称，从表中可以读到，BC1有两个16位参考RGB值（5位红色，6位绿色，5位蓝色），每个纹素都有2位的插值因子，用于从一个参考值或两个中间值中选择（作者注：另一种DXT1模式为透明像素保存四种可能的插值因子之一，将插值值的数量限制为3个——两个参考值及其平均值。），这表示与未压缩的24位RGB纹理相比，纹理压缩比为6:1；BC2以与BC1相同的方式编码颜色，但每个纹素增加4位（bpt），用于量化（原始）alpha；对于BC3，每个块具有与DXT1块相同编码方式的RGB数据，此外，alpha数据使用两个8位参考值和一个每纹素3位的插值因子进行编码，每个纹素可以选择一个参考alpha值或六个中间值中的一个；BC4只有一个通道，在BC3中编码为alpha；BC5包含两个通道，其中每个通道的编码方式与BC3相同。

BC6H用于高动态范围（HDR）纹理，其中每个纹素的每个R、G和B通道初始有的16位浮点值，这种模式使用16个字节，结果是8位，它有一种模式用于一条线路（类似于上面的技术）和另一种模式用于两条线路（译注：线参见下文），其中每个块可以从一小组分区中进行选择，还可以对两种参考颜色进行增量编码（delta-encoded），以获得更好的精度，而且根据所使用的模式，也可以具有不同的精度；在BC7中，每个区块可以有1到3条线，并存储8位，我们的目标是8位RGB和RGBA纹理的高质量纹理压缩，它的许多属性与BC6H相同，但它是LDR格式的纹理，而BC6H是HDR纹理的格式，注意BC6H和BC7在OpenGL中分别被称为BPTC FLOAT和BPTC，这些压缩技术可以应用于立方体或体积纹理，以及二维纹理。

这些压缩方案的主要缺点是它们是有损的（lossy），也就是说，原始图像通常不能从压缩版本中检索，在BC1-BC5的情况下，只使用4或8个插值值来表示16个像素，如果一个压缩纹理中有更多不同的值，就会有一些损失，在实践中，如果使用正确，这些压缩方案通常能提供可接受的图像保真度。

BC1-BC5的一个问题是用于一个块的所有颜色都位于RGB空间中的一条直线上，例如，红色、绿色和蓝色不能在单个块中表示，BC6H和BC7支持更多的线，因此可以提供更高的质量。

对于OpenGL ES，另一种称为Ericsson纹理压缩（ETC）[1714]的压缩算法被选择包含在API中，该方案具有与S3TC相同的快速解码、随机访问、无间接查找、以及固定速率等特点，它将一个由4×4纹素组成的块编码为64位，即每个纹素使用4位，其基本思想如图6.21所示。每个2×4块（或4×2，取决于哪个提供最好的质量）存储一个基本颜色，每个块还从一个小型静态查找表中选择四个常量，且块中的每个纹素可以选择添加一个在表中的值，这将修改每个像素的亮度，图像质量与DXTC相当。

图片包含 图示

描述已自动生成

**图 6.21. ETC（Ericsson纹理压缩）编码像素块的颜色，然后修改每个像素的亮度，以创建最终的纹素颜色。（图片由Jacob Ström压缩）**

在OpenGL ES 3.0中包含的ETC2[1715]中，使用从未使用过的比特组合为原始ETC算法添加更多模式，未使用的比特组合是压缩表示（例如64位），它被解压缩为与另一个压缩表示相同的图像，例如在BC1中，将两个参考色设置为相同是没有用的，因为这将指示一个恒定颜色块，而只要一个参考色包含该恒定色，就可以依次获得该色块。在ETC中，一种颜色也可以由第一种颜色加上带符号的数字增量编码，因此计算可以溢出或下溢，这种情况被用来指示其他压缩模式，ETC2添加了两种新的模式，每个块有四种颜色，以不同方式派生，一种最终模式是RGB空间中的一个平面，旨在处理平滑过渡。Ericsson alpha压缩（EAC）[1868]用一个成分（例如alpha）压缩一个图像，这种压缩类似于基本的ETC压缩，但只针对一个成分，生成的图像每个纹素存储4位，它可选择与ETC2组合，此外还可以使用两个EAC通道来压缩法线（下面将详细介绍这个主题），所有的ETC1、ETC2和EAC都是OpenGL 4.0、OpenGL ES 3.0、Vulkan和Metal的核心配置文件。

压缩法线贴图（在第6.7.2节中讨论）时需要格外小心，为RGB颜色设计的压缩格式通常不适用于普通的xyz数据，大多数方法都利用法线为单位长度这一事实，并进一步假设其z分量为正（一个对切线空间法线的合理假设），这允许只存储法线的x和y分量，z分量可以动态推导出来

|  |  |
| --- | --- |
|  | (6.4) |

这本身就导致了适度的压缩，因为只存储了两个分量，而不是三个。因为大多数GPU本身不支持三分量纹理，这也避免了浪费一个分量的可能性（或者不得不在第四个分量中包装另一个量），进一步的压缩通常通过在BC5/3Dc格式的纹理中存储x和y分量实现，见图6.22，由于每个块的参考值划分了x和y分量的最小值和最大值，它们可以被视为在xy平面上定义了包围盒，三位插值因子允许在每个轴上选择八个值，因此包围盒被划分为一个8×8的可能法线的网格，另外，可以使用EAC的两个通道（用于x和y），计算如上定义的z。

图示

描述已自动生成

**图 6.22. 左边：球体上的单位法线只需要对x和y分量进行编码。右边：对于BC4/3Dc，xy平面内的一个包围盒将法线包围起来，而盒子内每4×4块的法线可以使用8×8的法线（为了清晰起见，这里只显示4×4法线）。**

在不支持BC5/3Dc或EAC格式的硬件上，常用的备用方案[1227]是使用DXT5格式的纹理，并将两个分量存储在绿色和alpha分量中（因为它们的存储精度最高），并未使用其他两个分量。

PVRTC[465]是一种纹理压缩格式，可在Imagination Technologies的名为PowerVR的硬件上使用，它最广泛的应用是在iPhones和iPads上。它提供了每像素2位和4位的方案，并压缩4×4的纹素块，其关键思想是提供图像的两个低频（平滑）信号，这些信号是通过相邻块的纹素数据和插值得到的，然后在图像上的两个信号之间使用每个纹素1或2位的插值。

自适应可伸缩纹理压缩（Adaptive scalable texture compression, ASTC）[1302]的不同之处在于，它将一个n×m纹素的块压缩为128位，块大小从4×4到12×12不等，这导致了不同的比特率，每纹素从低至0.89位到高至8位，ASTC使用了广泛的技巧来实现紧凑的索引表示，并且每个块都可以选择行数和端点编码，此外，ASTC可以处理每个纹理1至4个通道以及LDR和HDR纹理的任何内容。ASTC是OpenGL ES 3.2及更高版本的一部分。

上面提到的所有纹理压缩方案都是有损的，当压缩一个纹理时，这个过程可花费不同的时间，可以花几秒钟甚至几分钟的时间进行压缩，以获得更高的质量，因此，这通常作为脱机预处理完成，并存储起来供以后使用；或者，可以只花几毫秒，结果质量较低，不过纹理可被接近实时的压缩并立即使用，一个例子是天空盒（第13.3节），当云可能轻微移动时，它每隔一秒左右就会重新生成一次。解压非常快，因为它是使用固定功能的硬件完成的，这种差异被称为数据压缩不对称（data compression asymmetry），即压缩所花费的时间可能比解压要长得多。

Kaplanyan[856]提出了几种可以提高压缩纹理质量的方法，对于包含颜色和法线贴图的纹理，建议使用每个分量16位的方式创建贴图。对于颜色纹理，执行直方图重归一化（histogram renormalization）（在这16位上），然后在着色器中使用缩放和偏差常数（bias constant）（逐个纹理）反转其效果，直方图归一化是一种将图像中使用的值扩展到整个范围的技术，这是实际上是一种对比度增强（contrast enhancemen）。每个分量使用16位可以确保在重新归一化后直方图中没有未使用的槽，这减少了许多纹理压缩方案可能引入的带状干扰（banding artifacts），如图6.23所示。此外，Kaplanyan建议，如果75%的像素高于116/255，则使用线性颜色空间，否则将纹理存储在sRGB空间中。对于法线映射，他还注意到BC5/3Dc经常独立于y压缩x，这意味着并不总是能找到最佳法线，所以他建议对法线使用以下误差度量（error metric）：

|  |  |
| --- | --- |
|  | (6.5) |

其中是原始法线，是被压缩后的相同法线，之后再解压缩。

需要注意的是，也可以在不同的颜色空间中压缩纹理，这可以用来加速纹理压缩，常用的转换是RGB→YCoCg[1112]：

|  |  |
| --- | --- |
|  | (6.6) |

其中是亮度项，和是色度项，反向变换也并不昂贵：

|  |  |
| --- | --- |
|  | (6.7) |

这仅相当于做了几次加法，这两个变换是线性的，可以看出公式6.6是一个矩阵-向量乘法，它本身是线性的（见公式4.1和4.2），这一点很重要，因为它可以存储YCoCg，而不是在纹理中存储RGB，纹理硬件仍然可以在YCoCg空间中执行滤波，然后像素着色器可以根据需要转换回RGB。应该注意的是，这个转换本身是有损的，这可能很重要，也可能不重要。

还有另一种可逆的RGB→YCoCg变换，概括为

|  |  |
| --- | --- |
|  | (6.8) |

其中代表右移，这意味着可以在24位RGB颜色和对应的YCoCg表示之间来回转换，而没有任何损失。需要注意的是，如果RGB中的每个分量都有n位比特，那么和都有n+1位，以保证可逆变换，只需要n位。Van Waveren和Castaño[1852]使用有损的YCoCg转换在CPU或GPU上实现了DXT5/BC3的快速压缩，它们将Y存储在alpha通道中（因为它具有最高的精度），而和存储在RGB的前两个分量中，因为Y是单独存储和压缩的，所以压缩变得很快，对于和分量，他们找到一个二维包围盒，并选择产生最佳结果的盒对角线。注意，对于在CPU上动态创建的纹理，也在CPU上压缩可能更好，当纹理在GPU上通过渲染创建时，通常最好也在GPU上压缩。YCoCg变换和其他亮度-色度变换常用于图像压缩，其中色度分量在2×2像素上平均，这样可以减少50%的存储空间，而且由于色度变化缓慢，所以通常效果很好，Lee-Steere和Harmon[1015]进一步将此转换为色相饱和度值（hue-saturation-value, HSV），在x和y上对色相和饱和度进行4倍的下采样，并将值存储为单通道的DXT1纹理，Van Waveren和Castaño也描述了法线映射压缩的快速方法[1853]。

Griffin和Olano[601]的一项研究表明，当将多个纹理应用到具有复杂阴影模型的几何模型时，纹理的质量通常很低，没有任何可感知的差异，因此，根据用例，质量的降低可能是可以接受的，Fauconneau[463]提出的DirectX 11纹理压缩格式的SIMD实现。

6.3 程序纹理化 Procedural Texturing

给定一个纹理空间位置，执行图像查找是生成纹理值的一种方法，另一种是计算一个函数，这样就定义了一个程序纹理（procedural texture）。

尽管程序纹理通常在离线渲染应用程序中使用，而图像纹理在实时渲染中更常见，这是由于现代GPU中图像纹理硬件极高的效率，其每秒可以执行数十亿次纹理访问，然而，GPU架构正在朝着更低耗的计算和（相对）更昂贵的内存访问的方向发展，这些趋势使得程序纹理在实时应用程序中得到了更好的应用。

考虑到体积图像纹理的高存储成本，对于程序纹理化，它是一个特别有吸引力的应用程序，这样的纹理可以通过多种技术合成，其中最常见的方法之一是使用一个或多个噪声函数来生成值[407, 1370, 1371, 1372]，参见图6.24。噪声函数通常以连续的二次方频率采样，称为八度（octaves），每个八度都有一个权重，通常随着频率的增加而下降，这些加权样本的总和被称为湍流（turbulence）函数。

图片包含 照片, 桌子, 不同, 食物

描述已自动生成

**图 6.24. 两个使用体积纹理的实时程序纹理示例。左边的大理石是一个半透明的体积纹理，使用光线行进（ray marching）渲染。在右边，该物体是由复杂的程序木质着色器[1054]生成的合成图像，并合成在现实环境中。（左图来自名为“玩大理石”的着色玩具，由Stéphane Guillitte提供，右图由Autodesk公司的Nicolas Savva提供。）**

由于计算噪声函数的成本高，通常预先计算三维阵列中的格子点并用于插值纹理值，有多种方法使用颜色缓冲混合快速生成这些数组[1192]。Perlin[1373]提出了一种快速、实用的方法采样该噪声函数，并展示了一些用途。Olano[1319]提供了噪声生成算法，允许在存储纹理和执行计算之间进行权衡。McEwan等人[1168]开发了在着色器中计算经典噪声和单一噪声的方法，而不需要任何查找，并提供了源代码。Parberry[1353]使用动态规划的方法将计算分摊到多个像素上，以加速噪声计算。Green[587]给出了一种更高质量的方法，但它对于接近交互的应用程序意味着更多，因为它使用50个像素的着色器指令进行一次查找，Perlin[1370, 1371, 1372]提出的原始噪声函数可以在此基础上进行改进。Cook和DeRose[290]提出了另一种表示方法，称为小波噪声，它避免了走样问题，只略微增加了评估成本。Liu等人[1054]使用各种噪声函数来模拟不同的木材纹理和表面光洁度。我们还推荐Lagae等人关于该主题的最新报告[956]。

其他程序化方法也是可行的，例如蜂窝状结构（cellular texture）是通过测量从每个位置到分散在空间中的一组“特征点”的距离而形成的，以各种方式例如改变颜色或着色法线，映射得到的最近距离，可以创建看起来像细胞、石板、蜥蜴皮肤和其他自然纹理的图案，Griffiths[602]讨论了如何在GPU上生成蜂窝状结构并有效地找到最近邻。

另一种程序纹理是物理模拟或其他互动过程的结果，如水波纹或扩展裂缝，在这种情况下，程序纹理可以有效地对动态条件产生无限的变化。

生成一个程序化的二维纹理时，参数化问题可能比编写纹理带来更多的困难，在编写的纹理中，拉伸或接缝伪像可以手动润色或处理，一种解决方案是通过直接在表面合成纹理来完全避免参数化，在复杂表面上执行这种操作在技术上具有挑战性，这是一个活跃的研究领域，参见魏等人[1861]对该领域的概述。

程序纹理的反走样比图像纹理的反走样更难，也更容易，一方面预计算方法如mipmapping等不可用，给程序员增加了负担，另一方面，程序纹理作者拥有关于纹理内容的“内部信息”，因此可以对其进行剪裁以避免走样，这尤其适用于由多个噪声函数之和创建的程序纹理，每个噪声函数的频率都是已知的，因此任何可能导致走样的频率都可以被丢弃，实际上使计算成本更低。对于其他类型的程序纹理，有多种反走样技术[407, 605, 1392, 1512]。Dorn等人[371]讨论了以前的工作，并提出了一些重构纹理函数的过程以避免高频，即带宽受限（band-limited）。

6.4 纹理动画 Texture Animation

应用到表面上的图像不一定是静态的，例如视频源可以用作一帧一帧之间变化的纹理。

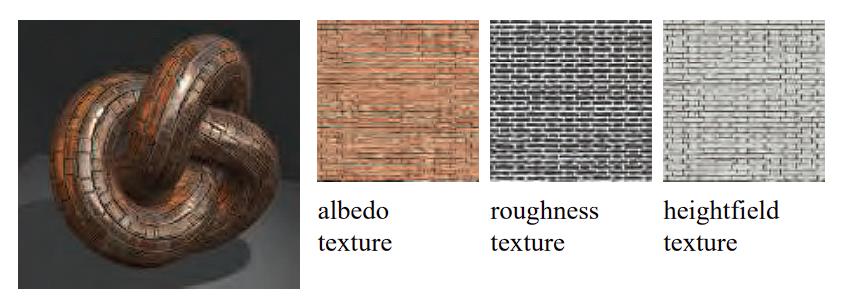
纹理坐标也不必是静态的，应用程序设计器可以显式地从逐帧修改纹理坐标，无论是在网格数据本身中，还是通过应用在顶点或像素着色器中的函数。想象一下，一个瀑布已经被建模，用一张看起来像流动中的水的图像纹理处理，假设v坐标是流动的方向，为了让水移动，必须从每一连续帧的v坐标中减去一定量，纹理坐标中的减法有使纹理本身看起来向前移动的效果。

通过将一个矩阵应用到纹理坐标上，可以创建更精细的效果。除了平移，还允许如缩放、旋转和剪切[1192, 1904]等线性变换，以及图像扭曲（image warping）和变换变形（morphing transforms）[1729]，和广义投影（generalized projections）[638]。通过在CPU或着色器中使用函数，可以创建许多更精细的效果。

通过使用纹理混合技术，可以实现其他动画效果，例如在大理石纹理上，以肉质纹理进行褪色，可以使雕像变得栩栩如生[1215]。

6.5 材质映射 Material Mapping

**纹理的一个常见用途是修改影响着色方程的材质属性，现实世界的物体通常具有随其表面变化的材质属性，为了模拟这样的对象，像素着色器可以从纹理中读取值，并使用它们在计算着色方程之前修改材质参数。纹理最常修改的参数是表面颜色，这种纹理被称为反照率颜色贴图（albedo color map）或漫反射颜色贴图（diffuse color map）。然而，任何参数都可以通过纹理进行修改：替换它、乘以它或以其他方式改变它，例如，在图6.25中，三种不同的纹理被应用到一个表面，替换常量值。**



**图 6.25. 金属砖和砂浆，右边是表面颜色、粗糙度（越浅越粗糙）和凹凸高度贴图（越浅越高）的纹理。（图片来自three.js示例webgl\_tonemapping[218]。）**

**材质的运用还可以更进一步，不仅是修改方程中的参数，纹理可以用来控制像素着色器本身的流（flow）和功能。通过让一个纹理指定表面的哪个区域有哪个材质，可以将具有不同着色方程和参数的两种或更多材质应用到一个表面上，从而为每种材质执行不同的代码。例如，有一些生锈区域的金属表面可以使用纹理来指示锈迹的位置，根据该纹理查找有条件地执行着色器的生锈部分，否则执行闪亮的金属着色器（第9.5.2节）。**

**着色模型的输入，如表面颜色，与着色器最终的颜色输出有线性关系，因此，包含这种输入的纹理可以用标准技术滤波，并避免走样，包含非线性着色输入的纹理，如粗糙度（roughness）或凹凸映射（bump mapping）（第6.7节），避免走样需要更多力气。考虑到着色方程的滤波技术可以改善这类纹理的结果，将在第9.13节讨论这些技术。**

6.6 Alpha映射 Alpha Mapping

**alpha值可以用于许多使用alpha混合或alpha测试的效果，如有效的渲染树叶、爆炸和远处的物体，仅举几例。本节讨论使用alpha纹理，注意提到的各种限制和解决方案。**

**一个与纹理相关的效果是贴花（decaling），举个例子，假设你想在茶壶上放一幅花的图片，你不想看到整个画面，只想看到花出现的部分，通过将alpha值0赋给纹素，可以使其透明，因此它没有任何效果。因此，通过适当设置贴花纹理的alpha，你可以用贴花替换或混合基础表面，通常，与透明边框一起使用的clamp对应函数将贴花的单个副本（而不是重复纹理）应用到表面。图6.26显示了如何实现贴花的一个例子，有关贴花的更多信息，请参阅第20.2节。**

图表, 图示

描述已自动生成

**图 6.26. 实现贴花的一种方法，首先用场景渲染framebuffer，然后渲染一个包围盒，对于盒内的所有点，贴花纹理被投影到framebuffer的内容中，最左边的纹素是完全透明的，因此它不会影响framebuffer，黄色的纹素是不可见的，因为它会被投影到表面的隐藏部分。**

**alpha的一个类似应用是制作剪切图（cutouts），假设你制作了一个灌木的贴花图像，并将其应用到场景中的一个矩形上，其原理与贴花相同，除了不与基础表面齐平，灌木将被绘制在其背后的所有几何图形的顶部。通过这种方式，你可以使用单个矩形渲染具有复杂轮廓的对象。**

棕榈树

描述已自动生成

**图 6.27. 左边是灌木丛纹理贴图和它下面的1位alpha通道贴图。在右边，灌木丛渲染在一个单独的矩形上，通过添加第二个旋转90度的矩形副本，我们得到了一个开销较小的三维灌木。**

**在灌木的案例中，如果你围绕它旋转观察，就会露馅，因为灌木没有厚度，一个解决方法是复制这个灌木矩形，并沿树干将其旋转90度，这两个矩形形成了一个开销较小的三维灌木，有时称其为“交叉树”[1204]，当从地面上看时，这种视觉效果是相当有效的，见图6.27。Pelzer[1367]讨论了一个类似的配置，使用三个剪切图来表示草，在第13.6节中我们讨论了一种称为billboarding的方法，用于将这种渲染减少到单个矩形。如果观察者移动到地面之上，这种视觉效果就会失效，因为从上面看灌木丛是两个剪切图，见图6.28。为了解决这个问题，可以用不同的方式添加更多的剪切图——切片、分支、层——以提供更令人信服的模型，第13.6.5节讨论了生成这种模型的一种方法，图19.31展示了另一种，参见第2页和第1049页的图片，了解最终结果的示例（译注：英文版页数）。**

卡通人物

中度可信度描述已自动生成

**图 6.28. 从离地表稍远的地方看“交叉树”灌木，此时观察角度再往上，视觉效果就失效了。**

**结合alpha映射和纹理动画可以产生令人信服的特殊效果，如闪烁的火炬，植物生长，爆炸和大气效果。**

**渲染对象时使用alpha映射有几个可选项，alpha混合（第5.5节）允许使用带小数的透明度值，这可以对物体的边缘和部分透明的对象进行反走样，然而，alpha混合要求在不透明三角形渲染后，以从后向前的顺序渲染混合三角形。一个简单的交叉树是没有正确渲染顺序的两个裁剪纹理的例子，因为每个四边形都在另一个四边形的一部分之前，即使理论上可以排序并得到正确的顺序，通常这样做效率也不高。例如，一块田地可能有成千上万的使用剪切图表示的草叶，每个网格对象可以由许多单独的叶片组成，显式地对每个叶片进行排序是非常不切实际的。**

**在渲染时，可以通过几种不同的方法来改善这个问题，其中之一是使用alpha测试，即有条件地丢弃像素着色器中alpha值低于给定阈值的片元的过程，具体做法如下**

|  |  |
| --- | --- |
|  | (6.9) |

**其中**是纹理查找得到的alpha值，参数是一个用户提供的阈值，它决定了哪些片元将被丢弃，这个二进制可见性测试允许以任意顺序渲染三角形，因为透明片元被丢弃了，我们通常希望对alpha为0.0的片元这样做，丢弃完全透明的片元还有额外的好处，可以进一步的节省着色器处理和合并的成本，以及避免错误地将z缓冲区中的像素标记为可见[394]。对于剪切图，我们通常将阈值设置为高于0.0，比如0.5或更高，然后进一步忽略alpha值，不使用它进行混合，这样做可以避免无序伪像（out-of-order artifacts），但是由于只有两个级别的透明（完全不透明和完全透明）可用，这样做的质量很低。另一种解决方案是对每个模型执行两次pass——一次用于实体剪切图，它被写入z缓冲区，另一次用于不写入z缓冲区的半透明样本。

应用程序, 日历, 地图

描述已自动生成

**图 6.29. 顶部：使用mipmapping的alpha测试，没有任何修正。底部：alpha测试，alpha值根据覆盖率缩放。（图片来自“见证者”（The Witness），由Ignacio Castaňo提供）**

alpha测试还有另外两个问题，即过度放大[1374]和过度缩小[234,557]，当alpha测试与mipmapping一起使用时，如果不进行不同的处理，效果可能无法令人信服，图6.29的顶部显示了一个例子，其中树的叶子比预期的透明度更高，这可以用一个例子来解释。假设我们有一个有四个alpha值的一维纹理，即(0.0, 1.0, 1.0, 0.0)，平均后，下一个mipmap级别变成(0.5, 0.5)，然后顶部级别是(0.5)，现在我们假设，访问mipmap级别0时，可以看到4个纹素中有1.5个纹素通过丢弃测试，但是在访问接下来两个级别时，由于将丢弃所有内容，另一个例子参见图6.30。

手机屏幕截图

中度可信度描述已自动生成

**图 6.30. 顶部是具有混合的不同mipmap级别的叶子图案，其具有更高的级别缩放可见性。底部为用0.5的alpha测试处理的mipmap，显示了物体后退时所拥有的像素减少的情况。****（图片由Ben Golus提供[557]。）**

Castaňo[234]提出了一个在mipmap创建过程中完成的简单的解决方案，效果很好，对于mipmap级别k，覆盖率定义为

|  |  |
| --- | --- |
|  | (6.10) |

其中为第k级mipmap的纹素数，为第k级mipmap在像素i处的alpha值，为公式6.9中用户提供的alpha阈值，这里我们假设的结果成立时为1，否则为0，注意表示最低的mipmap级别，即原始图像。之后对于每一个mipmap级别，我们找到一个新的mipmap阈值替代，使等于（或尽可能接近），这可以使用二分查找来实现。最后，mipmap第k级中所有纹素的alpha值按进行缩放，图6.29的底部使用了这个方法，NVIDIA的纹理工具也对此提供了支持。Golus[557]给出了一个不修改mipmap的变体，但是随着mipmap级别的增加，会在着色器中按比例放大alpha。

Wyman和McGuire[1933]提出一个不同的解决方案，其理论上将公式6.9的代码行替换为

|  |  |
| --- | --- |
|  | (6.11) |

随机函数在[0,1]中返回一个统一的值，这意味着平均而言，它会产生正确的结果，例如如果纹理查找的alpha值是0.3，片元将有30%的几率被丢弃。这是一种逐像素单个样本的随机透明度形式[423]，实际应用中将随机函数替换为哈希函数，以避免时空高频噪声：

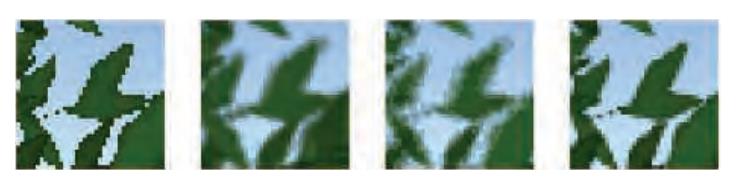
|  |  |
| --- | --- |
|  | (6.12) |

可以通过对上述函数的嵌套调用形成三维哈希，即，其返回一个之间的数字，hash的输入是对象空间坐标除以对象空间坐标的屏幕空间最大导数（x和y），随后进行clamping。需要进一步注意以获得z方向运动的稳定性，该方法最好与时间性反走样技术相结合，这种技术会随着距离的增加而减弱，因此近距离观察时我们根本不会得到任何随机效果，这种方法的优点是平均来说每个片元都是正确的，而Castaňo的方法[234]为每个mipmap级别创建了一个单独的，这个值在每个mipmap级别上可能会不同，这可能会降低质量，需要美工干预。

Alpha测试在放大情况下会出现涟漪伪像（ripple artifacts），这可以通过将Alpha映射预计算为距离场来避免[580]（参见第677页的讨论，译注：英文页）。

通道印射（Alpha to coverage），以及类似特征的透明度自适应反走样（transparency adaptive antialiasing），会取片元的透明度值，并将其转换为表示一个像素内覆盖了多少个样本的值[1250]。这个想法就像在5.5节种描述的屏幕门透明，但在子像素级别，假设每个像素有四个样本位置，且一个片元覆盖了一个像素，但由于剪切图纹理，它是25%透明的（75%不透明），alpha to coverage模式使片元变得完全不透明，但它只覆盖了四个样本中的三个，这种模式可用于重叠绿叶的剪切图纹理，例如[887,1876]。因为每个绘制的样本都是完全不透明的，所以最近的叶子会沿着边缘以一致的方式隐藏其后面的对象，不需要排序来正确地混合半透明边缘像素，因为alpha混合功能已经关闭。

Alpha to coverage对于反走样Alpha测试很好，但是在Alpha混合时会出现伪像，例如具有相同alpha覆盖百分比的两个alpha混合片元将使用相同的子像素模式，这意味着一个片元将完全覆盖另一个片元，而不是与其混合。Golus[557]讨论了如何使用fwidth()着色器指令为内容提供更清晰的边缘，见图6.31。



**图 6.31. 边缘具备部分alpha覆盖的叶片纹理的不同渲染技术，从左到右：alpha测试，alpha混合，alpha to coverage，具备锐化边缘的alpha to coverage。（图片由Ben Golus提供[557]。）**

对于任何alpha映射的使用，了解双线性插值如何影响颜色值是很重要的，想象两个彼此相邻的纹素：是纯红色的，而它的邻居是黑色的，几乎完全透明，两个纹素正中间位置的是多少？简单插值得到，得到的rgb值是一个暗红色，然而实际上结果并没有变暗，它是一个纯红色，已经预乘以它的alpha。如果你插值alpha值，为了正确的插值，你需要确保被插值的颜色在插值之前已经预乘以alpha。例如假设几乎透明的邻居被设置为，即淡淡的绿色，这个颜色没有预先乘上alpha，在插值时将得到结果——微小的绿色突然转变为（预乘的）黄色样本。这个相邻纹素的预乘版本是，它给出了正确的预乘结果，这个结果更有意义，最终的预乘颜色主要是红色，并带有一点难以察觉的绿色。

未使用预乘结果的双线性插值会导致贴花和剪切图对象边缘出现黑边，暗红色结果会被管线的其余部分视为未乘色，所以边缘变成黑色，即使使用alpha测试，也会看到这种效果。最佳策略是在双线性插值之前进行预乘法[490, 648, 1166, 1813]，WebGL的API支持这一点，因为合成（compositing）对网页很重要，然而双线性插值通常是由GPU执行的，在执行此操作之前，着色器不能对纹素值进行操作。PNG等文件格式的图像不会预乘，因为这样做会失去颜色精度，使用alpha映射时，上述两个因素结合在一起会导致默认情况下的黑色边缘。一种常见的解决方法是对剪切图图像进行预处理，用从周围不透明纹理派生的颜色为透明的“黑色”纹素着色[490, 685]，所有透明区域通常都需要以这种方式重新绘制，手工或自动，这样mipmap级别也可以避免边缘问题[295]。同样值得注意的是，当形成带有alpha值的mipmap时，应该使用预乘值[1933]。

6.7 凹凸映射 Bump Mapping

本节介绍一大类我们统称为凹凸映射（bump mapping）的小尺度细节表示技术，所有这些方法通常都是通过修改逐像素着色流程来实现的，它们提供了比纹理映射更立体的外观，且没有添加任何额外的几何图形。

物体的细节可以分为三个尺度：覆盖许多像素的宏观特征（macro-features），几个像素之间的细观特征（meso-features），以及基本在一个像素内的微观特征（micro-features），这些分类不是固定的，因为观察者可能在动画或交互会话中的多个距离中观察到相同对象。

宏观几何形态（Macrogeometry）由顶点、三角形或其他几何图元（primitive）表示，在创建三维角色时，四肢和头部通常是在宏观尺度上建模的。微观几何形态（Microgeometry）被封装在着色模型中，它通常在像素着色器中实现，并使用纹理映射作为参数，使用的着色模型模拟了表面微观几何结构的相互作用，例如光泽物体的微观是光滑的，而漫反射的表面微观是粗糙的。角色的皮肤和衣服似乎有不同的材质，因为它们使用不同的着色器，或者至少在这些着色器中使用不同的参数。

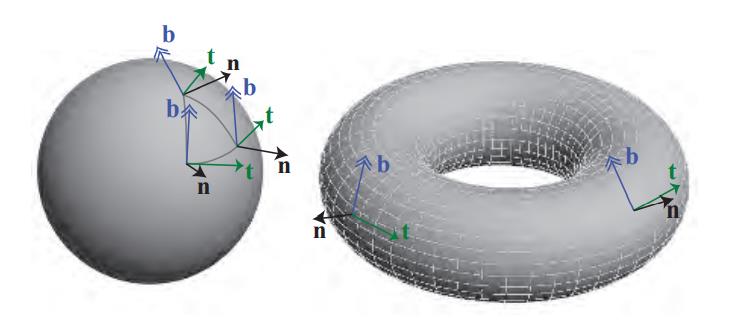
细观几何描述了这两个尺度之间的一切，它包含的细节过于复杂，无法有效地使用单个三角形渲染，但它足够大，可以让观察者在几个像素内区分表面曲率的单个变化，角色脸上的皱纹、肌肉组织细节、他们衣服的褶皱和接缝，都是中尺度的。一组统称为凹凸映射技术的方法常用于中尺度建模，这些像素级别的着色参数调整，使得观察者感知到远离基本几何的微小扰动，尽管其实际上保持平坦，不同类型的凹凸映射之间的主要区别是它们如何表示细节特征，变量包括细节特征的逼真程度和复杂程度。例如数字艺术家通常在模型中雕刻细节，然后使用软件将这些几何元素转换为一个或多个纹理，例如凹凸纹理和可能的裂缝暗化（crevice-darkening）纹理。

Blinn于1978年提出了在纹理中编码细观尺度细节的想法[160]，他观察到，如果在着色过程中，我们用一个轻微扰动的表面法线来代替真实的表面，那么表面似乎具有小尺度的细节，他将描述表面法线扰动的数据存储在阵列中。

关键思想是，我们不是使用纹理来改变光照方程中的颜色成分，而是访问纹理来修改表面法线，表面的几何法线保持不变，我们只是修改了用在照明方程中的法线，这个操作没有物理等价，我们在表面法线上进行改变，但表面本身在几何意义上保持光滑。正如每个顶点都拥有法线会给人一种三角形之间表面平滑的错觉，修改每个像素的法线会改变三角形表面本身的视觉感知，而不需要修改它的几何形状。

对于凹凸映射，法线必须相对于某些参考系改变方向，为此每个顶点上将存储一个切线坐标系（tangent frame），也称为切线空间基（tangent-space basis），这个参考系用来将光照转换到表面位置空间（surface location’s space）或反之，以计算干扰法线的效果。对于一个应用了法线映射的多边形曲面，除了顶点法线外，我们还存储了切线（tangent）向量和副切线（bitangent）向量，副切线向量也被错误地称为副法线向量（binormal vector）[1025]。

切线和副切线向量表示物体空间中的法线贴图本身的轴，因为目标是将光照转换为相对于贴图的光，参见图6.32。



**图 6.32. 图示是一个球面三角形，它的切线坐标显示在每个角上，如圆球和圆环的形状具有自然的切向空间基，正如圆环上的经纬线所示。**

这三个向量，法线，切线，和副切线，形成一个基础矩阵：

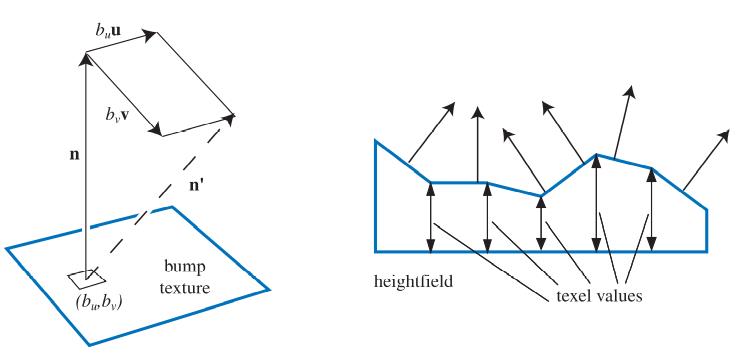
|  |  |
| --- | --- |
|  | (6.13) |

这个矩阵有时缩写为TBN，它将光的方向（对于给定顶点）从世界空间转换到切线空间，这些向量不需要真正地相互垂直，因为法线映射本身可能被扭曲以适应曲面，然而非正交基引入了纹理的倾斜，这可能意味着需要更多的存储空间，也可能有性能影响，即矩阵不能通过简单的转置来求逆[494]。节省内存的一种方法是只在顶点上存储切线和副切线，然后取它们的叉积来计算法向量，不过这种方法只有在矩阵的手性总是相同的情况下才有效[1226]。通常一个模型是对称的：一架飞机、一个人、一个文件柜和许多其他的对象，因为纹理消耗大量的内存，且它们经常被镜像到对称模型上，因此可以只存储对象纹理的一半，而纹理映射将它放置于整个模型，在这种情况下，对称两边的切线空间手性不同，不能被假设，不过如果在每个顶点上存储一个额外的比特信息来表示手性，仍有可能避免存储法线，如果这样做，该比特信息用于对切线和副切线的叉积取反，以生成正确的法线。如果切线坐标系是正交的，也可以将坐标基存储为四元数（第4.3节），这样既节省空间，又可以节省每像素的计算量[494, 1114, 1154, 1381, 1639]，质量上有轻微的损失是可能的，尽管在实践中很少见到。

切线空间的概念对于其他算法来说也很重要，正如下一章所讨论的，许多着色方程只依赖表面法线的方向，然而像拉丝铝或天鹅绒这样的材质，也需要知道观察者的相对方向和光照相对于表面的方向。切线坐标系有助于确定材质在表面上的方向，Lengyel[1025]和Mittring[1226]的文章对这一领域进行了广泛的报道。Schüler[1584]提出了一种在像素着色器中动态计算切线空间基的方法，且不需要存储每个顶点预先计算的切线坐标，Mikkelsen[1209]改进了该技术，推导出了一种不需要任何参数化，而是使用表面位置的导数和高度场的导数来计算扰动法线的方法，但是与使用标准的切线空间映射相比，这种技术可能导致显示的细节大大减少，也可能产生美术工作流程问题[1639]。

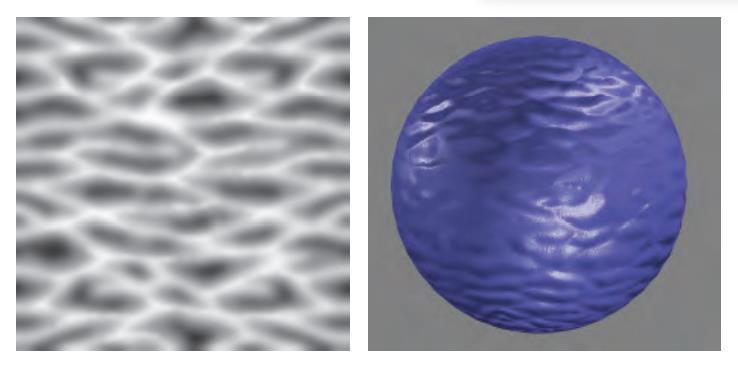
6.7.1 布林的方法 Blinn’s Methods

Blinn最初的凹凸映射方法在纹理的每个纹素中存储了两个带符号的值，和，这两个值对应于法线沿和图像轴方向的变化量，也就是说，这些纹理值通常是双线性插值的，用于缩放两个垂直于法线的向量，这两个向量加到法线上改变法线的方向，和两个值描述了曲面在该点的方向，见图6.33。这种类型的凹凸贴图纹理称为偏移矢量凹凸贴图（偏移矢量凹凸贴图）或偏移贴图（offset map）。



**图 6.33. 在左边，法向量在和方向上被从凹凸纹理中提取出的值修改，得到（未归一化）。在右边，显示了高度场和它对着色法线的影响，这些法线可以在高度之间插值，以获得更平滑的外观。**

另一种表示凹凸的方法是使用高度场（heightfield）来修改表面法线的方向，每个单色纹理值代表一个高度，所以在纹理中白色是高区域，黑色是低区域（反之亦然），参见图6.34中的示例。这是首次创建或扫描凹凸映射时常用的格式，也是Blinn在1978年引入的，高度场用于推导u和v的符号值，类似于第一种方法中使用的值，通过取相邻列之间的差来得到u的斜率，以及相邻行之间的差来得到v的斜率[1567]。一种变体是使用索贝尔（Sobel）滤波器，它赋予直接相邻的纹素更大的权重[535]。

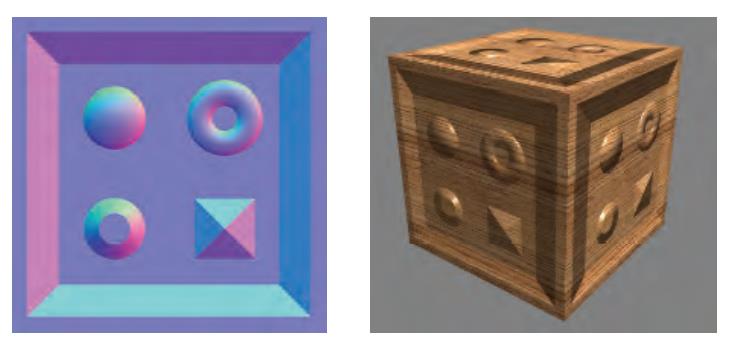


**图 6.34. 一种波浪形高度场凹凸图像及其在球体上的应用。**

6.7.2 法线映射 Normal Mapping

凹凸映射的一种常用方法是直接存储法线映射，算法和结果在数学上与Blinn的方法一致，只有存储格式和像素着色器的计算会改变。

法线映射编码映射到，例如对于8位纹理，x轴值0表示−1.0，而255表示1.0。一个例子显示在图6.35，颜色是浅蓝色，表示所示颜色映射的平滑表面，即法线。



**图 6.35. 使用法线映射的凹凸映射，每个颜色通道实际上都是一个表面法线坐标，红色通道是x方向的偏差，红色越多，法线越向右。绿色是y方向的偏差，蓝色是z方向的偏差。右边是使用法线贴图生成的图像，注意立方体顶部的扁平外观。（图片由Manuel M. Oliveira和Fabio Policarpo提供。）**

法线贴图的表示最初是作为世界空间法线贴图引入的[274, 891]，在实践中很少使用，对于这种类型的映射，扰动很简单：在每个像素处，从贴图中检索法线并直接使用它，连同光照的方向，计算表面上该位置的着色。法线映射也可以在对象空间中定义，这样模型就可以旋转，且法线仍然有效，然而世界空间和对象空间的表示都将纹理绑定到特定方向的特定几何图形上，这限制了纹理的重用。

我们通常在相对于表面本身的切线空间中检索扰动的法线，这允许表面的变形，以及正常纹理的大量重用，切线空间法线贴图也可以很好地压缩，因为z分量（与未受干扰的表面法线对齐）的符号通常可以被假设为正。

法线映射可以很好地增加真实感，见图6.36。



**图 6.36. 一个类似游戏的场景中使用法线贴图进行凹凸映射的例子，左上：不使用右边的两个法线贴图，左下：应用了法线贴图，右图：法线贴图。（3D模型和法线贴图由Dulce Isis Segarra López提供。）**

与滤波颜色纹理相比，法线贴图的滤波是一个难题，一般来说法线和着色颜色之间的关系不是线性的，所以标准的滤波方法可能会导致讨厌的走样。想象一下，你正在看由一片片闪亮的白色大理石砌成的楼梯，在某些角度，楼梯的顶部或侧面捕捉到光线，并反射出明亮的镜面高光，然而楼梯的平均法线是，比方说，45度角，它将从与原始楼梯完全不同的方向捕捉高光，当带有尖锐镜面高光的凹凸贴图在没有正确滤波的情况下渲染时，高光会根据样本所处的位置闪烁，从而产生分散注意力的闪烁效果。

Lambertian表面是一种特殊情况，其中法线贴图对着色几乎有线性影响，Lambertian着色几乎完全是一个点积，它是线性运算，求一组法线的平均值并对结果进行点积，等效于求单个法线的点积结果的平均值：

|  |  |
| --- | --- |
|  | (6.14) |

注意平均向量在使用前没有被归一化，公式6.14表明，标准滤波和mipmap对Lambertian表面几乎产生了正确的结果，但结果并不完全正确，因为Lambertian着色方程不是一个点积，它是一个clamp的点积——，clamp操作使它成为非线性的，这使得对于掠光方向，表面过于暗淡，但在实践中这通常是能接受的[891]。需要注意的是，一些通常用于法线贴图的纹理压缩方法（例如利用其他两个分量重构z分量）不支持非单位长度的法线，因此使用未归一化的法线贴图可能会带来压缩困难。

在非Lambertian表面的情况下，比起单独滤波法线贴图，将着色方程的输入作为一组数据进行滤波，可以产生更好的结果，该方法将在9.13节中讨论。

最后，从高度贴图派生法线贴图可能是有用的，具体做法如下[405]。首先，使用中心差计算在x方向和y方向的导数的近似值：

|  |  |
| --- | --- |
|  | (6.15) |

纹素处的非归一化法线为

|  |  |
| --- | --- |
|  | (6.16) |

这里必须注意纹理的边界。

地平线映射[1027]通过使用凹凸贴图，能够将阴影投射到其自身的表面上，可进一步增强法线贴图。这是通过预计算额外的纹理来实现的，其中每个纹理都与沿着表面平面的一个方向相关联，并为每个纹素存储该方向的水平角度，更多信息请参见第11.4节。

6.8 视差映射 Parallax Mapping

凹凸和法线映射的一个问题是，凹凸永远不会随着视角变化而改变位置，也不会相互遮挡。比如如果你沿着一堵真正的砖墙看，从某个角度你看不到砖块之间的砂浆，墙的凹凸贴图永远不会显示这种类型的遮挡，因为它只是改变法线，最好是在表面渲染每个像素时让凹凸实际影响某个位置。

视差映射的思想在2001年由Kaneko[851]提出，并由Welsh[1866]改进和推广，视差指的是，随着观察者移动，物体的位置相对于另一个物体移动。当观察者移动时，凹凸看起来应该有高度，视差映射的关键思想是通过检查可见物的高度，对像素中应该看到的东西进行有根据的猜测。

对于视差映射，凹凸存储在高度场纹理中，查看给定像素处的表面时，在该位置检索高度场数据，并用于移动纹理坐标以检索表面的不同部分，移动量基于检索到的高度和眼睛到表面的角度，参见图6.37。高度场值要么存储在一个单独的纹理中，要么打包在某些纹理的未使用颜色或alpha通道中（打包不相关的纹理时必须小心，这可能会对压缩质量产生负面影响），高度场值在用于移动坐标之前会被缩放和偏移，缩放比例决定了高度场在表面上方或下方延伸的高度，而偏差给出的是不发生移动处的“海平面”高度。给定纹理坐标位置，调整后的高度场高度h，以及高度值和水平分量的归一化观察向量，新的视差调整纹理坐标为

|  |  |
| --- | --- |
|  | (6.17) |

注意，不同于大多数着色方程，此处执行计算的空间很重要——观察向量需要在切线空间中。

图示

描述已自动生成

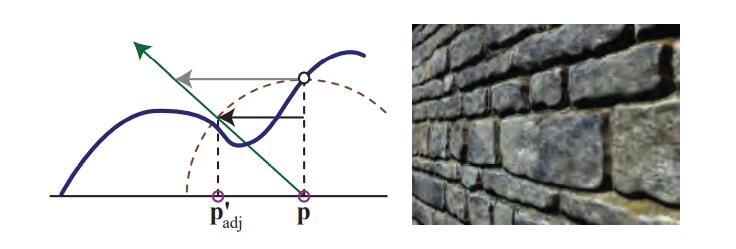
**图 6.37. 左边是想要达成的目标：表面上的实际位置是从观察向量与高度场相交的地方找到的。视差映射通过获取矩形位置的高度并使用它找到一个新的位置来实现目标的一阶近似。（由Welsh[1866]提供。）**

尽管这只是一个简单的近似，但如果在实践中凹凸高度变化相对缓慢，这种移动相当有效[1171]，这样一来，相邻的纹素有差不多的高度，所以用原位置的高度来估计新位置的高度是合理的。不过这种方法在浅视角下行不通，当观察向量接近表面的地平线时，一个小的高度变化会导致一个大的纹理坐标移动，这种近似就失效了，因为检索到的新位置与原始表面位置的高度相关性很小或没有。

为了改善这一问题，Welsh[1866]提出了偏移限制的想法，这个想法是限制移动量，使其永远不要大于检索到的高度，公式如下

|  |  |
| --- | --- |
|  | (6.18) |

注意这个方程比原方程计算起来更快，从几何上来解释，高度定义了一个半径，位置不能超过这个半径，如图6.38所示。



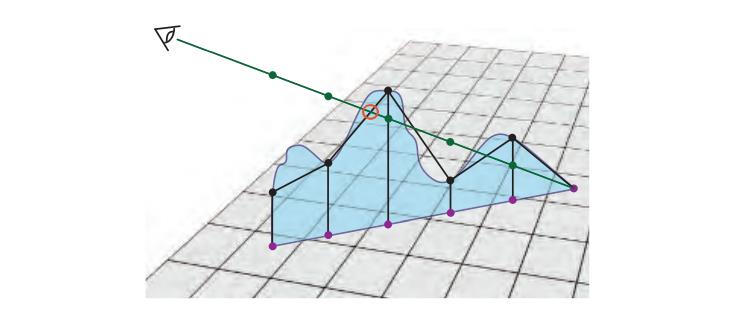
**图 6.38. 在视差偏移限制中，偏移量最多偏离原始位置的高度，显示为虚线圆弧，灰色偏移显示原始结果，黑色偏移显示有限制的结果，右边是用这种技术渲染出来的墙。（图片由Terry Welsh提供。）**

在陡峭（面朝）的角度下，这个方程几乎和原始方程一样，因为接近于1，在较浅的角度，偏移的效果受到限制。从视觉上看，这可以减少较浅角度的凹凸，但这比纹理的随机采样要好得多，问题依然存在，随着观察角度的变化，纹理会发生波动，而对于立体渲染，观察者同时感知两个视点，因此必须给出一致的深度线索[1171]。即使有这些缺点，带偏移限制的视差映射只需花费一点额外的像素着色器程序指令，就比基础的法线映射有了相当大的图像质量改善。Shishkovtsov[1631]通过移动在凹凸贴图法线方向的估计位置，来改善视差遮挡的阴影。

6.8.1 视察遮蔽映射 Parallax Occlusion Mapping

凹凸映射不会根据高度场修改纹理坐标，它只改变一个位置的着色法线，视差映射提供了高度场效应的一个简单近似，它的工作建立在假设是一个像素的高度与其相邻像素的高度大致相同的基础上。这种假设也许会迅速崩塌，凹凸之间可能从不相互遮挡或投下阴影，我们想要的是在像素处可见的东西，即观察向量与高度场的第一个交点。

为了更好地解决这个问题，一些研究人员建议使用射线沿着观察向量前行（ray marching），直到找到一个（近似的）交点，这项工作可以在像素着色器中完成，其中高度数据可以作为纹理访问，我们将这些方法的研究归为视差映射技术的一个子集，它们以某种方式利用射线行进[192, 1171, 1361, 1424, 1742, 1743]。



**图 6.39. 绿色的视线被投射到表面平面上，以固定的间隔采样（紫色点），然后检索高度。该算法找到视线与近似弯曲的高度场的黑色线段的第一个交点。**

这些类型的算法被称为视差遮挡映射（parallax occlusion mapping, POM）或浮雕映射（relief mapping）方法，或其他名字。其关键思想是首先沿着投影向量，测试一定数量的高度场纹理样本，通常掠射角的观察射线会生成更多的样本，以免错过最近的交点[1742, 1743]。沿着射线的每个三维位置都会被检索，转换到纹理空间，并被处理以确定它是高于还是低于高度场，一旦发现高度场以下的样本，这个在高度场以下的量，将和上一个在高度场以上的样本的量一起，用来寻找与高度场相交的位置，参见图6.39，然后使用附加的法线贴图、颜色贴图和其他纹理，在该位置给表面着色。多层高度场可用于产生悬岩（overhangs）、自重叠表面（overlapping surfaces）和双边浮雕贴图代替物（two-sided relief mapped impostors），参见13.7节。高度场跟踪方法也可用于让凹凸不平的表面在其自身上投下阴影，包括硬阴影[1171, 1424]和软阴影[1742, 1743]，它们的比较参见图6.40。



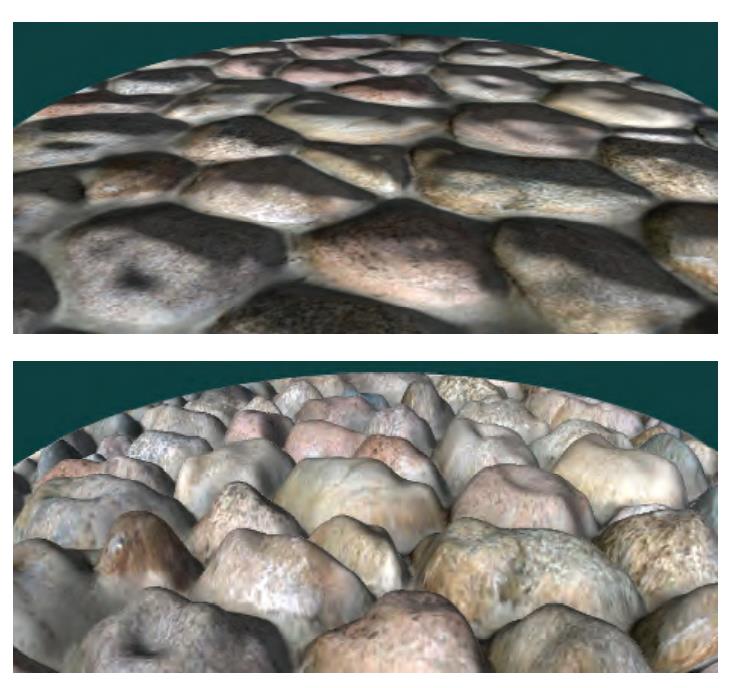
**图 6.40. 没有射线行进的视差映射（左）与有射线行进的视差映射（右）的对比，不使用射线行进时，在立方体的顶部会出现扁平，使用射线行进时，也会产生自阴影效应（self-shadowing effect）。（图片由Manuel M. Oliveira和Fabio Policarpo提供。）**

关于这个话题有大量的文献，虽然所有这些方法都沿着一条射线前进，但也有一些不同之处，你可以使用简单的纹理来检索高度，但也可以使用更高级的数据结构和更高级的求根方法，有些技术可能涉及着色器丢弃像素或深度缓冲区的写入，这会损害性能。下面我们总结了大量的方法，但请记住随着GPU的发展，最好的方法也在发展，这种“最佳”取决于内容和完成射线行进过程的步骤数。

确定两个符合条件的样本之间实际交点的问题是求根（root-finding）问题，在实践中高度场更多地被视为深度场，其中矩形的平面定义了表面的上限，这样一来，平面上的初始点就在高度场之上，Tatarchuk[1742, 1743]在找到高度场表面上的最后一个点，以及表面下的第一个点后，使用单步割线法求近似解。Policarpo等人[1424]在找到的两个点之间使用二分查找法来锁定更近的交点。Risser等人[1497]通过使用割线法迭代来加快收敛速度。折衷（trade-off）的方案是，常规采样可以并行进行，而迭代方法需要更少的整体纹理访问，但必须等待结果并执行较慢的依赖纹理获取，暴力求解的方法在总体上似乎表现更好（well）[1911]。

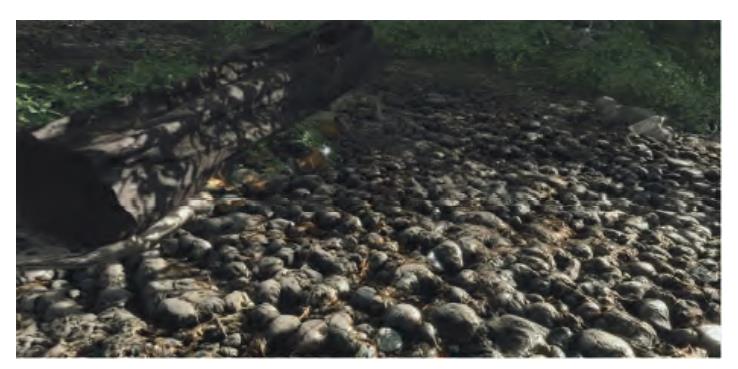
对高度场进行足够频繁的采样是至关重要的，McGuire和McGuire[1171]（译注：原文有误，两个McGuire）提出偏置mipmap查找，并使用各向异性的mipmap来确保高频高度场的正确采样，例如那些代表尖峰或毛发的高度场。你也可以以比法线贴图更高的分辨率存储高度场纹理，最终一些渲染系统甚至不存储法线贴图，转而使用交叉滤波器（cross filter）[40]从高度场动态的派生出法线，696页的公式16.1展示了这种方法（译注：英文页）。

另一种提高性能和采样精度的方法是，不以固定的间隔开始对高度场采样，而是尝试跳过中间的空白区域。Donnelly[367]将高度场预处理为一组体素（voxels），在每个体素中存储它与高度场表面的距离，这样一来可以快速跳过中间空间，代价是每个高度场需要更高的存储空间。Wang等人[1844]使用五维位移映射方案（five-dimensional displacement mapping scheme）来保持从各个方向及位置到表面的距离，这使得复杂曲面、自阴影和其他效果得以实现，代价是要消耗大量的内存。Mehra和Kumar[1195]使用了类似的方向距离贴图。Dummer[393]提出，Policarpo和Oliveira[1426]改进了圆锥阶跃映射（cone step mapping）的思想，这里的概念是还要为每个高度场位置存储一个圆锥半径（cone radius），这个半径定义了射线上的一个区间，其中最多只有一个与高度场的交点，此特性允许沿着涉嫌快速跳过而不错过任何可能的交点，尽管代价是需要依赖的纹理读取，创建圆锥阶跃映射的另一个缺点是需要预计算，这使得该方法无法用于动态变化的高度场。Schroders和Gulik[1581]提出了四叉树浮雕映射（quadtree relief mapping），这是一种在遍历过程中跳过体积（volume）的分层方法。Tevs等人[1760]使用“最大ipmap”允许跳过（体积）的同时最小化预计算成本。Drobot[377]还使用存储在mipmap中的类四叉树结构来加速遍历，并提出了一种混合不同高度场的方法，其中一种地形类型（terrain type）转换为另一种地形类型。



**图 6.41. 法线映射和浮雕映射，法线映射时不会发生自遮挡，浮雕映射在重复纹理的轮廓方面存在问题，因为矩形更多的是高度场的视图，而不是真正的边界定义。（图片由NVIDIA公司提供。）**

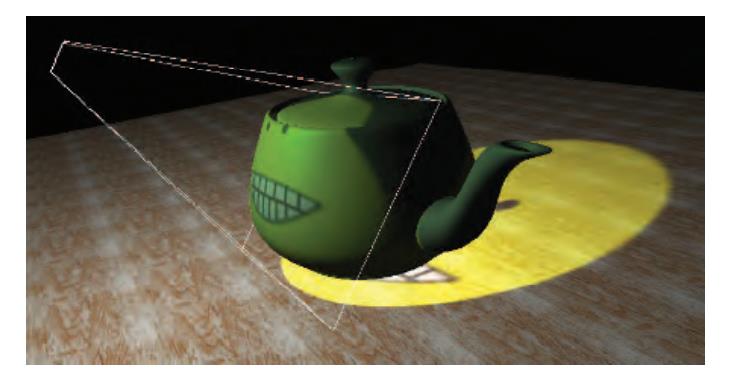
上述所有方法的一个问题是沿物体轮廓边缘观察时，幻想会被打破，这时将显示原始表面的平滑轮廓，见图6.41。关键思想是，渲染的三角形定义了哪些像素应该由像素着色程序计算，而不是表面实际的位置，此外对于曲面，轮廓的问题变得更加复杂。Oliveira和Policarpo[1325, 1850]描述并发展了一种方法，它使用二次轮廓逼近技术（quadratic silhouette approximation technique）。Jeschke等人[824]和Dachsbacher等人[323]都给出了更通用和更健壮的方法（并回顾了以前的工作），以正确处理轮廓和曲面，这种方法最初由Hirche[750]探索，其总体思路是将网格中的每个三角形向外挤压，以形成一个棱柱，渲染这个棱柱需要对高度场中可能出现的所有像素进行计算。这种方法称为外壳映射（shell mapping），因为扩展的网格在原始模型上形成了一个单独的壳，通过保留棱柱与光线相交时的非线性性质，可以实现无伪像的高度场渲染，尽管计算成本昂贵，这种技术的一个令人印象深刻的使用如图6.42所示。



**图 6.42. 视差遮挡映射，又称浮雕映射，用于路径上使石头看起来更真实。地面实际上是一组简单的三角形，加上了一个高度场。（图片来自游戏孤岛危机“Crysis”，由Crytek提供。）**

6.9 纹理灯光 Textured Lights

纹理还可以用来增加光源的视觉丰富性，并允许复杂的强度分布或聚光灯功能，对于所有亮度都限定在圆锥或平截圆锥的灯，可使用投影纹理来调节光强[1192, 1597, 1904]。这允许形状聚光灯、图案灯、甚至“幻灯片投影仪”效果（图6.43），这些灯通常被称为gobo或cookie灯，这是专业剧院和电影照明中使用的切割灯的术语。参见7.2节，讨论了以类似的方式使用投影映射（projective mapping）来投射阴影。



**图 6.43. 带投影纹理的光，纹理被投影到茶壶和地面上，并用于调节光线在投影截锥内的分布（截锥外被设置为0）。（图片由NVIDIA公司提供。）**

对于不局限于截锥内而是照亮所有方向的光，可以使用立方体映射来调节强度，而不是二维投影纹理，一维纹理可以用来定义任意距离的衰减函数，结合二维角度衰减图，可用于复杂的体积照明模式[353]。更普遍的可能是使用三维（体积）纹理来控制光的衰减[353, 535, 1192]，这允许任意体积的效果，包括光束，这种技术是内存密集型的（如所有的体积纹理一样）。如果光线体积的影响量沿三个轴对称，通过将数据镜像到每个八分之一，内存占用可以减少八倍。

纹理可以添加到任意光类型，以实现额外的视觉效果，带纹理光允许美工轻松控制照明，他们可以简单地编辑使用纹理。

进一步阅读和资源 Further Reading and Resources

Heckbert写了一篇很好的关于纹理映射理论的调查[690]和一篇更深入的关于该主题的报告[691]。Szirmay-Kalos和Umenhoffer[1731]对视差遮挡映射和位移方法进行了出色而彻底的研究。更多关于法线表示的信息可以在Cigolle等人[269]和Meyer等人[1205]的工作中找到。

《使用OpenGL的高级图形编程》（Advanced Graphics Programming Using OpenGL）[1192]一书广泛地介绍了使用纹理化算法的各种可视化技术。有关三维程序化纹理的广泛介绍，可参见《纹理和建模：一个程序方法》（Texturing and Modeling: A Procedural Approach）[407]。《高级游戏开发与可编程图形硬件》（Advanced Game Development with ProgrammableGraphics Hardware）[1850]一书有许多关于实现视差遮挡映射技术的细节，Tatarchuk的报告[1742,1743]和Szirmay-Kalos和Umenhoffer的调查[1731]也是如此。

对于程序化纹理（和建模），我们最喜欢的网站是Shader-toy，那里展示了许多有价值和迷人的程序化纹理函数，你也可以轻松地修改示例并查看结果。

访问本书的网站，realtimerendering.com，以获取许多其他资源。

第7章 阴影 Shadows

“生活的一切变化，一切魅力，一切美好，都是由光和影构成的”

——Tolstoy（托尔斯泰）

阴影对于创建逼真的图像和为用户提供关于物体位置的视觉提示非常重要，本章重点介绍阴影计算的基本原理，并介绍最重要和最流行的实时计算算法，我们还将简要讨论一些不太流行但包含重要原则的方法。本章不会花时间涉及所有的选择和方法，因为有两本全面的书深入研究阴影领域[412, 1902]，相反我们专注于研究自它们出版以来出现的文章和演示，偏向于经过实战检验的技术。

图示

描述已自动生成

**图 7.1. 阴影术语：光源、遮挡器、接收器、阴影、本影和半影。**

本章使用的术语如图7.1所示，其中遮挡器（occluders）是将阴影投射到接收器（receivers）上的物体，精确光源（Punctual light sources），即那些没有面积的光源，只会产生完全阴影的区域，有时称为硬阴影（hard shadows）。如果使用具有面积或体积的光源，就会产生软阴影（soft shadows），每个软阴影会有一个完全阴影的区域，称为本影（umbra），和一个部分阴影的区域，称为半影（penumbra），软阴影通过它们模糊的阴影边缘来识别，不过，重要的是要注意到，它们通常不能仅通过使用低通滤波器模糊硬阴影的边缘来正确渲染。如图7.2所示，对于正确的软阴影，阴影投射几何体越接近接收器，阴影就越清晰，软阴影的本影区域并不等同于由某个精确光源产生的硬阴影，相反，当光源变大时，软阴影的本影区域会变小，如果光源足够大或接收器距离遮挡器足够远时，它甚至可能消失。软阴影通常是更好的，因为半影的边缘让观察者知道这的确是一个阴影，硬阴影通常看起来不太真实，有时会被误解为实际的几何特征，如表面上的折痕，不过硬阴影比软阴影渲染速度更快。

图片包含 户外, 男人, 照片, 年轻

描述已自动生成

**图 7.2. 硬阴影和软阴影混合在一起，来自板条箱的阴影是清晰的，因为遮挡器靠近接收器，人的影子在地面接触点是****清晰的，随着与遮挡器的距离增加而软化，远处的树枝投下软阴影[1711]。（图片来自“全境封锁”“Tom Clancy’s The Division”，由育碧提供。）**

比具有半影更重要的是得有阴影，没有阴影作为视觉提示，场景往往不真实，也更难感知。正如Wanger[1846]所提到的，通常有一个不准确的阴影也比没有阴影要好，因为眼睛对阴影的形状是相当宽容的，例如将一个模糊的黑色圆圈作为纹理在地板上应用，可以将角色固定在地面上。

在接下来的章节中，我们将越过这些简单的阴影模型，介绍从场景中的遮挡器实时自动计算阴影的方法。第一节处理在平面上投射阴影的情况，第二节介绍更一般的阴影算法，即在任意表面上投射阴影，硬阴影和软阴影都将被介绍。作为总结，最后提出了一些适用于各种阴影算法的优化技术。

7.1 平面阴影 Planar Shadows

当物体在平面上投射阴影时，会出现一个简单的阴影，本节介绍了几种平面阴影的算法，每种算法都在阴影的柔和度和真实感方面有不同。

7.1.1 投影阴影 Projection Shadows

在此方案中，三维物体被渲染两次以创建阴影，可以通过推导一个矩阵，将物体的顶点投影到平面上[162, 1759]，考虑图7.3中的情况，其中光源位于处，待投影顶点位于处，投影顶点位于处，我们将推导出阴影平面为的特殊情况下的投影矩阵，然后将该结果推广到其它平面。

我们首先推导x坐标的投影，从图7.3左边的相似三角形中，我们得到

|  |  |
| --- | --- |
|  | (7.1) |

z坐标可以用同样的方法得到：，同时y坐标为零，现在这些方程可以转化为投影矩阵：

|  |  |
| --- | --- |
|  | (7.2) |

的证明很简单，同时这意味着确实是投影矩阵。

图示

描述已自动生成

**图 7.3. 左图：位于处的光源在平面上投下阴影，顶点被投影到平面上，投影点称为，相似三角形用于推导投影矩阵。右图：投影被投射到一个平面上，。**

一般情况下，阴影应该投射到的平面不是，而是，这种情况在图7.3的右边部分进行了描述。我们的目标是再找一个投影到的矩阵，为了达到这个目的，在处发射的射线穿过与平面相交，就得到了投影点：

|  |  |
| --- | --- |
|  | (7.3) |

这个方程也可以转化为投影矩阵，如式7.4所示，其中满足：

|  |  |
| --- | --- |
|  | (7.4) |

可以预见，当平面为时，此矩阵转换为公式7.2中的矩阵，即，。

要渲染阴影，只需将这个矩阵应用到应该在平面上投下阴影的对象上，并将这个投影对象渲染为深色且没有亮度。在实践中，你必须采取措施避免让投影的三角形在接收它们的表面下渲染，一种方法是在我们投影的平面上添加一些偏差，这样阴影三角形总是在表面的前面渲染。

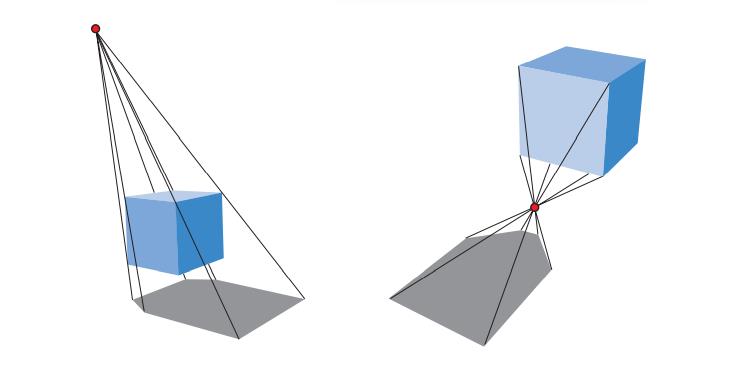
一个更安全的方法是先画出地平面，然后画出关闭z缓冲区写入的投影三角形，然后像往常一样渲染其余的几何图形，如此一来投影的三角形总是画在地平面的顶部，因为不进行深度比较。

如果地平面有一些限制，例如，它是一个矩形，投影的阴影可能会落在它的外面，从而破坏视觉效果，为了解决这个问题，我们可以使用模板缓冲区。首先，将接收器绘制到屏幕和模板缓冲区，然后关闭z缓冲区，只在接收器被绘制的地方绘制投影三角形，然后正常渲染场景的其余部分。

另一种阴影算法是将三角形渲染成纹理，然后应用到地平面上，该纹理是一种光照贴图（light map），它可以调节底层表面的光照强度（第11.5.1节）。可以看出，渲染阴影投影到纹理的想法也适用于半影和曲面上的阴影，这种技术的一个缺点是纹理会被放大，一个纹素覆盖多个像素，破坏了视觉效果。

如果阴影从一帧到另一帧不会发生变化，即光线和阴影的投射器之间彼此不会相对移动，那么这个纹理可以被重用，如果没有发生变化，大多数阴影技术都可以从重用帧与帧之间的中间计算结果中受益。

所有阴影投射器都必须在光和地面接收器之间，如果光源在物体上的最高点之下，则会产生反阴影（antishadow）[162]，因为每个顶点都通过光源上的点投影，正确的阴影和反阴影如图7.4所示，如果我们投影一个物体到接收平面之下，也会发生错误，因为它也不该投射阴影。



**图 7.4. 在左边，显示了一个正确的阴影，而在右边的图中，出现了一个反阴影，因为光源在物体的最高顶点之下。**

当然，明确地剔除和修剪阴影三角形以避免此类伪像是可能的，下面介绍的一个更简单的方法是使用现有的GPU管线来执行带剪切的投影。

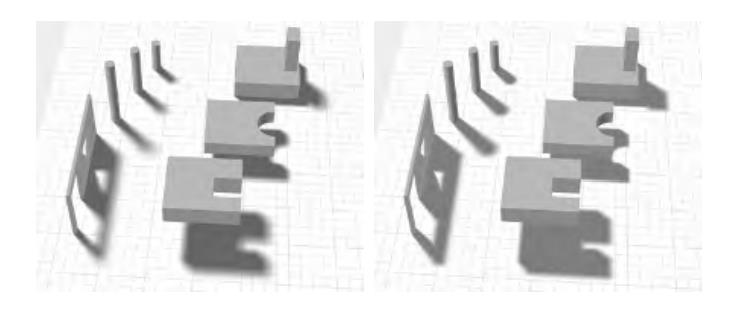
7.1.2 软阴影 Soft Shadows

通过使用各种技术，投射阴影可以变得柔和，这一节我们将描述Heckbert和Herf[697, 722]提出的一种生成软阴影的算法，该算法的目标是在地平面上生成显示软阴影的纹理，之后我们描述准确性稍低，但更快的方法。

只要光源有一个区域，就会出现软阴影，一种近似区域光效果的方法是在区域光的表面上放置几个精确光源进行采样，对于每一个精确光源，图像被渲染并积累到缓冲区中，这些图像的平均值就是一个软阴影的图像。请注意理论上，任何生成硬阴影的算法都可以与这种累积技术一起使用来生成半影，在实践中，由于涉及到执行时间，以交互速率这样做通常是难以为继的。

Heckbert和Herf使用一种基于截锥体的方法来制作他们的阴影，这个想法是把光当作观察者，地平面形成截锥的远切平面，截锥足够宽以包含遮挡器。

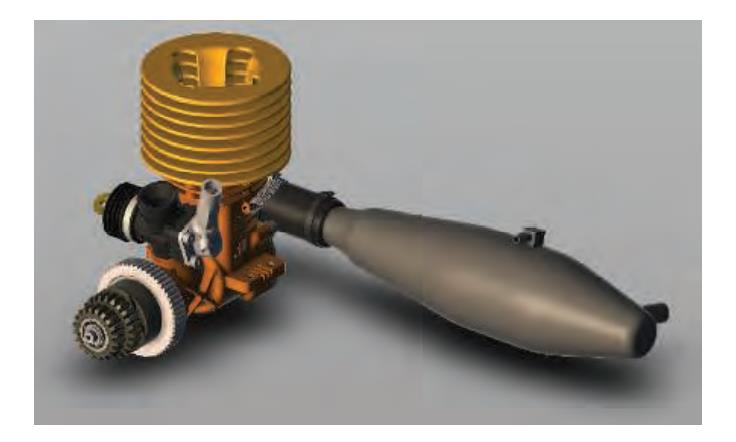
软阴影纹理是通过生成一系列地平面纹理而形成的，区域光在其表面采样，其中每个位置被用于着色代表地平面的图像，然后将阴影投射对象投影到该图像上，所有这些图像相加并平均后，产生一个地平面阴影纹理，参见图7.5左侧的示例。



**图 7.5. 在左边，用Heckbert和Herf的方法渲染，用了256个pass。在右边，用Haines的方法渲染了一个pass，用Haines的方法本影太大了，门口和窗户的周围特别明显。**

采样区域光方法的一个问题是，它往往看起来像是：来自精确光源的几个重叠的阴影，此外对于n个阴影pass，只能生成n+1个不同的阴影，大量的pass可以得到准确的结果，但代价过高，该方法可用于获得（字面意思上）“地面真实的”图像，来测试其他更快算法的质量。

一种更有效的方法是使用卷积，即滤波。在某些情况下，模糊处理从单个点生成的硬阴影就足够了，这可以生成与真实世界内容合成的半透明纹理，见图7.6，然而在接近物体与地面接触的地方，出现的均匀的模糊可能无法令人信服。



**图 7.6. 投影（Drop shadow），阴影纹理是通过在上方的阴影投射器渲染，然后模糊图像并将其在地平面上渲染而生成的。（图像在Autodesk的A360查看器中生成，模型来自Autodesk的Inventor示例。）**

还有许多其他方法可以给出更好的近似，但需要额外的成本。例如，Haines[644]从一个投影的硬阴影开始，然后用从中心的黑色到边缘的白色的梯度渲染轮廓边缘，以创建可信的半影，参见图7.5的右侧。然而这些半影在物理上是不正确的，因为它们应该延伸到轮廓边缘内部的区域。Iwanicki[356, 806]借鉴了球面谐波（spherical harmonics）的思想，用椭球体近似遮挡角色以生成软阴影。所有这些方法都有不同的近似程度和缺点，但都比对大量投影图像进行平均处理有效得多。

7.2 曲面上的阴影 Shadows on Curved Surfaces

将平面阴影的概念扩展到曲面的一个简单方法是使用生成的阴影图像作为投影纹理[1192, 1254, 1272, 1597]，从光的角度思考阴影，光所看见的地方就照亮，它看不见的东西都在阴影里，假设遮挡器从光源的角度看被渲染为黑色纹理，否则为白色纹理，这样该纹理可以投影到接收阴影的表面上。更高效地，接收器上的每个顶点都有一个为其计算的纹理坐标，并对其应用纹理，这些纹理坐标可以由应用程序显式计算，这与前一节中的地面阴影纹理略有不同，前一节中，对象被投影到特定的物理平面上，在这里，图像是从光的角度形成的，就像放映机中的一帧胶片。

渲染时，投影阴影纹理会改变接收器的表面，它还可以与其他阴影方法相结合，有时还主要用于帮助感知对象的位置。举个例子，在一款跨平台电子游戏中，主角的正下方总是会出现投影，即使角色是在完全的阴影中[1343]。更精确的算法能得到更好的结果，例如，Eisemann和Décoret[411]假设了一个矩形的头顶上光，并创建对象水平切片的阴影图像堆栈，然后将其转换为mipmaps或类似的东西，通过使用mipmap，每个切片的对应区域被与其到接收器的距离成正比地访问，这意味着越远的切片会投射出更软的阴影。

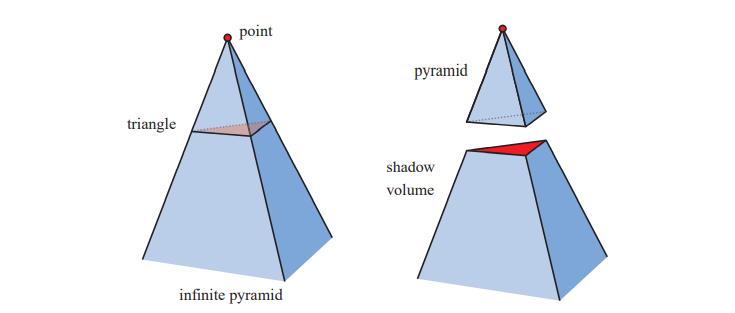
纹理投影方法有一些严重的缺点，首先，应用程序必须识别哪些对象是遮挡器，哪些是它们的接收器，程序必须使接收器比遮挡器离光更远，否则阴影就会“向后投射”。此外，遮挡物不能遮挡自己，接下来的两个节将介绍生成正确阴影的算法，而不需要这样那样的干预或限制。

注意，各种照明模式可以通过使用预先构建的投影纹理获取，聚光灯就是一个简单的正方形投影纹理，其内有一个圆圈来定义光线。威尼斯百叶窗（Venetian blinds）的效果可以通过由水平线组成的投影纹理来生成，这种类型的纹理被称为光衰减蒙版（light attenuation mask）、cookie纹理或gobo贴图。通过简单地将两个纹理相乘，可以将预构建的图案与动态创建的投影纹理相结合，这种光将在第6.9节中进一步讨论。

7.3 阴影体 Shadow Volumes

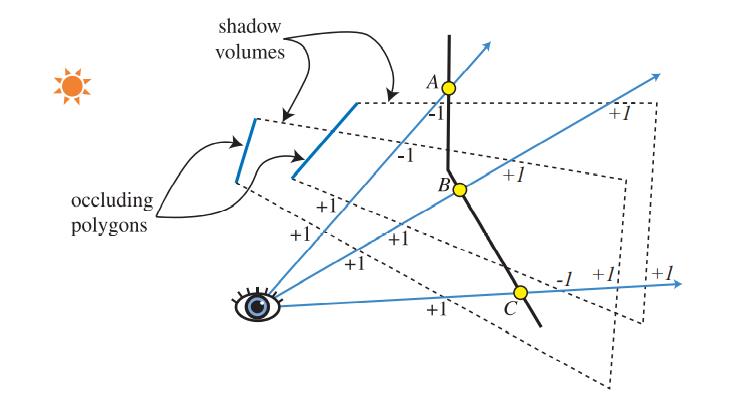
Heidmann在1991年[701]提出了一种基于Crow 's shadow volume[311]的方法，通过巧妙地使用模板缓冲区，可以将阴影投射到任意物体上，它可以在任何GPU上使用，因为唯一的要求是模板缓冲区，它不是基于图像的（不同于下面描述的阴影贴图算法），因此避免了采样问题，从而在各处产生正确清晰的阴影。这有时可能是一个缺点，例如，一个角色的衣服可能会有褶皱，产生薄而硬的阴影，从而严重走样，由于其不可预测的成本[1599]，体积阴影如今很少使用。我们在这里简单描述该算法，因为它阐述了一些重要的原理，基于这些原理研究还在继续。

首先，想象一个点和一个三角形，把从一点到三角形顶点的线延伸到无穷远处，就会得到一个无限大的三面棱锥体，三角形下面的部分，即不包含点的部分，是一个截断的无限棱锥，上面的部分只是一个棱锥，如图7.7所示。现在想象这个点实际上是一个点光源，这样的话，物体在截断棱锥的体积内（在三角形下）的任意部分都在阴影中，这个体积称为阴影体。



**图 7.7. 左边：来自点光源的线条通过三角形的顶点延伸形成一个无限的棱锥。右边：上面部分是一个棱锥，下面是一个无限截断的棱锥，也成为阴影体。阴影体积内的所有几何图形都在阴影中。**

假设我们观察一些场景，并跟随从眼睛穿过像素的射线，直到射线击中要显示在屏幕上的对象，当射线在到达该物体的路径上时，每当它穿过阴影体积的正面（即面向观察者）的一个面时，我们就增加一个计数器，因此，每当光线进入阴影时，计数器就增加一次。同样的方式，每当射线穿过截断棱锥的背面，我们减少相同的计数器，然后射线从阴影中射出，我们继续增加或减少计数器，直到射线击中要在该像素处显示的对象。如果计数器大于零，则该像素处于阴影中，否则就不是，这个原则也适用于多个三角形投射阴影时，参见图7.8。



**图 7.8. 使用两种不同计数方法计算阴影-体积交叉的二维侧视图。在z-pass体积计数中，当射线穿过阴影体积的正面三角形时，计数增加，当射线穿过背面三角形时，计数减少。因此在点A处，射线进入两个阴影体积所以计数+2，然后离开两个体积，留下净计数为0，所以点在光中。在z-fail体积计数中，计数从表面以外开始（这些计数以斜体字显示），对于B点的射线，z-pass方法通过穿过两个正面三角形得到+2的计数，而z-fail方法通过穿过两个背面三角形得到相同的计数。点C展示了为什么z-fail阴影体积必须有上限，从C点出发的射线首先击中一个正面的三角形，得到−1，然后从两个阴影体积中出来（通过它们的底面，这是此方法正常工作所必需的），给出净计数为+1，计数不是零，所以点在阴影中，这两种方法对观察表面上的所有点总是给出相同的计数结果。**

使用射线进行此操作非常耗时，但有一个更好的解决方案[701]：用一个模板缓冲区为我们计数，首先，清除模板缓冲区；第二，整个场景被绘制到帧缓冲区中，只使用未照亮（unlit）材质的颜色，以获得颜色缓冲区中的着色成分和z缓冲区中的深度信息；第三，关闭z缓冲区更新和颜色缓冲区写入（尽管z-buffer测试仍在进行），然后绘制阴影体积的正面三角形，在此过程中，模板操作被设置为在绘制三角形的地方增加模板缓冲区中的值；第四，用模板缓冲渲染另一个pass，这次只绘制阴影体积的背面三角形，这个pass中，当绘制三角形时，模板缓冲区中的值会递减，只有当渲染的阴影-体积面的像素可见（即不被任何真实的几何图形隐藏）时才进行递增和递减，此时模板缓冲区保存了每个像素的阴影状态；最后，再次渲染整个场景，这一次只渲染受光照影响的活性材质部分，并且只在模板缓冲区中的值为0的地方显示，值为0表示光线离开阴影的次数与它进入阴影体积的次数相同——即这个位置被光照亮。

这种计数方法是阴影体积背后的基本思想，阴影体积算法生成的阴影示例如图7.9所示，一些有效的方法可以在单次pass中实现该算法[1514]。然而，当物体穿透相机的近平面时，将发生计数问题，该问题的解决方案被称为z-fail，包括计算隐藏在可见表面后面而不是前面的交叉点[450,775]，图7.8显示了这个备选方案的简要摘要。



**图 7.9. 阴影体。在左边，一个人物投下了阴影。右边显示了模型的挤压三角形。（图片来自Microsoft SDK [1208]的示例“Shadow Volume”。）**

为每个三角形创建四边形会产生大量的透支（overdraw），也就是说，每个三角形将创建三个必被渲染的四边形，一个由一千个三角形组成的球体会产生三千个四边形，而每一个四边形都可能跨越屏幕。一种解决方案是只画那些沿着物体轮廓边缘的四边形，例如，我们的球体可能只有50个轮廓边缘，所以只需要50个四边形。几何着色器可用于自动生成这样的轮廓边缘[1702]，裁剪和clamping技术也可用于降低填充成本[1061]。

然而，阴影体积算法仍然有一个可怕的缺点：极端的可变性，想象一个单独的小三角形在视野中，如果相机和光源在完全相同的位置，阴影体积成本是最小的，形成的四边形不会覆盖任何像素，因为看到的是它们的侧视图，只有三角形本身重要，现在假设观察者绕着三角形运动，使它始终位于视图中，随着相机远离光源，阴影体积的四边形将变得更加可见，并覆盖屏幕中的更多部分，导致更多的计算。如果观察者碰巧移动到三角形的阴影中，阴影体积将完全填满屏幕，与我们原始的视图相比，这需要花费相当多的时间来计算，这种可变性使得阴影体积在交互式应用程序中不可用，在交互式应用程序中一致的帧率非常重要，与其他场景一样，面向光线观察会导致算法成本出现不可预测的巨大跳跃。

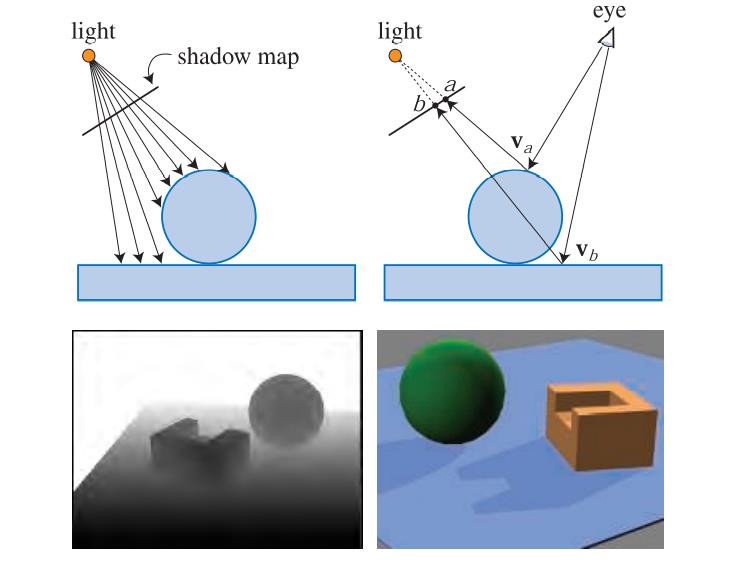
由于这些原因，阴影体在很大程度上被应用程序所抛弃，然而考虑到GPU上访问数据的新的不同的方法在不断发展，以及研究人员对这些功能的巧妙再利用，阴影体可能有一天会重新被广泛使用，例如，Sintorn等人[1648]概述了提高效率的阴影体积算法，并提出了自己的分层加速结构。

下面介绍的算法，阴影映射，具有更可预测的成本，非常适合GPU，因此在许多应用中成为阴影生成的基础。

7.4 阴影贴图 Shadow Maps

1978年，Williams[1888]提出了一种基于z缓冲区的通用渲染器，可用于在任意物体上快速生成阴影。想法是使用z缓冲区渲染场景，从光源的位置投射阴影，光“看到”的一切都是照亮的，其余的都在阴影中，生成这个图像时只需要z缓冲，光照、纹理和颜色缓冲区写入值都可以关闭。

现在z缓冲区中的每个像素都包含最靠近光源的对象的z深度，我们将这个z缓冲区的全部内容称为阴影贴图，有时也称为阴影深度贴图或阴影缓冲区。为了使用阴影贴图，场景将被第二次渲染，但这一次是相对于观察者，渲染每个绘制图元时，每个像素处的位置将与阴影贴图进行比较，如果渲染点距离光源的距离比阴影贴图中对应的值远，则该点在阴影中，否则不在阴影中。该技术通过使用纹理映射实现，参见图7.10。阴影映射是一种流行的算法，因为它是相对可预测的，构建阴影贴图的成本与渲染图元的数量大致成线性关系，而访问时间是恒定的。阴影贴图可以生成一次，并在光线和物体不移动的场景中每帧重复使用，例如计算机辅助设计（computer-aided design, CAE）软件。



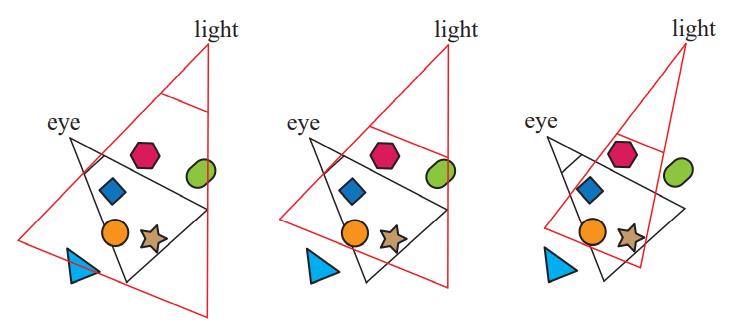
**图 7.10. 阴影映射。在左上角，通过存储视图中表面的深度来形成阴影贴图，在右上方，眼睛看向两个位置，球体在点va处被看到，这个点被发现位于阴影贴图上的纹素a处，存储在那里的深度并不（太）少于va点到光源的距离，所以这个点被照亮。矩形在点vb处被视线击中，该点距离光源的距离比存储在纹素b处的深度要远得多，所以位于阴影中。左下角是灯光视角下的场景，白色表示距离较远，右下角是用阴影贴图渲染的场景。**

当生成单个z缓冲区时，光线只能“看”向一个特定的方向，就像相机一样，对于遥远的方向光，如太阳光，光的视野设置为包括所有投射阴影到眼睛看见的观察体积中的物体，光使用正交投影，它的视野需要在x和y上足够宽和高，以看见那些对象。局部光源需要尽可能进行类似的调整，如果局部光距离投射阴影的物体足够远，一个视锥体可能就足以包含所有物体。如果局部光是一个聚光灯，它有一个与之相关的天然锥体，锥体以外的一切都被认为没有照亮。

如果局部光源位于场景内部，并且被阴影投射器包围，典型的解决方案是使用六视图立方体，这类似于立方体环境映射（cubic environment mapping）[865]，它们被称为全向阴影贴图（omnidirectional shadow maps），全向阴影贴图的主要挑战是避免沿两个独立贴图的接合处产生伪像。King和Newhall[895]深入分析了这个问题并提供了解决方案，Gerasimov[525]提供了一些实现细节。Forsyth[484, 486]为全向光照提出了一种通用的多截锥体分区方案，该方案还在需要的地方提供了更多的阴影贴图分辨率。Crytek[1590, 1678, 1679]设置一个点光源六个视图中的每个视图的分辨率，基于每个视图的投影截锥体的屏幕空间覆盖范围，所有的贴图都存储在一个纹理图集中。

并非场景中的所有对象都需要渲染到光线的观察体积中，首先，只有那些可以投射阴影的对象才需要被渲染，例如，如果已知地面只能接收阴影而不能投射阴影，那么它就不必渲染到阴影贴图中。

定义上讲，阴影投射器是光的视锥内的那些对象，这种截锥体可以通过多种方式扩大或收紧，使我们可以安全地忽略一些阴影投射器[896, 1812]。想想那些肉眼可见的阴影接收器，它们在沿着光的视野方向的某些最大距离内，任何超出这个距离的东西都不能在可见的接收器上投下阴影，同样，可见接收器的集合可能比光的原始x和y视图边界要小，见图7.11。另一个例子是，如果光源在眼睛的视锥体内，那么这个额外的视锥体之外的任何物体都不能在接收器上投下阴影，只渲染相关对象不仅可以节省渲染时间，还可以减少光的截锥体所需的尺寸，从而提高阴影贴图的有效分辨率，从而提高质量。此外它有助于使光截锥体的近平面距离光源尽可能的远，或远平面距离光源尽可能的近，这样做可以提高z缓冲区的有效精度[1792]（第4.7.2节）。



**图 7.11. 在左边，光线的视野包含了眼睛的视锥体。在中间，光的远平面被拉进到只包括可见的接收器，所以剔除了一个作为投射器的三角形，近平面也进行了调整。在右边，光的截锥体边缘被用于限制可见的接收器，所以剔除了绿色胶囊。**

阴影映射的一个缺点是阴影的质量取决于阴影贴图的分辨率（以像素为单位）和z缓冲区的数值精度，由于阴影贴图是在深度比较过程中采样的，因此算法容易出现走样问题，特别是接近物体之间的交点时。一个常见的问题是自阴影走样（self-shadow aliasing），通常被称为“表面粉刺”或“阴影粉刺”，在这种情况下，三角形被错误地认为是自身的阴影，这个问题有两个来源，一个只是处理器精度的数值限制，另一个来源是几何的，源于点样本的值被用来表示一个区域的深度，也就是说，为光生成的样本几乎从不与屏幕样本位于相同的位置（例如，像素通常在其中心采样），当光的存储深度值与被观察表面的深度值比较时，光的值可能略低于表面的深度值，导致自阴影。这种错误的影响如图7.12所示。



**图 7.12. 阴影映射偏差伪像。在左边，偏差太低，所以发生自阴影。在右边，高的偏差导致鞋子不投射接触阴影。阴影贴图的分辨率也太低了，使阴影具有块状外观。（图像来自Christoph Peters的阴影演示。）**

一种帮助避免（但不总是消除）各种阴影映射伪像的常见方法是引入偏差因子，当检查阴影贴图中发现的距离与被测试位置的距离时，从接收器的距离中减去一个小偏差，参见图7.13，这个偏差可以是一个常数值[1022]，但当接收器没有完全面向光源时，这样做可能会失败。一种更有效的方法是使用与接收器到光线的角度成正比的偏差，表面越远离光线，偏差就越大以避免这个问题，这种类型的偏差称为斜率比例偏差（slope scale bias），这两种偏差都可以通过使用OpenGL的glPolygonOffset()或类似的命令来应用，以将每个多边形移离光线。请注意，如果一个表面直接面对光，它完全不会因为斜率比例偏差而向后偏置，由于这个原因，一个恒定的偏差可以与斜率比例偏差一起使用，以避免可能的精度错误，斜率比例偏差也经常被clamp到一些最大值上，因为从光的角度观察，当表面几乎侧面朝向观察者时，切线值可能非常高。



**图 7.13. 阴影偏差。以头顶光将表面渲染成阴影贴图，垂直线表示阴影贴图的像素中心，遮挡器深度记录在×位置，我们想知道在三个显示为点的样本处，表面是否被照亮，每个最接近的阴影贴图深度值用相同颜色的×表示。在左边，如果不增加偏差，蓝色和橙色样本将被错误地确定为在阴影中，因为它们距离光源的距离比对应的阴影贴图深度更远。在中间，从每个样本中减去固定的深度偏差，使每个样本更靠近光源，蓝色样本仍然被认为在阴影中，因为它并不比它所测试的阴影贴图深度更接近光源。在右边，阴影贴图是通过将每个多边形以与其斜率有关的比例，远离光源而形成的，所有的样本深度现在都比它们的阴影贴图的深度更近，所以它们都被照亮。**

Holbert[759, 760]引入了法线偏移偏差（normal offset bias），它首先以光线方向和几何法线之间的夹角正弦值成正比地，将接收器的世界空间位置沿表面法线方向移动一点，见图7.24，这不仅改变了深度，还改变了在阴影贴图上测试的样本的x坐标、y坐标，随着光线到表面的角度越来越浅，这个偏移量会增加，我们希望样本离表面足够远以避免自阴影。这种方法可以被可视化为将样本移动到接收器上方的“虚拟表面”，这个偏移量是一个世界空间距离，因此Pettineo[1403]建议根据阴影贴图的深度范围缩放它，Pesce[1391]提出了沿相机观察方向倾斜的想法，这也可以通过调整阴影贴图的坐标来实现。7.5节中将讨论其他偏差方法，那里提到的阴影方法还需要测试几个相邻的样本。

太大的偏差会导致被称为光泄漏（light leaks）或Peter Panning的问题，这种情况下，物体似乎轻微漂浮在底层表面之上，这种伪像出现是因为物体接触点下面的区域，例如，脚下面的地面，被推得太前了，所以没有接收到阴影。

一种避免自阴影问题的方法是只渲染阴影贴图的背面，该方案被称为二次深度阴影映射（second-depth shadow mapping）[1845]，它在许多情况下都能很好地工作，特别是在无法手动调整偏差的渲染系统中，当物体是双面的、薄的或相互接触时，就会出现问题，如果一个物体是两面的网格都可见的模型，例如棕榈叶或一张纸，就会发生自阴影，因为背面和正面在同一位置。类似地，如果不执行偏置，问题就可能出现在轮廓边缘或薄的物体附近，因为在这些区域的背面接近正面，添加偏差可以帮助避免表面粉刺，但这样更容易受到光线泄漏的影响，因为接收器和遮挡器的背面的接触点之间并没有分离，参见图7.14。选择哪种方案取决于具体情况，例如，Sousa等人[1679]发现在他们的应用中，阳光阴影使用正面、室内光照使用背面效果最好。

形状

描述已自动生成

**图 7.14. 头顶光光源下的阴影贴图表面。在左边，面向光的表面用红色标记并发送到阴影贴图，表面可能被错误地确定为它们自己的阴影（“粉刺”），因此需要偏离光线。在中间，只有背面的三角形渲染到阴影贴图中，将这些遮挡器向下推的偏差可能会让光泄漏到位置a附近的地平面上，向前的偏差会使标记为b的轮廓边界附近的照明位置被认为是在阴影中。在右边，在阴影贴图的每个位置上，使用最接近的正面和背面三角形之间的中点形成一个中间表面，光泄漏可能发生在点c附近（这也可能发生在二次深度阴影映射中），因为最近的阴影贴图样本可能在该位置左侧的中层表面上，因此该点将更接近光。**

注意，对于阴影映射，对象必须是“水密的”（watertight，流形和封闭的，即固体，见第16.3.3节），或者必须将正面和背面都渲染到贴图上，否则对象可能无法完全投射出阴影。Woo[1900]提出了一种通用的方法，从字面上讲，它试图在仅使用正面或背面生成阴影之间找到一种折衷的方法，其想法是渲染固体物体到阴影贴图，并跟踪两个最接近光的表面，这个过程可以通过深度剥离或其他与透明相关的技术来完成，两个物体之间的平均深度形成一个中间层，该中间层的深度被用作阴影贴图，有时它被称为双重阴影贴图（dual shadow map）[1865]。如果物体足够厚，自阴影和光泄漏伪像将被最小化，Bavoil等人[116]讨论了处理潜在伪像的方法，以及其他实现细节，主要的缺点是与使用两个阴影贴图相关的额外成本，Myers[1253]讨论了遮挡器和接收器之间由美工控制的深度层。

随着观察者的移动，光的观察体积经常随阴影投射器的变化而变化，这样的变化反过来导致阴影在帧与帧之间的轻微移动，这是因为光的阴影贴图从光采样了一组不同的方向，而这些方向并没有与前一组对齐。对于定向光，解决方案是强制每个后续生成的阴影贴图在世界空间中保持相同的相对纹理光束位置（relative texel beam locations）[927, 1227, 1792, 1810]，也就是说，你可以将阴影贴图想象为在整个世界上施加一个二维网格参考系，其中每个网格单元表示贴图上的一个像素样本，当你移动时，阴影贴图将生成这些相同网格单元格的不同集合，换句话说，光线的视角投影被施加到这个网格上，以保持帧与帧的连贯性。

7.4.1 分辨率增强 Resolution Enhancement

雪地上有许多树

描述已自动生成

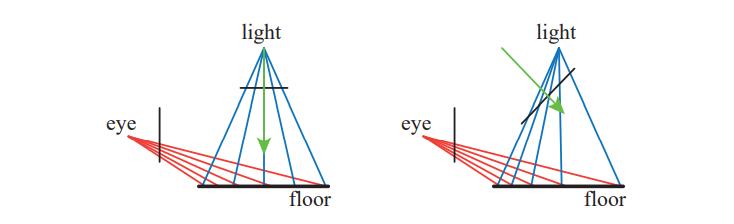
**图 7.15. 左边的图像是使用标准阴影映射创建的，右边的图像使用LiSPSM创建，每个阴影贴图纹素的投影都显示出来了，这两个阴影贴图具有相同的分辨率，不同之处在于LiSPSM改变了光的矩阵，以在更接近观察者的地方提供更高的采样率。（图片由维也纳科技大学的Daniel Scherzer提供。）**

类似于如何使用纹理，理想情况下，我们希望一个阴影贴图纹素覆盖大约一个图像像素，如果我们有一个与眼睛位于相同位置的光源，阴影贴图与屏幕空间像素完美地一对一映射（并且没有可见的阴影，因为光完全照亮眼睛所看到的东西），只要光的方向改变，每像素的比例就会改变，这就会导致伪像，一个示例如图7.15所示。因为前景（foreground）中的大量像素与阴影贴图中的各个纹素相关联，阴影会呈现块状且定义不清，这种不匹配称为透视走样（perspective aliasing）。如果一个表面几乎是边缘对着光线，但面向观察者，则一个阴影贴图纹素也可以覆盖多个像素，这个问题被称为投影走样（projective aliasing）[1792]，参见图7.16，块状化程度（Blockiness）可以通过增加阴影贴图的分辨率来减少，但要付出额外的内存和处理代价。



**图 7.16. 左边的光几乎在头顶上，由于与眼睛的视野相比分辨率较低，阴影的边缘有点粗糙。右边的光线接近地平线，所以每个阴影纹素水平覆盖了相当多的屏幕区域，因此导致了更锯齿状的边缘。（图片由Github上的TheRealMJP的“阴影”程序生成。）**

还有另一种方法创建光的采样图案，使其更接近相机的图案，这是通过改变场景向光投影的方式来实现的。通常我们认为观察视图是对称的，观察向量位于视锥的中心，然而观察方向仅仅定义了一个观察平面，而没有定义哪些像素被采样，定义视锥的窗口可以在这个平面上移动、倾斜或旋转，之后创建一个四边形，可以提供不同的世界空间到观察空间的映射，四边形仍然以规则的间隔采样，这是线性变换矩阵的性质和GPU对它的用法，通过改变光的观察方向和观察窗口的边界，可以修改采样率，见图7.17。



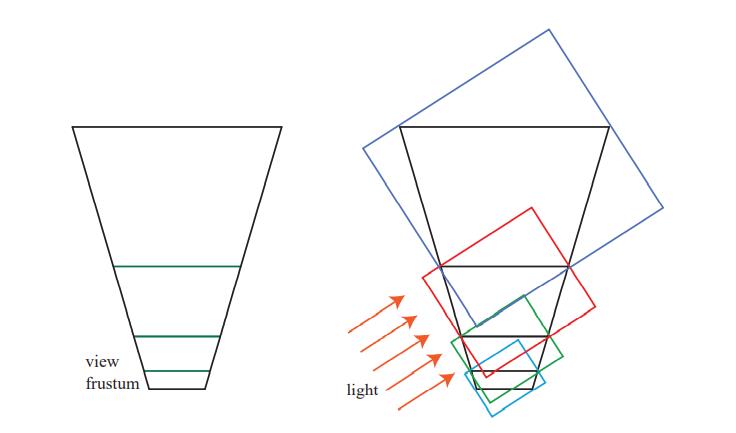
**图 7.17. 对于头顶光，左边地板上的采样频率与眼睛的频率不匹配，右边通过改变光的观察方向和投影窗口，采样频率朝着具有更高纹素密度的方向偏移，以更接近眼睛（译注：的频率）。**

将光的视角映射到眼睛的视角有22个自由度[896]，对这个解空间的探索导致了几种不同的算法，试图更好地匹配光的采样率与眼睛的采样率。方法包括透视阴影贴图（perspective shadow maps, PSM）[1691]，梯形阴影贴图（trapezoidal shadow maps, TSM）[1132]，光空间透视阴影贴图（light space perspective shadow maps, LiSPSM）[1893, 1895]，参见图7.15和图7.26的实例，这类技术往往称为透视变形（perspective warping）方法。

这些矩阵变形算法的一个优点是，除了修改光的矩阵外，不需要额外的工作，每种方法都有自己的优点和缺点[484]，因为每种方法都可以帮助匹配某些几何形状和光照情况下的采样率，不过恶化了其他情况下的采样率。Lloyd等人[1062, 1063]分析了PSM、TSM和LiSPSM之间的等效关系，对这些方法的采样和走样问题给出了完美的概述，当光的方向与观察的方向（例如在头顶）垂直时，这些方案的效果最好，因为这时透视变换可以被移动，使更多的样本更接近眼睛。

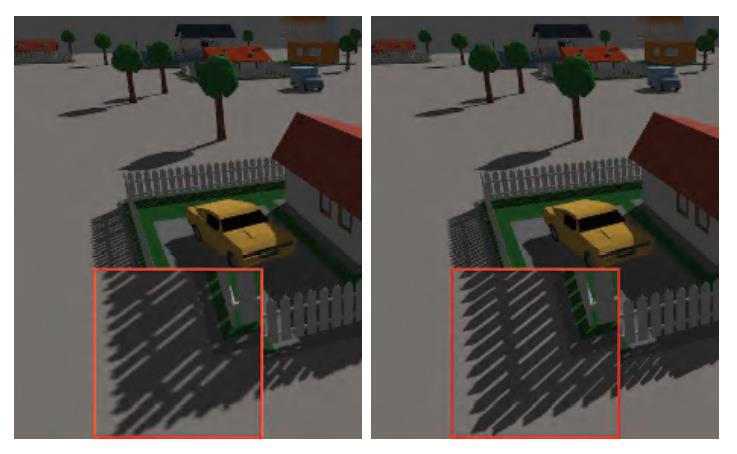
一种光照情况会使矩阵变形技术失效，即当光位于相机前方并指向它时，这种情况被称为视锥决斗（dueling frusta），或者更通俗的说法是“狭路相逢”（deer in the headlights），眼睛附近需要更多的阴影贴图样本，但线性扭曲只会使情况更糟[1555]。这个问题以及其他问题，如质量的突然变化[430]和“紧张的”（nervous），指的是在相机运动过程中产生的不稳定的阴影质量[484, 1227]，使这些方法不受欢迎。

在观察者所在的位置添加更多样本是一个很好的想法，这产生了为给定观察视图生成多个阴影贴图的算法，当Carmack在2004年Quakecon的主题演讲中描述这个方法时，它第一次产生了明显的影响。Blow独立实现了这样的一个系统[174]，想法很简单：生成一组固定的阴影贴图（可能具有不同的分辨率），覆盖场景的不同区域，在Blow的方案中，四个阴影贴图嵌套在观察器周围，这样一来，高分辨率的贴图就可用于附近的物体，而远处的物体分辨率会下降。Forsyth[483, 486]提出了一个相关的想法，为不同的可见对象集合生成不同的阴影贴图，在他的设置中，避免了如何处理跨越两个阴影贴图边界的对象的过渡问题，因为每个对象都有且只有一个关联到的阴影贴图。Flagship Studios开发了融合这两种想法的系统，一个阴影贴图用于附近的动态对象，另一个用于靠近观察者的静态对象的网格部分，还有第三个阴影贴图用于将场景中的静态对象作为一个整体，第一个阴影贴图每帧生成一次，另外两个可以只生成一次，因为光源和几何体是静态的。虽然所有这些特定的系统现在都相当老了，但自那些算法开发以后，针对不同的对象和情况（有些是预先计算的，有些是动态的）使用多个贴图的想法是一个共同的主题。



**图 7.18. 在左侧，来自眼睛的视锥体被分成四个体积。在右边，为体积创建了包围盒，它决定了被方向光的四个阴影贴图中的每一个渲染的体积。（Engel提出[430]。）**

在2006年Engel[430]，Lloyd等人[1062, 1063]和Zhang等人[1962, 1963]独立研究了相同的基本思想（注：Tadamura等人[1735]早在七年前就提出了这个想法，但直到其他研究人员探索了它的用途，它才产生了影响），这个想法是通过平行于观察方向将视锥体的体积切成几块，参见图7.18。随着深度的增加，每个后续体积的深度范围大概是前一个体积的二到三倍[430, 1962]，对于每个观察体积，光源可以生成一个将它紧密绑定的截锥，然后生成阴影贴图。通过使用纹理图集或数组，不同的阴影贴图可以被视为一个大的纹理对象，从而最小化缓存访问延迟，得到的质量改进的对比如图7.19所示。Engel为该算法命名为级联阴影贴图（cascaded shadow maps, CSM），比Zhang的术语并行——分割阴影贴图（parallel-split shadow map）更常用，但两者都出现在文献中，而且实际上是相同的[1964]。



**图 7.19. 在左侧，场景的广阔可视区域导致一个单独的阴影贴图在2048****×2048分辨率下显示透视走样。在右侧，沿着观察轴放置4张1024×1024的阴影贴图显著提高了画质[1963]。插入的红色方框中显示了栅栏前角落的缩放。（图片由香港中文大学的Fan Zhang提供。）**

这种算法易于实现，可以覆盖巨大的场景区域，具有合理的结果，且结果稳定，通过在靠近眼睛的地方以更高的频率采样，可以解决视锥决斗问题，且不会出现严重的最坏情况问题（worst-case problems），由于这些优点，级联阴影映射被用于许多应用程序。

虽然可以使用透视变形将更多的样本打包到单个阴影贴图的细分区域中[1783]，但规范做法是为每个级联使用分开的阴影贴图，如图7.18所示。图7.20显示了从观察者的视角出发，每张贴图覆盖的区域是不同的，更小的观察体积为更近的阴影贴图在它们需要的地方提供更多的样本，确定贴图中的z-深度范围如何被切分——是一项被称为z-分区（z-partitioning）的任务——可能非常简单，也可能非常复杂[412, 991, 1791]。一种方法是对数分区（logarithmic partitioning）[1062]，其中对每个级联贴图，远近平面距离的比例是相同的：

|  |  |
| --- | --- |
|  | (7.5) |

其中n和f是整个场景的近平面和远平面，c是贴图的数量，r是结果比例。例如，如果场景中最近的物体距离为1米，最大距离为1000米，我们有三个级联贴图，那么，最近视图的近远平面距离为1和10，下一个间隔为10到100以保持这个比例，最后一个间隔为100到1000米。初始近平面深度对这种分区有很大的影响，如果最近的深度只有0.1米，那么10000的立方根是21.54，一个相当大的数字，例如，从0.1到2.154到从46.42到1000，这意味着生成的每个阴影贴图必须覆盖更大的区域，降低了精确度，在实践中，这样的分区给了近平面附近的区域相当大的分辨率，如果该区域内没有物体，则是一种浪费。一种避免这种不匹配的方法是将分割距离设置为对数和等距分布的加权混合[1962, 1963]，但如果我们仍能确定场景的紧密视图边界（tight view bounds）就更好了。



**图 7.20. 级联阴影可视化。紫色、绿色、黄色和红色代表从最近到最远的级联。（图片由Unity Technologies提供。）**

此方法的挑战在于设置近平面，如果设置得离眼睛太远，物体可能会被这个平面裁切，这是非常糟糕的伪像，对于画面剪辑，美工可以提前精确地设置这个值[1590]，但对于交互环境来说，这个问题更具挑战性。Lauritzen等人[991, 1403]提出了样本分布阴影贴图（sample distribution shadow maps, SDSM)，它使用前一帧的z深度值，并通过两种方法之一来确定更好的分区。

第一种方法是查看z深度的最小值和最大值，并使用它们设置近平面和远平面，这是通过使用在GPU上被称为reduce的操作来执行的，此过程中，由计算着色器或其他着色器分析一系列越来越小的缓冲区，其输出缓冲区作为输入的反馈，直到剩下一个1×1的缓冲区。通常情况下，这些值会根据场景中物体的移动速度进行一点调整，除非采取纠正措施，否则从屏幕边缘进入的附近物体仍可能造成帧的问题，尽管在下一帧中会很快得到纠正。

第二种方法也分析深度缓冲区的值，制作一个称为直方图（histogram）的图，其记录z深度沿范围分布的情况，除了找到紧密的近平面和远平面之外，在完全没有物体的地方，图中还可能存在空隙。通常添加到此类区域的任意分区平面都可以捕捉到物体实际所在的位置，从而为级联贴图集提供更多的z-深度精度。

在实践中，第一种方法是通用且快速的（通常在每帧1毫秒的范围内），并提供了良好的结果，因此它已在几个应用程序中被采用[1405, 1811]，参见图7.21。



**图 7.21. 深度边界的效果。在左边，没有使用特殊处理来调整近平面和远平面。在右边，使用了样本分布阴影贴图（SDSM）以寻找更紧的边界，注意每个图像的左边缘附近的窗框，二楼的花盒下面的区域，以及一楼的窗户，由于宽松的视图边界导致的欠采样（undersampling）造成伪像。指数阴影贴图（Exponential shadow maps）用于渲染这些特殊的图像，但是提高深度精度的想法对所有阴影贴图技术都是有用的。（图片由Ready at Dawn Studios提供，版权归索尼互动娱乐所有。）**

与单个阴影贴图一样，由于光照样本在帧与帧之间移动而产生的闪烁伪像是一个问题，当对象在级联之间移动时，情况可能会更糟，有多种方法可用来在世界空间中保持稳定的采样点，每种方法都有各自的优点[41, 865, 1381, 1403, 1678, 1679, 1810]。当一个物体跨越两个阴影贴图之间的边界时，阴影的质量会发生突然的变化，一种解决方案是让观察体积稍微重叠，在这些重叠区域采集的样本从两个相邻的阴影贴图中获取结果并进行混合[1791]，或者可以使用抖动（dithering）在该区域采集单个样本[1381]。

由于它的流行（译注：应该指前文中的级联阴影贴图），人们为提高效率和质量付出了相当大的努力[1791, 1964]，如果阴影贴图的截锥体内没有任何变化，则不需要重新计算该阴影贴图。对于每个光源，可以通过寻找哪些物体对于光可见，以及其中哪些可在接收器上投射阴影，来预先计算阴影投射器的列表[1405]。因为很难感知阴影是否正确，所以可以采取一些适用于级联和其他算法的捷径，一种技术是使用低层次的细节模型作为替代物来进行实际的阴影投射[652, 1812]，另一种方法是不考虑很微小的遮挡器[1381, 1811]。理论上更远的阴影贴图的更新频率可能低于每帧一次，因为这样的阴影不那么重要，这个想法冒着由大型移动物体引起的伪像的风险，因此需要小心使用[865, 1389, 1391, 1678, 1679]。Day[329]提出了在帧与帧之间“滚动”远距离贴图的想法，这基于大部分静态阴影贴图在帧与帧之间是可以重复使用的，只有边缘可能会变化，所以需要渲染，像《毁灭战士》（DOOM, 2016)等游戏维持了大量的阴影贴图图集，只有那些物体移动过的贴图才重新生成[294]。更远的级联贴图可以设置为完全忽略动态对象，因为这样的阴影可能对场景贡献不大，在某些环境中，可以使用高分辨率的静态阴影贴图来代替这些更远的级联贴图，可以显著减少工作负担[415, 1590]。稀疏纹理系统（sparse texture system）（第19.10.1节）可用于单个静态阴影贴图的非常庞大的世界[241, 625, 1253]。级联阴影映射可以与烘焙光贴图纹理（baked-in light-map textures）或其他更适合特定情况的阴影技术相结合[652]。Valient的报告[1811]值得注意，因为它描述了各种电子游戏的不同阴影系统定制和技术。第11.5.1节详细讨论了预计算光和阴影的算法。

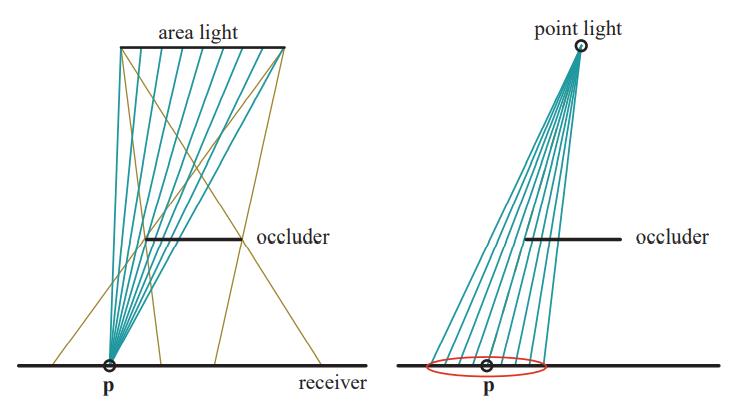
创建几个独立的阴影贴图意味着要为每个贴图遍历一些几何图形，许多提高效率的方法都建立在，在单独的pass中将遮挡器渲染到一组阴影贴图的思想之上。几何着色器可以用来复制对象数据，并将其发送到多个视图[41]，实例化的几何着色器允许一个对象输出到多达32个深度纹理[1456]。多视口扩展可以执行一些操作，比如将对象渲染到特定的纹理数组切片（texture array slice）等[41, 154, 530]，第21.3.1节更详细地讨论了它们在虚拟现实中的使用，视口共享技术的一个可能的缺点是，生成的所有阴影贴图的遮挡器必须沿管线发送，而不是与每个阴影贴图建立联系[1791, 1810]。

你自己现在就在全世界数十亿个光源的阴影中，光只能从其中的几处到达你。在实时渲染中，如果所有的光一直都处于活跃状态，那么包含大量光源的大型场景会被计算量淹没。如果视锥体内部有一块肉眼看不到的体积空间，则被这个接收器体积遮挡的对象不需要被计算[625, 1137]，Bittner等人[152]使用遮挡剔除（occlusion culling）（第19.7节）从眼睛出发寻找所有可见的阴影接收器，然后从光的观察点出发，将所有潜在的阴影接收器渲染到一个模板缓冲区遮板（mask）上，这个遮板编码从光源看来可见的阴影接收器。为了生成阴影贴图，他们使用遮挡剔除来渲染光中可见的物体，并使用遮板来剔除那些接收器没有位于其上的物体。各种剔除策略也可以用于光，由于辐射度（irradiance）随距离的平方递减，一种常见的技术是在一定阈值距离后剔除光源，例如，第19.5节中的门户剔除（portal culling）技术可以找到哪些光影响了哪些单元格，这是一个活跃的研究领域，因为其性能上的优势相当可观[1330, 1604]。

7.5 百分比渐近滤波 Percentage-Closer Filtering

阴影贴图技术的一个简单扩展是可以提供伪软（pseudo-soft）阴影，这种方法还可以帮助改善分辨率问题，即当单个光样本单元格覆盖多个屏幕像素时，会导致阴影看起来是块状的。解决方案与纹理放大类似（第6.2.1节），与从阴影贴图中提取单个样本不同，该方案检索四个最近的样本，该技术并不在深度本身之间，而是它们与表面深度比较的结果之间插值，也就是说，将表面的深度分别与四个纹素的深度进行比较，然后针对每个阴影贴图样本，确定该点在光中还是阴影中，得到的结果即阴影为0，光线为1，然后进行双线性插值，计算光对该表面位置的实际贡献度。这种滤波产生了人为的柔和阴影，这些半影根据阴影贴图的分辨率、相机位置和其他因素而变化，例如，较高的分辨率使得边缘的软化范围更窄，同样的，有一点半影和平滑度总比什么都没有好。

这种从阴影贴图中检索多个样本并将结果混合的想法被称为百分比渐近滤波（percentage-closer filtering, PCF）[1475]。区域光产生软阴影，到达表面某一位置的光的数量是该位置可见光的面积的比例的函数，PCF试图通过反转这个过程来近似一个精确（或定向）光的软阴影，它并不是从一个表面位置开始寻找到光的可见区域，而是从原始位置附近的一组表面位置出发寻找精确光的可见性，参见图7.22。“百分比渐近滤波”这个名字指的是最终的目标，是找到在光下可见的样本的百分比，之后这个比例用于有多少光被用来着色表面。

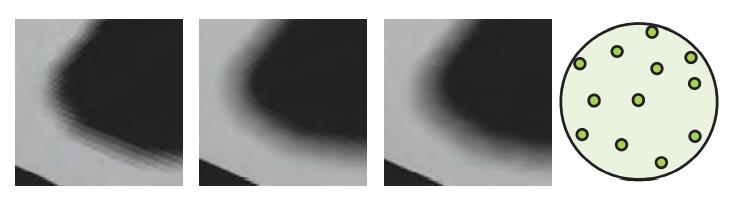


**图 7.22. 在左边，来自区域光源的棕色线显示了半影的形成位置，对于接收器上的单个点p，可以通过测试区域光表面上的一组点来计算接收到的照亮度，并找到没有被任何遮挡器遮挡的点。在右边，一个点光源不会投射出半影，PCF通过反转这个过程：在给定的位置，对阴影贴图上的一个可比较区域进行采样，以得出被照亮样本数量的百分比。来近似区域光的效果。红色椭圆显示阴影贴图上采样的区域，理想情况下，这个圆盘的宽度与接收器和遮挡器之间的距离成正比。**

在PCF中，会在一个表面位置附近，几乎相同的深度，但阴影贴图中不同的纹素位置处生成多个位置信息，每个位置的可见性都被检查，其结果布尔值表示照亮或不照亮，之后混合得到一个软阴影。注意这个过程是非物理的：这个过程不是直接对光源采样，而是依赖于对表面本身采样，到遮挡器的距离不影响结果，所以不同阴影有相似大小的半影，尽管如此，这种方法在许多情况下提供了合理的近似。

一旦确定了要采样区域的宽度，重要的是要以一种避免走样伪像的方式进行采样，对阴影贴图位置的附近进行采样和滤波的方法有许多，变量包括要采样的区域有多宽、要使用多少个样本、哪种采样模式以及如何对结果进行加权。对于功能较差的API，可以通过类似于双线性插值的特殊的纹理采样模式来加速采样过程，该模式访问四个相邻位置，相较于混合结果，将四个样本中的每个样本与给定值进行比较，并返回通过测试的比率[175]。然而，在常规网格模式中执行最近邻采样可能会产生明显的伪像，使用模糊结果但保持对象边缘的联合双边滤波器（joint bilateral filter）可以提高质量，同时避免阴影泄漏到其他表面[1343]，有关这种滤波技术的更多信息，请参见12.1.1节。

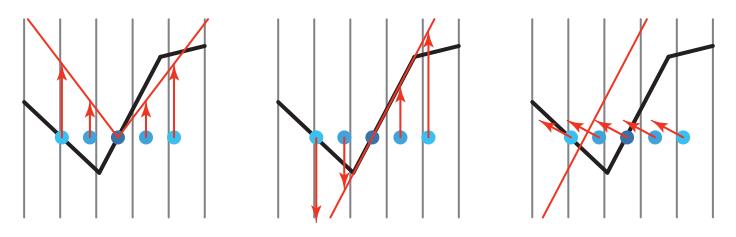
DirectX 10为PCF引入了单指令双线性滤波的支持，提供了更平滑的结果[53, 412, 1709, 1790]，这相比最近邻采样提供了相当大的视觉改进，但常规采样的伪像仍然是一个问题。一种最小化网格图案（译注：这里的网格图案应该是一种伪像）的解决方法是使用预先计算的泊松分布图案（Poisson distribution pattern）对区域进行采样，如图7.23所示，这种分布将样本分散开来，使它们彼此之间既不接近，也不具有规则图案。众所周知对每个像素使用相同的采样位置，无论其分布如何，都会产生图案[288]，这样的伪像可以通过围绕其中心随机旋转样本分布来避免，这将走样转变成噪声。Castaûo[235]发现泊松采样产生的噪声因其平滑性、程式化的内容而特别引人注目，他基于双线性采样提出了一种高效的高斯加权采样方案。



**图 7.23. 最左边显示了4×4网格图案中的PCF采样，使用最近邻采样。最右边显示了磁盘上12点的泊松采样图案，使用此图案对阴影贴图进行采样，可以在左中间的图得到改进后的结果，尽管伪像仍然可见。在右中间的图，在像素与像素之间采样图案围绕其中心随机旋转，结构化的阴影伪像变成（不那么令人反感的）噪声。（图片由ATI研究公司的John Isidoro提供。）**

自阴影问题和光线泄漏，即“粉刺”和“Peter Panning”，会因PCF变得更糟，斜率比例偏差仅仅基于它与光的角度将表面推得远离光，假设样本在阴影贴图上不超过一个纹素，通过从表面上的单一位置在更大范围内进行采样，一些测试样本可能会被真实的表面所阻塞。

人们还发明了一些不同的附加偏差因子，并成功地使用它们来减少自阴影的风险。Burley[212]描述了偏差圆锥（bias cone），其中每个样本以其与原始样本的距离成正比地向光的方向移动，Burley建议斜率为2.0，并带有较小的常数偏差，参见图7.24。



**图 7.24. 附加阴影偏差法。对于PCF，在原始样本位置（五个点的中心）周围采集几个样本，所有这些样本都应该被照亮。在左图中，形成了一个偏置圆锥，样本被移动到它上面，可以增加圆锥的陡度，以使右边的样本足够接近被照亮的地方，这样做冒着其他真正位于阴影中（没有显示）的地方的样本增加光泄露的风险。在中间的图中，所有的样本被调整到位于接收器的平面上，这样做对凸面效果很好，但对凹面适得其反，如左侧所示。在右图中，法向偏移偏置使样品沿表面法线方向移动，与法线和光线夹角的正弦成正比，对于中心的样本，这可以被认为是移动到原表面之上的一个假想表面，这种偏差不仅会影响深度，还会改变用于测试阴影贴图的纹理坐标。**

schüler[1585]，Isidoro[804]和Tuft[1790]提出了基于观察的技术，即接收器本身的斜率应该用来调整其余样本的深度，在这三种方法中，Tuft的公式[1790]最容易用于级联阴影贴图，Dou等人[373]进一步改进和扩展了这一概念，解释了z深度如何以非线性方式变化，这些方法假设附近样本的位置在由三角形构成的同一平面上，称为接收器深度偏差（receiver plane depth bias）或其他类似的术语，这种技术可以在多种情况下相当精确，因为这个假想平面上的位置确实在表面上，或者如果模型是凸的，则在该表面的前面。如图7.24所示，凹面附近的样本可能会被隐藏，常数、斜率比例、接收器平面、视野偏差和法线偏移偏差的组合已经被用来解决自阴影问题，尽管手动调整每个环境仍然是必要的[235, 1391, 1403]。

PCF的一个问题是，由于采样区域的宽度保持不变，阴影将显示出均匀的软阴影，所有的半影宽度都相同，在某些情况下，这是可以接受的，但在遮挡器和接收器之间有地面接触的情况下，这是不正确的，参见图7.25。



**图 7.25. 百分比渐近滤波和百分比渐近软阴影。在左边是带有一点PCF滤波的硬阴影。中间是等宽的软阴影。在右边，在物体与地面接触的地方，有适当硬度的可变宽度的软阴影。（图片由NVIDIA公司提供）**

7.6 百分比渐近软阴影 Percentage-Closer Soft Shadows

2005年Fernando[212, 467, 1252]发表了一种颇具影响力的方法，称为百分比渐近软阴影（percentage-closer soft shadows, PCSS），它试图通过搜索阴影贴图上（译注：一个位置）的附近区域以寻找所有可能的遮挡器，这些遮挡器到该位置的平均距离被用来定义样本区域宽度：

|  |  |
| --- | --- |
|  | (7.6) |

其中是接收器到光的距离，是遮挡器的平均距离。换句话说，当平均遮挡器距离接收器越来越远，以及离光越来越近时，样本表面区域的宽度就会增加。请查看图7.22，并思考移动遮挡器的效果，看看这是如何发生的，图7.2、7.25和7.26显示了相关示例。

如果一个地方没有找到遮挡器，则该位置是完全照亮的，没有进一步处理的必要，类似地，如果位置完全被遮挡，处理也会结束，否则，对感兴趣的区域进行采样，并计算光的近似贡献。为了节省处理成本，可以使用样本区域的宽度来改变采样的数量，也可以应用其他技术，例如，对不太可能重要的远距离软阴影使用较低的采样率。

这种方法的一个缺点是，它需要在阴影贴图的一个相当大小的区域采样以找到遮挡器，使用旋转泊松圆盘图案可以帮助隐藏欠采样的伪像[865, 1590]，Jimenez[832]注意到泊松采样在运动下可能是不稳定的，并发现通过使用介于抖动和随机之间的中间函数形成的螺旋模式能在帧与帧之间提供更好的结果。

Sikachev等人[1641]详细讨论了使用SM5.0中的特性更快地实现PCSS的方法，该方法由AMD引入，通常称为接触点硬化阴影（contact hardening shadows, CHS）。这个新版本还解决了基本PCSS的另一个问题：半影的大小受阴影贴图分辨率的影响，参见图7.25，通过首先生成阴影贴图的mpmap，然后选择最接近用户定义的世界空间内核大小的mip级别，可以将这个问题最小化，只需要16个GatherRed()函数的纹理调用，采样一个8×8的区域来寻找平均阻塞器（blocker，译注：这个词之前未出现过，可能是遮挡器同义词）深度。一旦找到预估计的半影，则将更高分辨率的mip级别用于阴影的锐利区域，而低分辨率的mip级别用于较软的区域。

CHS已被用于大量电子游戏中[1351, 1590, 1641, 1678, 1679]，且研究仍在继续。例如，Buades等人[206]提出了可分离软阴影映射（separable soft shadow mapping, SSSM），将采样网格的PCSS过程分割为可分离的部分，并在像素之间尽可能重用元素（element）。

事实证明，分层的最小/最大阴影贴图（min/max shadow map）有助于加速每个像素需要多个样本的算法，虽然通常无法平均阴影贴图的深度，但每个mipmap级别的最小值和最大值是有用的，也就是说，可以生成两个mipmap，一个保存每个区域（有时称为HiZ）中找到的最大z-深度，另一个保存最小z-深度。给定纹素位置、深度和要采样的区域，mipmap可以用于快速确定完全光照和完全阴影的情况，例如，如果纹素的z-深度大于mipmap对应区域存储的最大z-深度，那么纹素必定在阴影中——不需要进一步的采样，这种类型的阴影贴图使得确定光可见性的任务更加高效[357, 415, 610, 680, 1064, 1811]。

像PCF这样的方法是通过采样接收器附近的位置来工作的，PCSS的工作原理是寻找遮挡器附近的平均深度，这些算法不直接考虑光源的面积，而是采样附近的表面（译注：可能指的是接收器遮挡器附近），并受到阴影贴图分辨率的影响。PCSS背后的一个主要假设是，平均阻塞器是半影大小的合理估计，当两个遮挡器，比如一盏路灯和一座远山，在一个像素处部分遮挡同一个表面时，这个假设就被打破了，并会导致伪像。理想情况下，我们想确定从一个单独的接收器位置出发有多少区域光源可见，一些研究人员已经探索了使用GPU的反向投影（backprojection），想法是将每个接收器的位置视为视点，将区域光源视为观察平面的一部分，并将遮挡器投影到该平面上。Schwarz、Stamminger[1593]和Guennebaud等人[617]总结了前人的工作，并提出了自己的改进。Bavoil等人[116]采用了不同的方法，使用深度剥离来创建多层阴影贴图。反向投影算法可以得到很好的结果，但每像素的高成本（到目前为止）意味着它们还没有在交互式应用程序中得到采用。

7.7 滤波阴影贴图 Filtered Shadow Maps

Donnelly和Lauritzen提出了一种允许对生成的阴影贴图进行滤波的算法，叫做方差阴影贴图（variance shadow map, VSM）[368]，该算法将深度存储在一个贴图中，将深度的平方存储在另一个贴图中，在生成贴图时，可以使用MSAA或其他反走样方案。这些贴图可以进行模糊化处理、mipmap处理、放入求和区域表[988]等，将这些贴图视为可过滤的纹理是一个巨大的优势，因为在从这些贴图中检索数据时，可以使用整个采样和滤波技术合集。

我们将在这里深入描述VSM，以了解这个过程是如何工作的，同时，这类算法中的所有方法都使用了相同类型的测试，有兴趣进一步了解这一领域的读者应该访问相关参考资料，我们还推荐Eisemann等人[412]的书，它为这个主题提供了相当大的空间。

首先，VSM在接收器的位置采样深度贴图（仅一次），以返回最近的光遮挡器的平均深度，当这个称为第一矩（first moment）的平均深度大于阴影接收器t的深度时，接收器被认为完全在光中，当平均深度小于接收器深度时，使用下面的公式：

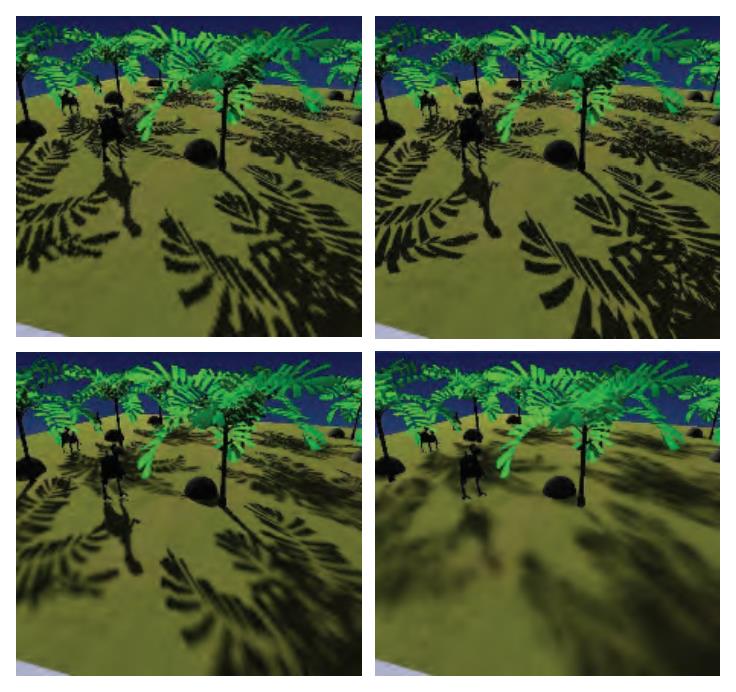
|  |  |
| --- | --- |
|  | (7.7) |

其中是光照下样本的最大百分比，为方差，t为接收器深度，为阴影贴图的平均期望深度，深度平方阴影贴图的样本，称为第二矩（second moment），用于计算方差：

|  |  |
| --- | --- |
|  | (7.8) |

是接收器可见性百分比的上限，实际照度百分比不能大于这个值，这个上限来自切比雪夫不等式（Chebyshev’s inequality）片面的变体（one-sided variant），该方程试图利用概率论来估计表面位置上分布的遮挡器有多少超出了表面与光的距离，Donnelly和Lauritzen表明，对于固定深度的平面遮挡器和平面接收器，，因此公式7.7可以很好地近似许多真实的阴影情况。

Myers[1251]建立了一种关于为什么这种方法有效的直觉，在阴影边缘区域的方差会增加，深度的差异越大，方差也越大。项则是可见性百分比的显著决定因素，如果这个值略高于零，意味着平均遮挡器深度比接收器稍微更接近光，而此时接近1（表示完全照亮），这将发生在半影光线充足的边缘。朝半影移动，平均遮挡器深度越来越接近光，所以这个项变大而下降，与此同时，在半影内方差本身也在变化，沿边缘从几乎为零到最大方差，在这里虽然遮挡器的深度不同，但它们平分区域，这些项相互抵消，在半影上形成线性变化的阴影，与其他算法的比较见图7.26。

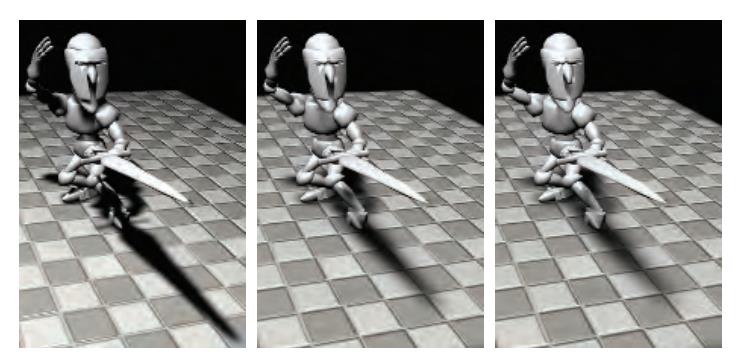


**图 7.26. 左上角，标准阴影映射。右上角，透视阴影映射，增加了阴影贴图在观察者附近的纹素密度。左下角，百分比渐近软阴影，随着遮挡器与接收器距离的增加，阴影变得柔和。右下角，具有恒定软阴影宽度的方差阴影映射，每个像素使用单一的方差贴图样本阴影处理。（图片由Nico Hempe, Yvonne Jung和Johannes Behr提供。）**

方差阴影映射的一个显著特点是它能以一种优雅的方式处理几何引起的表面偏差问题，Lauritzen[988]推导了如何使用表面的斜率来修改第二矩的值，偏差和其他由数值稳定性导致的问题对方差映射来说可能都是难题。例如，公式7.8用一个较大的值减去另一个相似的值，这种类型的计算往往会放大基础数值表示的精度不足问题，使用浮点纹理有助于避免这个问题。

总的来说，由于GPU优化的纹理功能得到了有效的利用，VSM因其处理时间花费（较少）从而显著的提高了质量。比起PCF在生成更柔和的阴影时，为了避免噪声，需要更多的样本，以及因此需要更多的时间，VSM可以仅用一个单独的、高质量的样本来决定整个区域的效果，并生成平滑的半影，这种能力意味着在算法的限制下，阴影可以被任意地变软，而不需要额外的代价。

与PCF一样，滤波核的宽度决定了半影的宽度，通过寻找接收器和最近的遮挡器之间的距离，可以改变核的宽度，从而提供令人信服的软阴影，mipmap操作后的样本对宽度缓慢增加的半影的覆盖率估算较差，会产生盒装伪像，Lauritzen[988]详细介绍了如何使用和面积表来提供更好的阴影，一个示例如图7.27所示。



**图 7.27. 方差阴影映射，其中到光源的距离从左到右增加。（图片来自NVIDIA SDK 10 [1300]示例，由NVIDIA公司提供。）**

一个方差阴影映射失效的地方是沿着半影区域，当两个或多个遮挡器覆盖一个接收器，且一个遮挡器靠近接收器时，从概率论得出的切比雪夫不等式将产生一个与正确光照百分比无关的最大光值，最接近的遮挡器只部分隐藏了光，使方程的近似不成立，这就导致了光出血（light bleeding，又称光泄漏）现象，即完全被遮挡的区域仍然能接收到光，参见图7.28。通过在更小的区域取更多的样本，可以将方差阴影映射转化为PCF的一种形式，使得这个问题可以得到解决，与PCF一样，速度和性能是需要权衡的，但对于阴影深度复杂度低的场景，方差映射工作得很好。Lauritzen[988]提出了一种美工控制的方法来改善这个问题，即将低百分比（的区域）视为完全的阴影，并将其余百分比范围重新映射为0%到100%，这种方法使光出血变暗，代价是缩小了整个半影。即便如此轻微出血也是严重的限制，VSM适合从地形中产生阴影，因为这样的阴影很少涉及多个遮挡器[1227]。



**图 7.28. 在左边，方差阴影映射应用于一个茶壶。在右边，一个（没有显示的）三角形在茶壶上投下阴影，导致在地面的阴影中产生了令人反感的伪像。（图片由Marco Salvi提供。）**

能够使用滤波技术快速生成光滑阴影的承诺，使人们对滤波阴影映射产生了很大的兴趣，主要的挑战是解决各种流血问题。Annen等人[55]引入了卷积阴影贴图（convolution shadow map），扩展了Soler和Sillion的平面接收器算法[1673]背后的思想，该思想是用傅里叶展开（Fourier expansion）来编码阴影深度，与方差阴影映射一样，这样的贴图可以被滤波，该方法收敛于正确答案，从而减少了漏光问题。

卷积阴影映射的一个缺点是需要计算和访问多个项，这大大增加了执行和存储成本[56, 117]。Salvi[1529, 1530]和Annen等人[56]同时独立地提出了使用基于指数函数的单一项的想法，这种方法称为指数阴影贴图（exponential shadow map, ESM）或指数方差阴影贴图（exponential variance shadow map, EVSM），它将深度的指数及其第二矩保存到两个缓冲区中。指数函数更接近于阴影贴图执行的步长函数（例如，在光下或不在光下），所以它可以显著减少流血伪像，它避免了卷积阴影映射存在的另一个称为振铃（ringing）的问题，即略微超过原始遮挡器深度的特定深度会发生轻微的光泄漏。

存储指数值的一个限制是，第二矩的值可能变得非常大，因此超出了浮点数的使用范围。为了提高精度，并允许指数函数更急剧下降，可以生成z深度以使其为线性[117, 258]。

由于指数阴影贴图方法相比VSM提高了质量，且与卷积贴图相比，它的存储容量更小，性能更好，因此在三种滤波方法中它最受关注。Pettineo[1405]注意到一些其他改进，例如使用MSAA来改善结果和获得一些有限的透明度，他还描述了如何使用计算着色器来提高滤波性能。

最近，由Peters和Klein提出了矩阴影映射（moment shadow mapping）[1398]，它提供了更好的质量，尽管是以使用四个或更多的矩（moment）为代价，增加了存储成本，这种成本可以通过使用16位整数来存储矩来降低。Pettineo[1404]实现了这种新方法，并将其与ESM进行了比较，提供了一个探索多种变体的代码库。

级联阴影贴图技术可应用于滤波贴图以提高精度[989]，与标准级联贴图相比，级联ESM的一个优点是可以为所有级联设置单一偏置因子[1405]，Chen和Tatarchuk[258]详细讨论了各种光泄漏问题以及级联ESM遇到的其他伪像，并提出了一些解决方案。

滤波后的贴图可以被认为是PCF的一种廉价形式，它需要的样本很少，像PCF一样，这样的阴影具有固定的宽度，这些滤波方法都可以与PCSS结合使用，以提供可变宽度的半影[57, 1620, 1943]，矩阴影映射的扩展还包括提供光散射和透明效果的能力[1399]。

7.8 体积阴影技术 Volumetric Shadow Techniques

透明物体会衰减并改变光的颜色，对于某些透明对象集合，可以使用类似于第5.5节中讨论的技术来模拟这些效果，例如，在某些情况下，可以生成第二种类型的阴影贴图，将透明对象渲染到它上面，并存储最接近的深度和颜色或alpha覆盖率。如果接收器没有被不透明阴影贴图遮挡住，则测试透明深度贴图，如果遮挡，则根据需要检索颜色或覆盖率[471, 1678, 1679]。这个想法让人联想到第7.2节中的阴影和光的投影，存储的深度避免投影到透明物体和光之间的接收器上，这样的技术不能应用于透明对象本身。

自阴影对于像头发和云这样的物体的真实渲染是至关重要的，这些物体要么很小，要么是半透明的，单深度阴影贴图不适用于这些情况，Lokovic和Veach[1066]首先提出了深度阴影贴图（deep shadow maps）的概念，在这个概念中，每个阴影贴图纹素存储了一个描述光如何随深度衰减的函数，这个函数通常由一系列不同深度的样本近似，每个样本都有一个不透明度值。贴图中包含给定位置深度的两个样本用于查找阴影的效果，GPU面临的挑战是如何高效地生成和计算这些函数，这些算法使用了与顺序无关的透明算法（第5.5节）类似的方法，并遇到了一些与该算法遇到的挑战类似的挑战，例如如何正确地表达每个函数所需数据的紧凑存储方式。

Kim和Neumann[894]首先提出了一种基于GPU的方法，他们称之为不透明度阴影贴图（opacity shadow maps），只存储不透明度的贴图是在一个固定深度集上生成的。Nguyen和Donnelly[1274]给出了这种方法的更新版本，生成了如图17.2所示的图像，然而，深度切片都是平行且一致的，因此需要许多切片来隐藏由于线性插值而产生的切片之间的不透明度伪像。Yuksel和Keyser[1953]通过创建更接近模型形状的不透明度贴图来提高效率和质量，这样做可以减少所需的层数，因为每一层的计算对最终图像更重要（译注：这句话应该是质量比数量重要的意思）。

人们为了避免依赖于固定的切片设置，提出了更多的自适应技术。Salvi等人[1531]引入了自适应体积阴影贴图（adaptive volumetric shadow maps），其中每个阴影贴图纹素都存储了不透明度和图层深度，在光栅化时使用像素着色器操作对数据流（表面不透明度）进行有损压缩（lossily compress），这避免了用无法估计的内存来收集所有样本并在一个集合中处理它们，该技术类似于深度阴影贴图[1066]，但压缩步骤是在像素着色器中动态完成的。将函数表示限制在一个较小的、固定数量的存储不透明度/深度的对（pair）上，可以使GPU上进行的压缩和检索更加高效[1531]，这种方法的成本比简单的混合要高，因为需要读取、更新和写回曲线，成本取决于用以表示曲线的点的数量，在这种情况下，该技术还要求支持UAV和ROV功能的最新硬件（第3.8节末尾），一个相关示例参见图7.29。

图片包含 游戏机

描述已自动生成

**图 7.29. 用自适应体积阴影贴图渲染头发和烟雾[1531]。（经Marco Salvi和Intel公司授权转载，版权归Intel公司所有，2010年。）**

在游戏GRID2中使用自适应体积阴影映射进行真实烟雾渲染，其平均成本低于2毫秒/帧[886]。Fürst等人[509]描述并提供了用于视频游戏的深度阴影贴图实现的代码，他们使用链表来存储深度和alpha值，并使用指数阴影映射来提供光照区域和阴影区域之间的软过渡。

阴影算法的探索仍在继续，各种算法和技术的结合变得越来越普遍。例如，Selgrad等人[1603]研究了使用链表存储多个透明样本，并使用带有分散写入（scattered writes）功能的计算着色器来构建贴图，他们的工作使用了深度阴影贴图的概念，以及经过滤波的贴图和其他元素，这为生成高质量的软阴影提供了一个更通用的解决方案。

7.9 不规则Z-缓冲区阴影 Irregular Z-Buffer Shadows

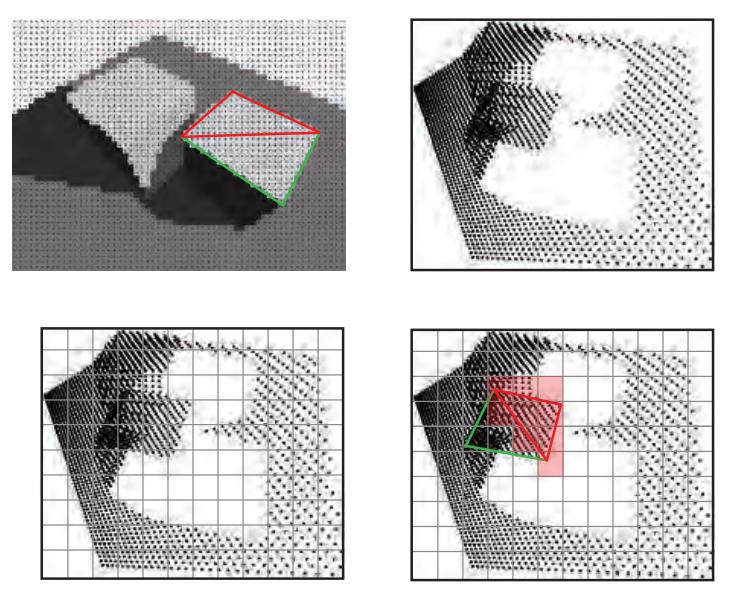
各种各样的阴影贴图方法有几个受欢迎的原因，它们的成本是可预测的，并且很好的适应不断增加的场景规模，最坏的情况是与图元的数量成线性，它们很好地映射到GPU上，因为它们依赖光栅化来均匀地采样光的世界视图。然而，由于这种离散采样，问题出现了，因为眼睛看到的位置与光看到的位置不是一一对应的，当光对表面的采样频率低于眼睛时，就会出现各种各样的走样问题，即使在采样率相当的情况下，也存在由于表面采样的位置与眼镜看到的位置略有不同而导致的偏差问题。

阴影体积提供了一个精确的、可分析的解决方案，因为光与表面的相互作用产生了一系列三角形，定义了任意给定的位置是被照亮还是在阴影中，在GPU上实现时算法不可预测的成本是一个严重的缺点，近年来探索的改进[1648]是诱人的，但目前还没有证据表明有商业应用采用这些方法。

从长远来看，另一种可分析的阴影测试方法可能具有潜力：光线追踪（ray tracing），将在第11.2.2节中详细描述，其基本思想非常简单，特别是对于阴影，从接收器位置发射一束射线到光源，如果发现有任何阻挡光线的物体，接收器就处于阴影中。快速射线跟踪器的大部分代码都专门用于生成和使用分层数据结构，以最小化每条射线所需测试的对象数量，每帧为动态场景构建和更新这些结构是一个几十年的老话题，也是一个持续研究的领域。

另一种方法是使用GPU的光栅化硬件来观察场景，但不仅仅存储z深度，而是存储关于光的每个网格单元中遮挡器边缘的附加信息[1003, 1607]，例如，设想在每个阴影贴图纹素中存储与网格单元格重叠的三角形列表，这样的列表可以通过保守光栅化（conservative rasterization）来生成，在保守光栅化中，如果三角形的任意部分与像素重叠，就会生成一个片元，而不像平常一样仅重叠像素的中心时才生成（第23.1.2节）。这种方案的一个问题是通常需要限制每个纹素的数据量，这反过来会导致不能准确地确定每个接收器位置的状态，考虑到现代GPU的链表原理[1943]，每个像素存储更多数据是可能的，但是，除了物理内存限制，在每个纹素的列表中存储不同数量数据的一个问题是，GPU处理会变得非常低效，因为单个warp可能包含几个需要检索和处理许多项的片元线程，而其余的线程是空闲的，没有工作要做。构造一个避免由于动态“if”语句和循环而导致线程发散的着色器对性能至关重要。

除了在阴影贴图中存储三角形或其他数据并针对它们测试接收器位置之外，还有一种选择是翻转问题，存储接收器位置，然后针对每个接收器位置测试三角形，这种保存接收器位置的概念首先由Johnson等人[839]和Aila和Laine[14]探索，称为不规则z缓冲区（IZB），这个名字有点误导人，因为缓冲区本身有正常的、规则的阴影贴图形状，相反，缓冲区的内容是不规则的，因为每个阴影贴图纹素会有一个或多个接收器位置存储其中，或者可能什么都没有，参见图7.30。



**图 7.30. 不规则z-缓冲区。在左上角，从眼睛看到的视图在像素中心产生了一组点，图示两个三角形组成一个立方体面。在右上角，从光线的视角显示这些点。在左下方，施加了一个阴影贴图网格，在每个纹素上，将生成其网格单元格中的所有点的列表。在右下方，通过保守光栅化对红色三角形进行阴影测试，在每个接触到的纹素（显示为浅红色）上，列表中的所有点都要根据光线来测试三角形的可见性。（底层的光栅图像由Timo Aila和Samuli Laine提供[14]。）**

可以使用由Sintorn等人[1645]和Wyman等人[1930, 1932]提出的方法，一种多pass算法创建IZB，并测试其内容在光的观察下的可见性。首先，从眼睛出发渲染场景，以找到眼睛看到的表面的z深度，这些点被转换为光线的场景视图，并由此设置光的截锥体的紧密边界，然后这些点被存入光的IZB中，每个点被放入对应纹素的列表中，注意有些列表可能是空的，在那些眼睛看不到任何表面但光线可以看到的空间。遮挡器被保守光栅化到光的IZB来确定是否有任何点被隐藏，如果有的话也因此在阴影中，保守光栅化可以确保，即使一个三角形没有覆盖一个光照纹素的中心，它也会对可能重叠的点进行测试。

可见性测试发生在像素着色器中，测试本身可被可视化为一种光线追踪，从图像上点的位置发出射线到光，如果一个点在三角形内部并且比三角形所在的平面更远，那么它被隐藏，一旦所有遮挡器都被光栅化，光的可见性结果将被用来着色表面，这种测试也称为截锥追踪（frustum tracing），因为该三角形可被视为，定义一个视锥体用于检查包含在其体积中的点（译注：最后一句直译，没懂什么意思）。

要使这种方法在GPU上很好地工作，仔细的编码是至关重要的，Wyman等人[1930, 1932]注意到他们的最终版本比最初原型快了两个数量级，性能提升的部分原因是算法的直接改进，比如剔除表面法线背向光线（因此总是没有光照）的图像点，以避免为空纹素生成片元。其他性能提升来自于GPU数据结构的改进，以及通过在每个纹素中使用短且长度相似的点列表来最小化线程发散，图7.30显示了一个具有较长列表的低分辨率的阴影贴图，用于说明的目的。理想情况是每个列表一个图像点，更高的分辨率提供更短的列表，但遮挡器生成的用于计算的片元的数量也越多。

从图7.30的左下图中可以看出，由于透视效果的影响，左边地平面上可见点的密度要比右边高得多，使用级联阴影贴图有助于降低这些区域的列表大小，通过将更多的光照贴图分辨率聚集到眼睛附近。

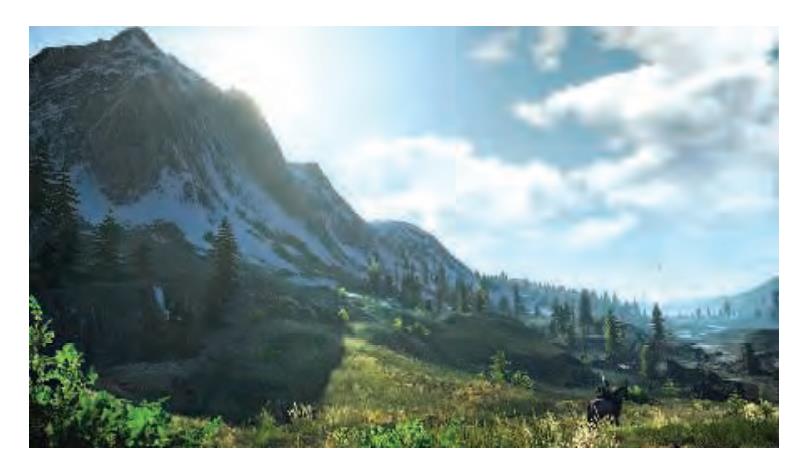
这种方法避免了其他方法的采样和偏差问题，并提供了非常清晰的阴影，出于审美和感知的原因，通常需要柔和的阴影，但附近的遮挡器可能有偏差问题，如Peter Panning。Story和Wyman[1711, 1712]探索了混合阴影技术，核心思想是根据遮挡器距离来混合IZB和PCSS阴影，当遮挡器很近时使用硬阴影结果，当遮挡器较远时使用软阴影结果，见图7.31。阴影质量对于附近的物体通常是最重要的，所以可以只在选定的子集上使用IZB技术来降低成本，该解决方案已成功应用于电子游戏中，本章就从这样的一个图像开始，如图7.2所示。



**图 7.31. 在左侧，PCF为所有对象提供了一致均匀的软阴影。在中间，PCSS根据到遮挡器的距离软化阴影，但树枝阴影重叠在板条箱的左上角产生伪像。在右边，来自IZB的清晰阴影与来自PCSS的柔和阴影混合，得到了改进的结果[1711]。（图片来自育碧的“全境封锁”。）**

7.10 其它应用 Other Applications

将阴影贴图视为一定体积的空间，将光明与黑暗分开，也有助于确定物体的哪些部分在阴影中，Gollent[555]描述了CD Projekt的地形阴影系统如何为每个区域计算被遮挡的最大高度，该高度不仅可以用于地形阴影，还可以用于场景中的树木和其他元素的阴影。为了找到每个高度，可为太阳渲染可见区域的阴影贴图，然后从太阳出发检查每个地形高度场位置的可见性，如果在阴影中，可以通过以固定步长增加世界高度，直到太阳进入视野，之后执行二分查找，以此来估计太阳首次看到这个位置的高度。换句话说，我们沿着一条垂线前进，并迭代缩小它与阴影贴图上区分光和暗的表面相交的位置，相邻的高度被插值以寻找任意位置的该遮挡高度。这种技术用于地形高度场的软阴影的一个例子可以在图7.32中看到，我们将在第14章中看到更多的光线行进（ray marching）穿过光和暗区域的使用。



**图 7.32. 为每个高场位置计算最先看到太阳的高度来照明地形，注意沿着阴影边缘的树木是如何恰当地产生阴影的[555]。（CD PROJEKT®，Witcher®是CD PROJEK资本集团的注册商标，巫师游戏CD PROJEKT S.A.由CD PROJEKT S.A.开发，保留所有权利。《巫师》游戏基于Andrzej Sapkowski的小说，所有其他版权和商标均为其各自所有者的财产。）**

最后一个值得一提的方法是渲染屏幕空间阴影（screen-space shadows），阴影贴图通常不能在小的特征上产生精确的遮挡，因为它们的分辨率有限，这在渲染人脸时尤其成问题，因为我们特别容易注意到人脸上的任何视觉伪像，例如，渲染发光的鼻孔（如果不是故意的）看起来不和谐。虽然使用更高分辨率的阴影贴图或仅使用感兴趣区域的单独阴影贴图会有帮助，但另一种可能是充分利用现有的数据，在大多数现代渲染引擎中，来自更早的预渲染pass的相机视角的深度缓冲，在渲染过程中可用，存储在其中的数据可以被视为一个高度场，通过对这个深度缓冲区进行迭代采样，我们可以执行光线行进过程（第6.8.1节），并检查朝向光的方向是否不被遮挡。虽然成本很高，因为它涉及重复采样深度缓冲区，但这样做可以为剧情动画中的特写提供高质量的结果，在这种情况下，花费额外的毫秒通常是合理的，该方法由Sousa等人提出[1678]，并在今天的许多游戏引擎中普遍使用[384, 1802]。

总结整个章节，某种形式的阴影映射是迄今为止用于将阴影投射到任意表面形状上的最常用算法。当阴影投射在大面积区域，如室外场景时，级联阴影贴图可以提高采样质量；求近平面的最大距离时通过SDSM可以进一步提高精度；百分比渐近滤波（PCF）使阴影更柔软；百分比渐近软阴影（PCSS）及其变体使接触处变硬，不规则z缓冲区可以提供精确的硬阴影；滤波的阴影贴图提供了快速的软阴影计算，且当遮挡器远离接收器（如地形）时效果得特别好；最后，屏幕空间技术可以用于额外的精确度，尽管要付出明显的成本。

在本章中，我们聚焦于当前应用程序中使用的关键概念和技术，每一种都有其优势，如何选择取决于世界的大小、内容组成（静态内容vs动画）、材质类型（不透明、透明、头发或烟雾），和光照的数量和类型（静态或动态；局部或远光；点光、聚光或区域光），以及诸如基础纹理能否隐藏任意伪像等因素。GPU的能力不断发展和提高，因此我们期待在未来几年继续看到与硬件有良好映射关系的新算法出现，例如，第19.10.1节中描述的稀疏纹理技术已应用于阴影贴图存储以提高分辨率[241, 625, 1253]；Sintorn，Kämpe和其他人[850, 1647]提出了一种创造性的方法，探索将用于光照的二维阴影贴图转换为三维体素集（voxel，小盒子状，参见第13.10节）的想法，使用体素的一个优点是它可被归类为照亮或处于阴影，因此需要最小存储空间，高度压缩的稀疏体素八叉树表示法可存储具有大量光照和静态遮挡器的阴影；Scandolo等人[1546]将他们的压缩技术与使用双重阴影贴图的基于区间的方案结合起来，提供了更高的压缩率；Kasyan[865]使用体素圆锥追踪（voxel cone tracing，第13.10节）从区域光中生成柔和的阴影，一个例子参见图7.33，更多的圆锥追踪阴影如图13.33所示。



**图 7.33. 顶部是用基础的软阴影近似生成的图像。底部是在一个体素化的场景，使用圆锥追踪的基于体素的区域光阴影，请注意汽车的阴影更加分散，照明也因一天中时间的变化而不同。（图片由Crytek提供[865]。）**

进一步阅读和资源 Further Reading and Resources

本章的重点是阴影算法的基本原理和它需要什么样的质量——可预测的质量和性能——才对交互式渲染有用，我们避免对渲染的这一领域的研究进行详尽的分类，因为有两篇文章已解决了这个问题，Eisemann等人[412]的《实时阴影》（Real-Time Shadows）一书直接关注交互渲染技术，讨论了大量的算法及其优点和成本，SIGGRAPH 2012课程提供了本书的摘录，同时还增加了对较新作品的引用[413]，可以从他们的网站www.realtimeshadows.com上获得SIGGRAPH 2013课程的演示文稿。Woo和Poulin的《阴影算法数据矿工》（Shadow Algorithms Data Miner）[1902]一书概述了用于交互式和批处理渲染的多种阴影算法。两本书都提供了该领域数百篇研究文献的参考。

Tuft的两篇文章[1791, 1792]对常用阴影映射技术及其涉及的问题进行了极好的概述。Bjørge[154]提出了一系列适用于移动设备的流行阴影算法，以及比较各种算法的图像。Lilley的报告[1046]给出了实用阴影算法的坚实而广泛的概述，重点是用于GIS系统的地形渲染。Pettineo[1403,1404]和Castaño[235]的博客文章因它们的实用提示和解决方案，以及演示代码库特别有价值。Scherzer等人[1558]简短总结了专门针对硬阴影的工作。Hasenfratz等人[675]对软阴影算法的研究已经过时，但在一定深度上涵盖了广泛的早期工作。

第8章 光照和颜色 Light and Color

“会拆开彩虹，正像它不久前曾经使身体柔弱的拉弥亚化为一道虚影。”

——John Keats

（译注：拉弥亚，Lamia，希腊神话中人首蛇身的女妖，济慈这句诗的意思应该指的是哲学破坏了美感。）

前几章中讨论的许多RGB颜色值表示光的强度和着色，在本章中，我们将通过测量这些值学习各种物理光量，为后续章节奠定基础，后续章节将从更基于物理的角度讨论渲染。我们还将更多地了解渲染过程中经常被忽视的“下半场”：将代表场景线性光量的颜色转换为最终显示颜色。

8.1 光量 Light Quantities

任何基于物理的渲染方法的第一步都是以精确的方式量化光，（基于此）首先提出了辐射度量学（Radiometry），这是关于光的物理传输的核心领域。我们接着讨论光度测定学（photometry），它处理由人眼灵敏度加权的光值，我们对颜色的感知是一种精神物理学（psychophysical）现象：对物理刺激的心理感知，我们将在比色法一节中讨论颜色感知。最后，我们讨论了用RGB颜色值进行渲染的有效性。

8.1.1 辐射度量学 Radiometry

辐射度量学研究的是电磁辐射量的测量，正如我们将在9.1节中详细讨论的那样，这种辐射以波的形式传播，不同波长（wavelengths，具有相同相位的两个相邻点之间的距离，例如两个相邻的波峰）的电磁波——往往具有不同的特性，在自然界中，电磁波存在巨大的波长范围，从长度不足一纳米的百分之一的伽马波到长达数万公里的极低频无线电波（extreme low frequency , ELF）。人类所能看到的波只占这一范围的一小部分，从大约400纳米的紫色光到700多纳米的红色光，参见图8.1。

图片包含 文本

描述已自动生成

**图 8.1. 可见光的波长范围，显示在完整的电磁波谱中。**

辐射量的存在是为了测量电磁辐射的各个方面：总能量、功率（随时间变化的能量）和相对于面积、方向或两者的功率密度，表8.1总结了这些量。

表格

描述已自动生成

**表 8.1. 辐射量和单位。**

在辐射度量学中，基本单位是辐射通量（radiant flux）Φ，辐射通量是辐射能量在时间-功率上的流动，计量单位为瓦（watts, W）。

辐照度（Irradiance）是辐射通量相对于面积的密度，即，辐照度是相对一个区域来定义的，这个区域可能是空间中一个想象的区域，通常是物体的表面，它的单位是瓦特每平方米。

在介绍下一个量之前，我们需要先介绍立体角（solid angle）的概念，它是角度概念的三维扩展。角度可以被认为是一个平面上连续方向的集合大小的度量，其弧度值等于这个方向集合与半径为1的封闭圆相交的圆弧的长度，类似地，立体角测量三维空间中一个连续方向集合的大小，用立体弧度（steradians，缩写为“sr”）来表示，立体弧度由半径为1的封闭球体上（与连续方向集合）相交的块的面积定义[544]，立体角用符号ω表示。

在二维空间中，2π弧度的角覆盖了整个单位圆，将其扩展到三维空间，4π立体弧度的立体角将覆盖单位球的整个面积，一个立体角的立体弧度的大小如图8.2所示。

卡通画

描述已自动生成

**图 8.2. 从球体的剖面图中去掉一个表示立体角的立体弧度的圆锥，其形状本身与测量无关，球体表面的覆盖才是关键。（译注：这里的意思应该是不要关注里面是圆锥还是什么，最主要的是圆锥的底面，即表面。）**

现在我们可以介绍辐射强度（radiant intensity）I，这是辐射通量相对于方向的密度，更准确地说，是相对于立体角，它的单位是瓦特每立体弧度。

最后是辐射率（radiance）L（译注：radiance还有光辉、光芒的意思），它是一种测量单条射线的电磁辐射的方法，更准确地说，它被定义为辐射通量相对于面积和立体角的密度，这个面积是在垂直于射线的平面上测量的，如果辐射率被应用到一个带有某些方向的表面（译注：也就是不垂直），那么必须使用余弦校正因子，你可能会遇到使用术语“投影面积”定义的辐射率，这时参考这个校正因子。

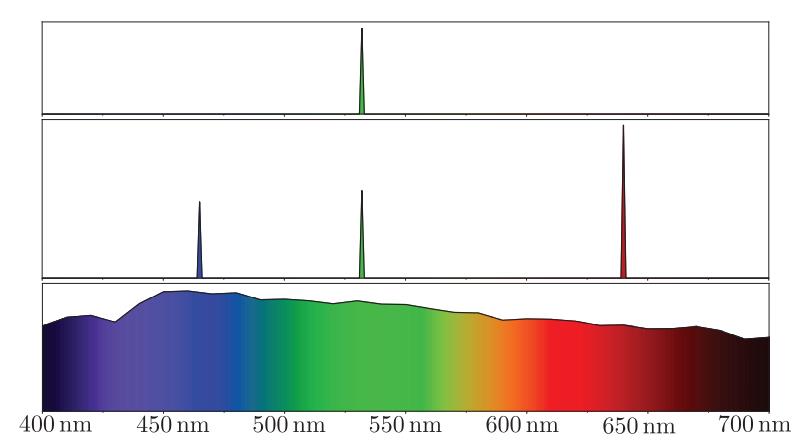
辐射率是传感器（如眼睛或相机）测量（更多细节参见9.2节）的量，所以对渲染来说它是最重要的量，计算着色方程的目的就是计算沿给定射线的辐射率，射线从着色表面点到相机，沿着这条射线的L值在物理上等同于第5章中的量，辐射率的公制单位是瓦特每平方米每立体弧度。

环境中的辐射率可以被认为是五个变量（或六个，包括波长）的函数，称为辐射分布（radiance distribution）[400]，其中三个变量指定位置，另外两个表示方向，这个函数描述了在空间中任意位置传播的所有光。思考渲染过程的一种方法是，眼睛和屏幕定义一个点和一组方向（例如，一条射线穿过每个像素），而这个函数在眼睛处对每个方向进行计算，在13.4节中讨论的基于图像的渲染使用了一个相关的概念，称为光场（light field）。

在着色方程中，辐射率通常以或的形式出现，分别表示从x点发出或进入x点的辐射率，方向向量**d**表示射线的方向，按照惯例，它总是指向远离x的方向，虽然这个惯例在的情况下可能有点令人困惑，由于**d**指向光传播的相反方向，可以很方便进行诸如点积之类的计算。

辐射率的一个重要特性是它不受距离的影响，忽略了大气（如雾）的影响，换句话说，无论表面与观察者的距离如何，它都具有相同的辐射率，当距离越远时，表面覆盖的像素就越少，但从表面发出的每个像素处的辐射率是恒定的。

大多数光波包含许多不同波长的混合物，这通常被可视化为光谱功率分布（spectral power distribution, SPD），这是一个显示光的能量如何在不同波长分布的图表，图8.3显示了三个示例，值得注意的是，尽管图8.3中的中间和底部的SPD间存在巨大差异，它们被视为相同的颜色，显然，人的眼睛不适合做光谱仪，我们将在8.1.3节详细讨论色彩视觉。



**图 8.3. 三种不同光波的光谱功率分布（SPD）。顶部的SPD针对绿色激光，它的光谱分布非常窄，它的波形类似于图9.1中的简单正弦波。中间的SPD是由相同的绿色激光加上两个额外的，一个红一个蓝的激光组成的光，这些激光的波长和相对强度对应于显示中性白色的RGB激光投影显示（译注：这句话的意思应该是，那些激光在光谱上的位置是以中性白色为基准）。底部SPD用于标准D65光源，这是一个典型的中性白色参考光，旨在代表室外照明，这种SPD的能量连续分布在可见光谱上，是典型的自然光照。**

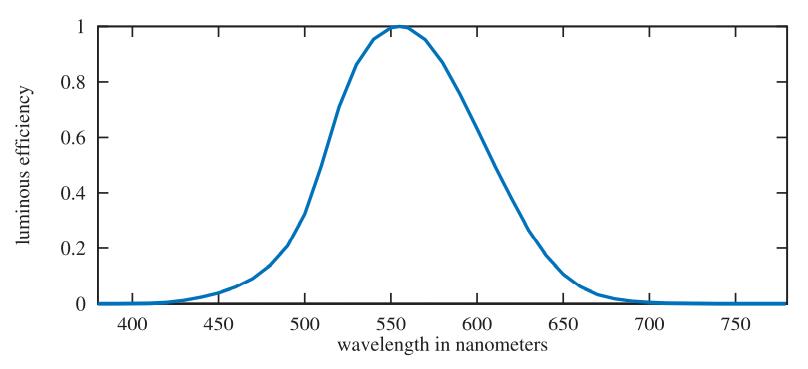
所有辐射量（radiometric quantities）都有光谱分布（译注：不是上述的光谱功率分布，这里的意思应是所有辐射量都有类似SPD的光谱分布图），因为这些分布是除以波长的密度，所以它们的单位是原始量除以纳米，例如，辐照度的光谱分布以瓦特每平方米每纳米为单位。

由于完整的SPD难以用于渲染，特别是在交互速率下，在实践中辐射量被表示为RGB三元组，在第8.1.3节中，我们将解释这些三元组与光谱分布的关系。

8.1.2 光度测定学 Photometry

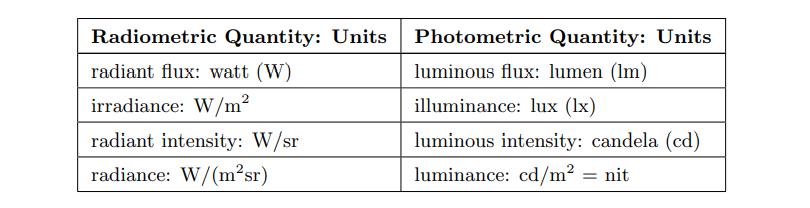
辐射度量学只处理物理量，不考虑人的感知，光度测定学是一个相关的领域，它与辐射度量学类似，不过是根据人眼的灵敏度来衡量一切。通过乘以CIE光度曲线（photometric curve）——一个以555nm为中心的钟形曲线——代表眼睛对不同波长的光的反应[76, 544]，辐射计算结果被转换为光度单位，参见图8.4。

（注：更准确的全名是“CIE适光光谱发光效率曲线”（CIE photopic spectral luminous efficiency curve）。“适光”一词指的是大于每平方米3.4坎德拉（candela，发光强度单位，也叫烛光）的光照条件——黄昏或更亮，在这种情况下，眼睛的视锥细胞是活跃的。有一个对应的“适暗”（scotopic）CIE曲线，以507nm为中心，用于当眼睛适应黑暗时，即低于每平方米0.034坎德拉——无月之夜或更暗的情况，这样的条件下眼睛的杆状细胞活跃。）



**图 8.4. 光度曲线。**

转换曲线和测量单位是光度测定理论和辐射度量学唯一的区别，每个辐射量都有一个等效的光度测定量，表8.2显示了它们的名称和单位，所有的单位都有如预期的关系（例如，勒克斯（lux）是每平方米流明（lumen））。虽然逻辑上流明应该是基本单位，但历史上坎德拉被定义为基本单位，其他单位都是从它派生出来的，在北美，照明设计师使用被弃用的帝国计量单位——英尺蜡烛（foot-candle, fc）来代替勒克斯测量照度。在任何情况下，大多数照度（illuminance）是测光计测量的，它在照明工程中是很重要的。



**表 8.2. 辐射度量学和光度测定学的量与单位。**

亮度（Luminance）常被用来描述平滑表面明亮的程度（brightness），例如，高动态范围（HDR）电视屏幕的峰值亮度通常在500到1000尼特（nit）之间，相比之下，晴空的亮度约为8000尼特，60瓦的灯泡亮度约为120000尼特，地平线上的太阳亮度为600000尼特[1413]。

8.1.3 比色法 Colorimetry

（译注：本节及之前章节涉及到了许多颜色相关的概念，它们的直译有混淆，在此总结如下：

1.色相-hue；

2.色调-tone；

3.色彩-tint；

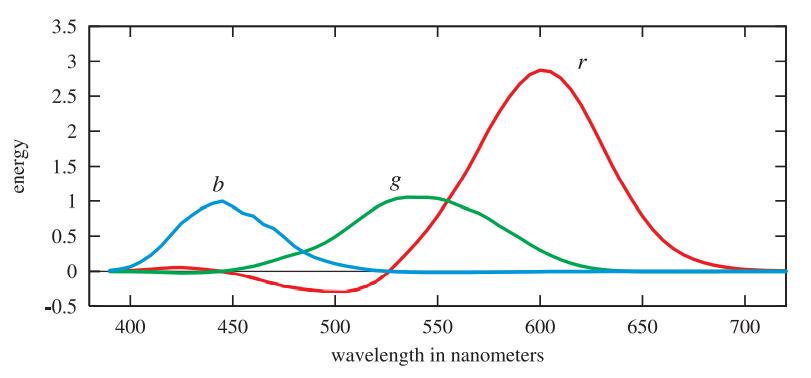
4.色度-chromaticity，任何颜色由色度与亮度表示；

5.颜色-color，颜色是一个通用概念，严格定义：人类视觉系统对可见光不同频率的辐射，所引发的感觉。但生活中使用时与上述概念并不做严格区分。）

在8.1.1节中，我们已经看到，我们对光的颜色的感知与光的SPD（光谱功率分布）密切相关，我们也看到这不是一个简单的一一对应关系，图8.3中底部和中间的SPD完全不同，但被视为完全相同的颜色，比色学研究的是光谱功率分布与颜色感知之间的关系。

人类可以分辨出大约1000万种不同的颜色，为了感知颜色，眼睛在视网膜上有三种不同类型的锥状受体，每种受体对不同波长的反应不同，其他动物有不同数量的颜色感受器，在某些情况下多达15个[260]。因此，对于一个给定的SPD，我们的大脑只从这些受体接收到三种不同的信号，这就是为什么只用三个数字就可以精确地表示任何颜色刺激[1707]。

哪三个数字呢？CIE（国际照明委员会）提出了一套测量颜色的标准，并利用它们进行了颜色匹配实验，在实验中，三种颜色的光投影到白色的屏幕上，使它们的颜色加在一起形成一块区域（译注：这个区域就有混合后的颜色了），要匹配的测试颜色被投影到这个区域旁边，测试颜色具有单一波长，然后观察者可以使用刻度到加权范围[−1, 1]的旋钮改变三种颜色的光（译注：的比例，也因此使最终结果变色），直到测试颜色匹配。一些测试颜色需要负权重来匹配，这样的权重意味着相应的光被添加到测试颜色区域的波长中，r，g和b三个光的一组测试结果如图8.5所示。光几乎是单色的，每个光的能量分布狭窄地聚集在以下波长之一：r为645nm，g为526nm，b为444nm，将每组匹配权值与测试区域波长联系起来的函数称为颜色匹配函数（color-matching function）。



**图 8.5. r，g和b的2维颜色匹配曲线，由Stiles和Burch提供[1703]，不要将这些颜色匹配曲线与配色实验中使用的光源的光谱分布混淆，虽然它们都是纯波长。**

这些函数提供了一种将光谱功率分布转换为三个值的方法，给定单一波长的光，创建的照明条件将从屏幕上的两个光照区域得到相同的感知，可以从图中读出三种颜色光的设置，即旋钮设置。对于任意的光谱分布，颜色匹配函数可以乘以每个结果曲线下的分布与面积（即积分），得出设置彩色光的相对数量，以匹配光谱产生的感知颜色。很多不同的光谱分布可以解析为相同的三个权重，也就是说，它们在观察者看来是相同的，权值匹配的光谱分布称为同色光（metamers）。

三个加权的r，g和b光不能直接代表所有可见的颜色，因为它们的配色函数对不同波长都有负的权重，CIE提出了三种不同的光源假设，它们的颜色匹配函数对所有可见波长都是正的，这些曲线是原始的r，g和b颜色匹配函数的线性组合，这要求光源的光谱功率分布在某些波长为负，因此这些光是不可实现的数学抽象，它们的配色函数分别表示为，和，如图8.6所示。颜色匹配函数与光度曲线（图8.4）相同，因为辐射率通过此曲线转换为亮度。

图表

描述已自动生成

**图 8.6. Judd-Vos改进的CIE(1978) 2维颜色匹配函数，注意两个是同一条曲线的一部分。**

与前面的一组颜色匹配函数一样，，和都可以通过乘法和积分将任何SPD简化为3个数字：

|  |  |
| --- | --- |
|  | (8.1) |

这些X、Y和Z三刺激值（tristimulus values）是CIE XYZ空间中定义颜色的权重，把颜色分为亮度（明亮程度）和色度（chromaticity）通常很方便，色度是一种颜色的特性，与它的亮度无关，例如，两种蓝色，一种深一种浅，尽管亮度不同，但可以具有相同的色度。

图表, 雷达图

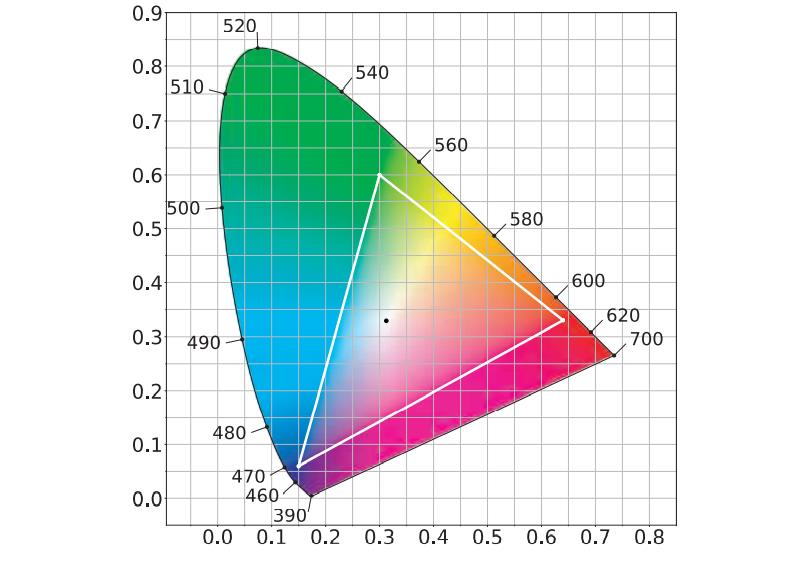
描述已自动生成

**图 8.7. 在XYZ空间中显示的CIE RGB基色的RGB颜色立方体，以及它在平面上的投影（紫色）。蓝色的轮廓包含可能的色度值的空间，从原点辐射出来的每条线都有一个恒定的色度值，只在亮度上有变化。**

为此，CIE通过将颜色投影到平面上定义了二维色度空间，参见图8.7，这个空间中的坐标为x和y，其计算方法如下：

|  |  |
| --- | --- |
|  | (8.2) |

z值没有提供额外的信息，所以通常省略它。色度坐标x和y值的图表（plot）称为CIE 1931色度图（chromaticity diagram），参见图8.8，图中的曲线轮廓表示可见光谱的颜色所在，连接光谱两端的直线称为紫线（purple line），黑点表示光源D65的色度，这是一个常用的白点——用来定义白色或消色差的（achromatic，无色）色度的刺激。



**图 8.8. CIE 1931色度图。曲线用对应的纯色波长标记，白色三角形和黑点分别表示色域（gamut）和白点，用于sRGB和Rec. 709颜色空间。**

总而言之，我们从一个实验开始，使用三种单波长的光（混合），并测量每种光需要多少以匹配其他波长的光的颜色（译注：原文此处是外观，两种光看起来一样，但并不相同，其中一个由三种波长的光混合而得），有时，这些纯光必须添加到被观察的样品中，以便与之匹配，这得出了一组颜色匹配函数，它们被组合起来创建一个没有负值的新集合，有了这个非负的颜色匹配函数集，我们可以将任意光谱分布转换为XYZ坐标，该坐标定义了颜色的色度和亮度，可以简化为xy来仅描述色度，而保持亮度不变。

给定一个颜色点，从白点通过该点画一条线到边界（光谱线或紫色线），（白点）颜色点的相对距离与（白点）到区域边缘的距离的比值就是颜色的激发纯度（excitation purity），区域边缘上的点决定了主波长（dominant wavelength）。这些比色术语在图形学中很少遇到，相反，我们使用饱和度（saturation）和色相（hue），它们分别与激发纯度和主波长松散相关（correlate loosely），更精确的饱和度和色相的定义可以在Stone[1706]和其他人[456, 789, 1934]的书中找到。

色度图描述了一个（二维）平面，完整描述一种颜色所需的第三个维度是Y值——亮度，它们定义了所谓的xyY坐标系，色度图对于理解渲染中如何使用颜色以及渲染系统的限制非常重要。电视或电脑显示器通过使用R、G和B颜色值的某些设置来显示颜色，每个彩色通道控制一个显像基色（display primary），其发出具有特定光谱功率分布的光，三种基色的每一种都按其各自的颜色值进行缩放，并将这些颜色值加在一起，形成观察者所能感知的单一的光谱功率分布。

色度图中的三角形表示典型电视或电脑显示器的色域（gamut），三角形的三个角是基色，它们是屏幕所能显示的饱和度最高的红、绿、蓝色，色度图的一个重要特性是，这些极限颜色可以用直线连接起来，以展示显示系统作为一个整体的极限，直线代表混合这三种基色所能显示的颜色的极限。白点表示当R、G和B颜色值相等时显示系统产生的色度，重要的是要注意，显示系统的全色域是一个三维体，色度图只显示了这个体在二维平面上的投影，更多信息请参见Stone的书[1706]。

在渲染中有几个RGB空间，每个空间都由R、G和B基色和一个白点定义，为了比较它们，我们将使用一种不同类型的色度图，称为CIE 1976 UCS（uniform chromaticity scale，均匀色度标尺）图，该图表是CIELUV颜色空间的一部分，被CIE（以及另一个颜色空间CIELAB）采用，目的是为XYZ空间提供感知上更加统一的替代方案[1707]。在CIE XYZ空间中，感知正好相反（互拆）的颜色对在CIE XYZ空间的距离可以相差20倍，CIELUV在此基础上进行了改进，将这一比率降低到最多4倍，这种增加的感知统一性，使得1976年的图表在比较RGB空间的色域时比1931年的图表更好。对感知统一颜色空间的研究持续进行，最近产生了[364]和[1527]空间，这些颜色空间在感知上比CIELUV更统一，特别是对于现代显示器的标志性高亮度和饱和颜色，然而，基于这些颜色空间的色度图还没有被广泛采用，因此我们在本章中使用CIE 1976 UCS图，例如图8.9。

（译注：CIE提供的这些颜色空间基于人为定义，用于更好的描述颜色，比如将互斥颜色对定义到两端，具体可参考GAMES101-第20集中关于颜色空间的描述。）

图表, 表面图

描述已自动生成

**图 8.9. CIE 1976 UCS图，显示三个RGB颜色空间：sRGB，DCI-P3和ACEScg的基色和白点。sRGB图也可以用于Rec. 709（空间），因为两种颜色空间有着相同的基色和白点。**

在图8.9所示的三个RGB空间中，sRGB是目前在实时渲染中最常用的，重要的是注意在本节中我们使用“sRGB颜色空间”指的是具有sRGB基色和白点的线性颜色空间，而不是在第5.6节中讨论的非线性sRGB颜色编码。大多数计算机显示器是为sRGB颜色空间设计的，同样的基色和白点也适用于Rec. 709颜色空间，这是用于HDTV显示的，因此对于游戏主机来说非常重要。然而，越来越多具有更宽色域的显示器正被生产出来，一些用于照片编辑的计算机显示器使用Adobe1998颜色空间（未显示）。DCI-P3色彩空间最初是为剧情电影制作而开发，现在正得到更广泛的应用，从iPhones到Macs，苹果的整个产品线都采用了这种颜色空间，其他制造商也纷纷效仿，尽管超高清（ultra-high definition, UHD）内容和显示器被指定使用极宽色域Rec. 2020颜色空间，但在许多情况下DCI-P3也被用作UHD的实际颜色空间。Rec. 2020没有显示在图8.9中，但它的色域与图中第三个颜色空间ACEScg非常接近，ACEScg颜色空间是由美国电影艺术与科学学院（AMPAS）为剧情电影的计算机图形渲染开发的，它不打算用作显示器颜色空间，而是用作渲染的工作颜色空间（working color space），在渲染后将颜色转换为适当的显示器颜色空间。

虽然目前sRGB颜色空间在实时渲染中无所不在，但更宽的颜色空间的使用也呈增加趋势，最直接的好处是针对宽色域显示器的应用程序[672]，但即使是针对sRGB或Rec. 709显示器的应用程序也有优势。当在不同的颜色空间中执行乘法等常规渲染操作时，会得到不同的结果[672, 1117]，有证据表明，在DCI-P3或ACEScg空间执行这些操作比在线性sRGB空间执行这些操作产生更准确的结果[660, 975, 1118]。

从RGB空间到XYZ空间的转换是线性的，可以使用由RGB空间的基色和白点派生的矩阵来完成[1048]，通过矩阵的反转和连接，可以得到从XYZ转换到任意RGB空间，或两个不同RGB空间之间转换的矩阵。注意在这样的转换后，RGB值可能是负数或大于1，这些都是色域外的颜色，即在目标RGB空间中不可复制的颜色，可以使用各种方法将这些颜色映射到目标RGB色域[785, 1241]。

一种常用的转换是将RGB颜色转换为灰度亮度值（grayscale luminance），因为亮度和Y系数相同，这个操作只是RGB到XYZ转换的“Y部分”，换句话说，它是RGB系数和RGB-XYZ转换矩阵的中间行的点积，在sRGB和Rec. 709空间下，这个方程为[1704]

|  |  |
| --- | --- |
|  | (8.3) |

这又把我们带回到光度曲线上，见图8.4，这条曲线表示标准观察者的眼睛对不同波长的光的反应，乘以三种基色光的光谱功率分布，对每条曲线结果积分，得到的三个结果权重就是上面亮度方程的系数，红、绿、蓝各部分的灰度强度值不相同的原因是眼睛对不同波长的光有不同的敏感度。

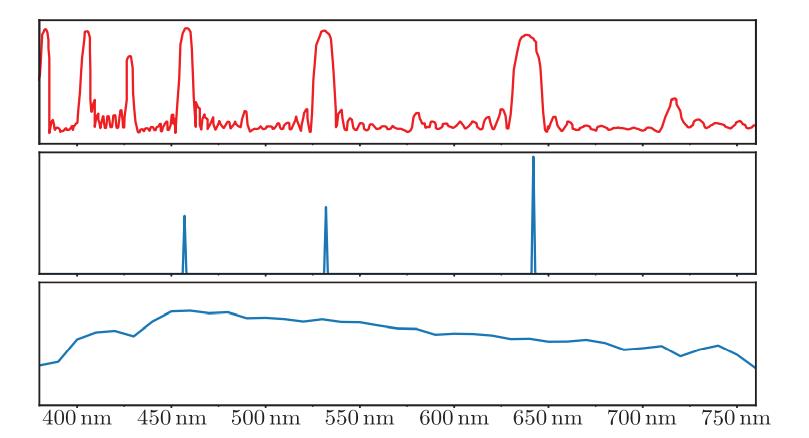
比色法可以告诉我们两种颜色刺激是否匹配，但不能预测它们的外观，给定XYZ颜色刺激的外观在很大程度上取决于诸如照明、周围颜色和先前的条件等因素，颜色外观模型（Color appearance models, CAM）如CIECAM02，试图处理这些问题，并预测最终的颜色外观[456]。

颜色外观建模是视觉感知（visual perception）的更广泛领域的一部分，包括遮蔽（masking）等效果[468]，它是在物体上使用高频、高对比度的图案试图隐藏缺陷，换句话说，像波斯地毯这样的纹理将有助于掩盖色带（color banding）和其他着色伪像，这意味着渲染这类表面需要的花费更少。

8.1.4 使用RGB颜色渲染 Rendering with RGB Colors

严格来说，RGB值代表的是感知量而不是物理量，使用它们进行基于物理的渲染在技术上是一种分类错误，正确的方法是使用光谱量执行所有渲染计算，然后通过密集采样或在适当的基础上投影，并在最后转换为RGB颜色显示。

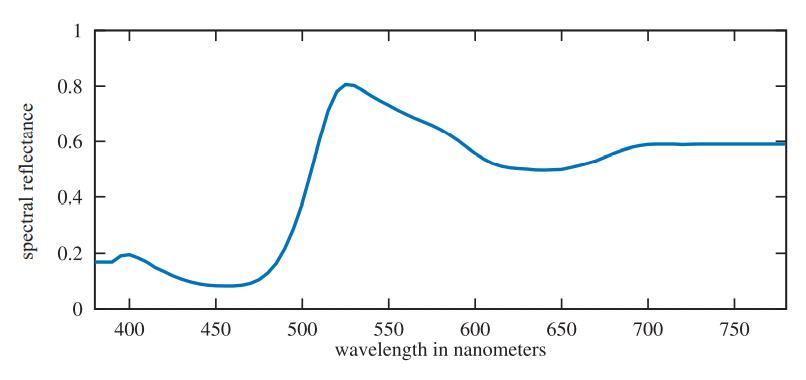
例如，最常见的渲染操作之一是计算从物体反射的光，物体的表面通常会更多的反射某些波长的光（比起其他波长），这可以用它的光谱反射率（spectral reflectance）曲线来描述，严格正确的计算反射光颜色的方法是将每个波长的光谱反射率乘以入射光的SPD，得到反射光的SPD，然后将其转换为RGB颜色，而不是在RGB渲染器中将光线和表面的RGB颜色相乘以得到反射光的RGB颜色，在一般情况下，这不会得到正确的结果，为了说明这一点，我们将看一个有些极端的例子，参见图8.10。



**图 8.10. 上面的图显示了一种用于屏幕投影的材料的光谱反射率。下面两图显示了两种具有相同RGB颜色的光源（illuminants）的光谱功率分布：中间图是一个RGB激光投影仪，底部图是D65标准光源。这种屏幕材质将反射大约80%的来自激光投影仪的光，因为它的反射率峰值与投影仪的基色（基本）对齐，但是，它将反射来自D65光源的不到20%的光，因为大多数光源的能量在屏幕的反射率峰值之外。如果该场景使用的是RGB渲染，可以预测屏幕对两种光的反射强度相同（但实际根本不同）。**

我们的例子展示了一种用于激光投影仪的屏幕材料，它在窄波段具有高反射率，与激光投影仪的波长相匹配，同时在大多数其他波长具有低反射率，这导致它反射来自投影仪的大部分光，但吸收来自其他光源的大部分光，在这种情况下，RGB渲染器会产生严重错误。

然而，图8.10所示的情况远非典型，实际中遇到的表面的光谱反射率曲线要平滑得多，例如图8.11所示，典型光源的SPD类似于D65光源，而不是例子中的激光投影仪，当光源SPD和表面光谱反射率都很平滑时，RGB渲染引入的误差相对较小。



**图 8.11. 黄色香蕉的光谱反射率[544]。**

在预渲染（predictive rendering）应用程序中，这些细微的错误可能很重要，例如，两条光谱反射率曲线可能在一个光源下具有相同的颜色外观，而在另一个光源下不同，这个问题称为同色光失效（metameric failure）或光源条件等色（illuminant metamerism），在例如涂装修理过的汽车车身部分时，这是一个严重的问题，RGB渲染并不适用于尝试预测这种类型效果的应用程序中。

但是，对于大多数渲染系统，特别是那些交互式应用程序，它们的目标并不是产生预测模拟，RGB渲染工作得非常好[169]，即便剧情电影的离线渲染也只是最近才开始使用光谱渲染，而且它还远未普及[660, 1610]。

本节只涉及了颜色科学的基础知识，主要是让人们认识到光谱与颜色三元组的关系，并讨论设备的局限性，下一节将讨论一个相关的主题，即渲染场景的颜色到显示值的转换。

8.2 场景到屏幕 Scene to Screen

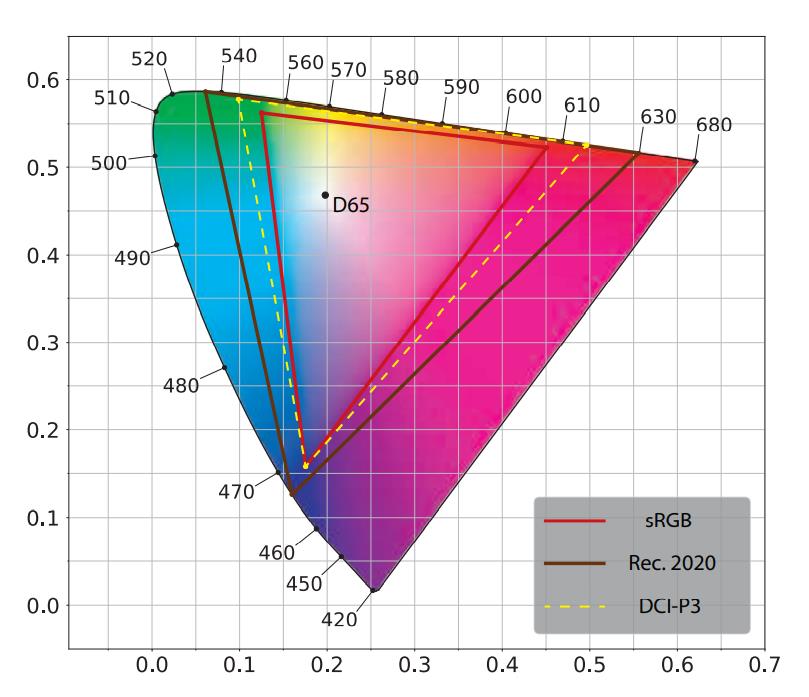
本书接下来的几章将重点讨论基于物理渲染的问题，给定一个虚拟场景，基于物理渲染的目标是假设这个场景是真实的情况下，计算其中显示的辐射率，然而，在这一点上要做的工作还有很多，最终的结果——显示的framebuffer中的像素值—仍然需要确定，在本节中，我们将讨论这一结果涉及的一些注意事项。

8.2.1 高动态范围显示编码 High Dynamic Range Display Encoding

本节的内容建立在第5.6节的基础上，内容涵盖显示编码，我们决定将高动态范围（HDR）显示的讨论推迟到本节，因为它需要一些有关的背景知识，比如色域，这些内容在本书的那一部分中还没有被讨论。

第5.6节讨论了标准动态范围（SDR）显示器的显示编码，SDR显示器通常使用sRGB显示标准，而SDR电视机则使用Rec. 709和Rec. 1886标准，这两套标准都有相同的RGB色域和白点（D65），以及有点类似（但不完全相同）的非线性显示编码曲线，它们也有大致相似的参考白色亮度等级（sRGB为80 cd/，Rec. 709/1886为100 cd/），不过显示器和电视机的制造商并没有严格遵守这些亮度规范，他们在实践中倾向于制造白色亮度等级更高的显示器[1081]。

HDR显示器使用Rec. 2020和Rec. 2100标准，Rec. 2020定义了一个色域明显更宽的颜色空间，如图8.12所示，以及与Rec. 709和sRGB颜色空间相同的白点（D65）。Rec. 2100定义了两种非线性显示编码：感知量化器（perceptual quantizer, PQ）[1213]和混合对数-伽马（hybrid log-gamma, HLG），HLG编码在渲染环境中使用不多，因此在这里我们将重点放在PQ上，它定义了10,000 cd/的峰值亮度值。



**图 8.12. CIE 1976 UCS图显示了Rec. 2020和sRGB/Rec. 2020的色域和白点（D65）。还显示了DCI-P3颜色空间的色域以进行比较。**

虽然峰值亮度和色域规格对编码目的来说很重要，但就实际显示而言，它们有些遥远，在撰写本文时，甚至很少有消费级的HDR显示器的峰值亮度超过1500 cd/，在实际中，显示色域更接近于DCI-P3（也在图8.12中显示）而不是Rec. 2020，基于这个原因，HDR显示器通过内部色调和色域将标准规格映射实际显示能力，这种映射可能会受到应用程序传递的元数据的影响，以指示内容的实际动态范围和色域[672, 1082]。

从应用程序端来看，有三种路径用于将图像传输到HDR显示器上，尽管根据显示器和操作系统的不同，不是所有路径都可用：

1. HDR10—广泛支持HDR显示器以及PC和控制台的操作系统，其帧缓冲格式是每像素32位，每个RGB通道有10个无符号整数位，alpha通道有2个无符号整数位，它采用PQ非线性编码和Rec. 2020颜色空间，每个HDR10显示模型都执行自己的色调映射，这种映射没有标准化或文档化。
2. scRGB（线性变体）—仅支持Windows操作系统，名义上它使用sRGB基色和白色级别，但两者都可以超出，因为标准支持RGB值小于0或大于1，帧缓冲格式是每个通道16位，并存储线性RGB值，它可以在任何HDR10显示器上工作，只需将驱动程序转换为HDR10，使用它主要为了方便以及向后兼容sRGB。
3. 杜比视觉（Dolby Vision）—专有格式，尚未在显示器或任何控制台上得到广泛支持（撰写本文时），它使用自定义的每通道12位的帧缓冲格式，并使用PQ非线性编码和Rec. 2020颜色空间，显示内部色调的映射是跨模型标准化的（但没有文档化）。

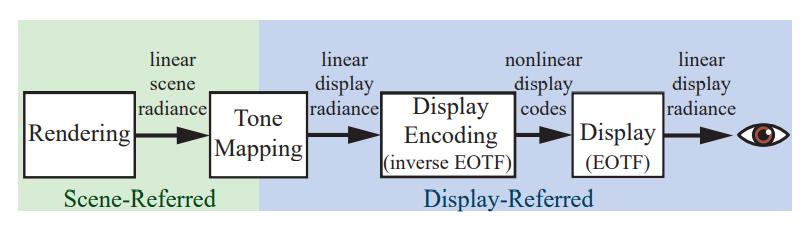
Lottes[1083]指出实际上还有第四种选择，如果仔细调整曝光和颜色，那么HDR显示器可以通过常规SDR信号路径来驱动，效果良好。

对于除了scRGB以外的任意选项，作为显示-编码步骤的一部分，应用程序需要将像素的RGB值从渲染工作空间转换到Rec. 2020，这需要一个3×3矩阵变换，然后应用PQ编码，这比Rec. 709或sRGB编码函数[497]稍昂贵一点，Patry[1360]给出了PQ曲线的一种廉价近似。在HDR显示器上合成用户界面（UI）元素时需要特别注意，确保用户界面易读且亮度舒适[672]。

8.2.2 色调映射 Tone Mapping

在第5.6节和8.2.1节中，我们讨论了显示编码，即将线性辐射率值转换为显示硬件的非线性编码值的过程，显示编码所应用的函数是显示器的光电传递函数（electrical optical transfer function, EOTF）的逆函数，它确保输入的线性值与显示器发出的线性辐射率匹配，我们前面的讨论忽略了发生在渲染和显示编码之间的一个重要步骤，现在我们准备探讨这个步骤。

色调映射或者说色调重现（tone reproduction），是将场景辐射率值转换为显示的辐射率值的过程，在此步骤中应用的转换称为端到端传递函数（end-to-end transfer function），或场景到屏幕的转换（scene-to-screen transform）。图像状态（image state）的概念是理解色调映射的关键[1602]，有两种基本的图像状态：场景相关（Scene-referred）图像是根据场景辐射率值定义的，而显示相关（display-referred）图像是根据显示辐射率值定义的，图像状态与编码无关，两种状态下的图像都可以被线性或非线性编码。图8.13显示了图像状态、色调映射和显示编码如何在成像管线（imaging pipeline）中组合在一起，处理从初始渲染到最终显示的颜色值。



**图 8.13. 合成（渲染）图像的成像管线。我们渲染（场景相关状态）的线性辐射率值，色调映射将其转换为（显示相关状态）的线性值，显示编码应用EOTF的逆将线性显示值转换为非线性编码值（代码），将其传递给显示器，最后，显示硬件应用EOTF将非线性显示值转换为从屏幕发射到眼睛的线性辐射。**

关于色调映射的目标有几个常见的误解，它并不是确保从场景到屏幕的转换是恒等变换，并在显示时完美重现场景辐射率值；也不是将每个比特的信息从场景的高动态范围“挤”到显示的低动态范围中，尽管场景和显示的动态范围之间的差异确实发挥了重要作用。

要理解色调映射的目标，最好将其视为图像重现（image reproduction）的一个实例[757]，图像重现的目标是创建一个显示相关状态的图像，在给定显示属性和观察状态下，尽可能接近地重现观察者在观察原始场景时会产生的感知印象，参见图8.14。



**图 8.14. 图像重现的目的是确保重现的图像所唤起的知觉印象（右）尽可能接近于原始场景（左）。**

有一种类型的图像重现目标略有不同，首选图像重现（Preferred image reproduction）旨在创建一个某种意义上，看起来比原始场景更好的显示相关状态的图像，首选图像重现将在后面的8.2.3节中讨论。

考虑到典型场景中的亮度范围超过显示能力几个数量级，重现与原始场景相似的感知印象是一个具有挑战性的目标，场景中也可能有些颜色的饱和度（纯度）远远超过显示能力，然而，正如文艺复兴时期的画家们所做的那样，摄影、电视和电影确实能够对原始场景产生令人信服的感知模拟，这一成就是利用人类视觉系统的某些特性实现的。

视觉系统会补偿绝对亮度的差异，这种能力被称为适应性（adaptation），由于这种能力，在昏暗房间里的屏幕上显示的室外场景的重现可以产生与原始场景相似的感觉，尽管重现的亮度不到原始场景的1%。然而，适应性所提供的补偿不完美，在较低的亮度水平下，感知到的对比度是降低的（Stevens效应），感知到的“色彩”也是如此（Hunt效应）。

其他因素也会影响重现的实际或感知对比度，显示器的环境（显示器矩形外的亮度水平，如房间照明的亮度）可能增加或减少感知对比度（Bartleson-Breneman效应）。显示耀斑（Display flare），是通过显示器缺陷或屏幕反射而添加到被显示图像上的多余光，通常会在相当大的程度上降低图像的实际对比度，这些效果意味着，如果我们想要保留与原始场景相似的感知效果，我们必须提高显示相关状态下图像的对比度和饱和度值[1418]。

然而，这种反差的增加加剧了一个现存的问题，由于场景的动态范围通常比显示的范围大得多，我们不得不选择一个较窄的亮度窗口来重现，并将该窗口上方和下方的值剪切为黑色或白色，提高对比度会进一步缩小窗口，为了部分地抵消暗和亮值的裁剪，可以用一个柔软的变化（soft roll-off）带回一些阴影和高光的细节。（译注：在形容颜色的变化时，此处和后文提到了roll off，个人认为这是一种贴切的形容，但没有很好的中文对应词，所以翻译为变化。）

所有这些导致了一个S形的（sigmoid）色调重现曲线，类似于光化学胶片（film）提供的曲线[1418]，这并非偶然，柯达和其他公司的研究人员仔细调整了光化学胶片乳液的特性，以产生有效和喜人的图像重现，由于这些原因，“电影般的（filmic）”这个形容词经常出现在色调映射的讨论中。

曝光（exposure）的概念是色调映射的关键，在摄影中，曝光指的是控制落在胶片或传感器上的光量，然而，在渲染中，曝光是在应用色调重现变换之前，对场景相关状态的图像进行的线性缩放操作。曝光的棘手之处在于确定应用的比例因子，色调重现变换和曝光是紧密联系在一起的，通常色调变换的设计会期望将它们应用于已用某种方法曝光的场景相关状态的图像。

通过曝光进行缩放，然后应用一个色调重现变换的过程，组成全局色调映射的一种形式（global tone mapping），在这种映射中，相同的映射应用于所有像素，相比之下，局部色调映射（local tone mapping）过程基于像素周围和其他因素，在像素间使用不同的映射，实时应用程序几乎只使用全局色调映射（有少数例外[1921]），因此我们将重点讨论这种类型，首先讨论色调重现变换，然后讨论曝光。

重要的是要记住，场景相关状态的图像和显示相关状态的图像是根本不同的，物理操作只有在对场景相关的数据执行时才有效，由于显示的限制和我们讨论过的各种感知的影响，两个图像状态之间总是需要一个非线性变换。

**色调重现变换 Tone Reproduction Transform**

色调重现变换通常表示为将场景相关的输入值映射为显示相关的输出值的一维曲线，这些曲线可以独立应用于R、G和B值，或亮度，在前一种情况下，结果自动位于显示色域中，因为每个显示相关的RGB通道值将在0到1之间。然而，在RGB通道上执行非线性操作（特别是剪切）除了导致亮度发生所需变化外，还可能导致饱和度和色相变化，Giorgianni和Madden[537]指出，饱和度的变化可能对感知有益，大多数重现变换使用对比度增强来抵消Stevens效应（以及环境与观察耀斑效应），这将导致相应的饱和度增强，也将抵消Hunt效应，然而，通常认为色相变化是不可取的，而现代色调变换试图通过在色调曲线后应用额外的RGB调整来减少它们（即饱和度）。

通过将色调曲线应用到亮度，可以避免色相和饱和度的变化（或者至少减少），然而，这可能导致显示相关状态的颜色超出显示器的RGB色域，在这种情况下它需要被映射回来。

色调映射的一个潜在问题是，将非线性函数应用到场景相关状态的像素颜色可能会导致一些反走样技术的问题，这个问题（以及解决它的方法）在第5.4.2节中讨论。

Reinhard色调重现操作[1478]是用于实时渲染的早期色调变换之一，它保持较暗的值基本不变，而较亮的值渐近地变成白色。Drago等人[375]提出了一种有点类似的色调映射操作，该操作能够根据输出显示亮度进行调整，这可能使其更适合HDR显示器。Duiker为电子游戏创造了一种近似柯达胶片的响应曲线的曲线[391, 392]，这条曲线后来被Hable[628]修改，添加了更多用户控制，并在游戏《神秘海域2》中使用，Hable对这条曲线的阐述很有影响力，导致“Hable胶片曲线”被运用在一些游戏中，Hable[634]后来提出了一种新的曲线，与他早期的工作相比，其具有许多优点。

Day[330]提出了一种S形的色调曲线，它被用于Insomniac Games工作室的片头，以及游戏《使命召唤：高级战争》中。Gotanda[571, 572]创建了模拟胶片的响应和数码相机传感器的色调转换，《星之海洋4》等游戏使用了这些方法。Lottes[1081]指出，显示器耀斑对显示器有效动态范围的影响是显著的，并且高度依赖于房间照明条件，因此，为用户提供对色调映射的调整项是很重要的，他提出了一种支持此类调整的色调重现转换，可用于SDR和HDR显示器。

学院颜色编码系统（Academy Color Encoding System, ACES）是由美国电影艺术与科学学院的科技委员会创建，作为管理电影和电视行业色彩的建议标准。ACES系统将场景到屏幕的转换分为两个部分，第一种是相关状态的渲染转换（reference rendering transform, RRT），它在一个被称为输出颜色编码规范（output color encoding specification, OCES）的标准的，与设备无关的输出空间内将场景相关的值转换为显示相关的值。第二部分是输出设备变换（output device transform, ODT），它将颜色值从OCES转换到最终显示编码，有许多不同的ODT，每一个都是针对特定的显示设备和观察状态设计的。RRT和适当ODT的连接创建了整体转换，这种模块化结构便于处理各种不同的显示类型和观察状态，Hart[672]建议在需要同时支持SDR和HDR显示器的应用程序中使用ACES色调映射转换。

尽管ACES是为电影和电视而设计的，但它的变换在实时应用程序中得到了越来越多的应用，ACES色调映射在虚幻引擎[1802]中是默认启用的，Unity引擎也支持它[1801]。Narkowicz给出了适用于具有SDR和HDR的ODT的ACES RRT的廉价曲线[1260, 1261]，以及Patry[1359]、Hart[672]提出了ACES ODT的参数化版本，以支持一系列设备。

用于HDR显示器的色调映射需要注意，因为显示器也会应用一些它们自己的色调映射。Fry[497]提出了一套用于寒霜游戏引擎的色调映射变换，它们为SDR显示器应用了相对积极的色调重现曲线，而为使用HDR10信号路径（根据显示器的峰值亮度有一些变体），以及使用杜比视觉路径且没有色调映射的显示器（换句话说，它们依赖于显示器内置的杜比视觉色调映射），应用不那么积极的曲线。寒霜引擎的色调重现变换被设计成中性的，没有显著的对比度或色相变化，目的是通过颜色分级应用任何想要的对比度或色相修改（第8.2.3节），为此，色调重现变换要在颜色空间中应用[364]，颜色空间是为了色度轴和亮度轴之间的感知一致性和正交性而设计的，寒霜会改变亮度的色调映射，并随着显示的亮度变白而逐渐降低色度的饱和度，这提供了一个不需要色相变化的干净转换。

讽刺的是，随着资源（assets）（如火焰效果）出现的问题会有权在它们之前的变换中影响色相的变化，所以最终寒霜团队修改了变换，允许用户重新引入某种程度的色相变化到显示相关的颜色，图8.15显示了寒霜的变换与本节提到的其他几个方法的对比。



**图 8.15. 应用了四种不同色调变换的场景，差异主要体现在圈出的区域，那里的场景像素值特别高。左上：剪切（加上sRGB OETF）；右上：Reinhard[1478]；左下：Duiker[392]；右下：寒霜（保留色相的版本）[497]。Reinhard、Duiker和寒霜的变换都保留了由剪切丢失的高光信息，然而，Reinhard曲线倾向于降低图像较暗部分的饱和度[628, 629]，而Duiker变换增加较暗区域的饱和度，这有时被认为是一个合适的特征[630]，通过设计，寒霜变换同时保留了饱和度和色相，避免了在其他三张图片的左下角圆圈中可以看到的强烈的色相变化。（图片由©2018 Electronic Arts公司提供。）**

**曝光 Exposure**

一组常用的计算曝光的技术依赖于分析场景相关的亮度值，为了避免引入停顿，该分析通常通过采样前一帧来完成。

根据Reinhard等人[1478]的建议，在早期的实现中使用的一个指标是对数平均场景亮度，通常，曝光是通过计算帧的对数平均值来确定的[224, 1674]，这个对数平均值是通过执行一系列的下采样后处理pass来计算的，直到计算出帧最终的单一值。

使用平均值往往对异常值过于敏感，例如，少量明亮的像素可能会影响整个帧的曝光，后来的实现通过使用亮度值的直方图来改善这个问题，直方图可以计算中位数，而不是平均值，这样更具健壮性，直方图中的其他数据点可以用于改进结果，例如，在Valve的《橙盒》中，基于第95百分位和中位数的启发式方法（heuristics）被用于确定曝光率[1821]。Mittring描述了使用计算着色器来生成亮度直方图的方法[1229]。

到目前为止讨论的技术问题是，像素亮度是驱动曝光的错误量，如果我们看看摄影实践，比如Ansel Adams的区域系统[10]以及如何使用入射光来设置曝光的话，很明显，最好是单独使用照明（不受表面反射率的影响）来确定曝光[757]，这样做是有效的，因为粗略地说，照相曝光是用来中和光线的，这导致打印（出来的图片）主要显示物体表面的颜色，这与人类视觉系统的颜色恒定（color constancy）特性相对应，以这种方式处理曝光还可以确保将正确的值传递给色调变换，例如，在电影或电视行业中使用的大多数色调变换都被设计成将曝光的场景相关值0.18映射成显示相关值0.1，并希望用0.18表示主场景照明中18%的灰色卡片[1418, 1602]（译注：灰卡是摄影中的名词）。

尽管这种方法在实时应用程序中还不常见，但它已经开始得到使用，例如，游戏《合金装备V：地面部队》便拥有基于光照强度的曝光系统[921]。在许多游戏中，静态曝光水平是根据已知的场景光照值为环境的不同部分手动设置的，这样做可以避免曝光发生意想不到的动态变化。

8.2.3 颜色分级 Color Grading

在第8.2.2节中，我们提到了首选图像重现的概念，即生成在某种意义上比原始场景更好的图像，这通常涉及图像颜色的创造性操作，这个过程称为颜色分级。

数字颜色分级已经在电影工业中使用了一段时间，早期的例子包括电影《哦，兄弟，你在哪里？》（2000）和《Amélie》（2001）。颜色分级通常是通过交互式的控制示例场景图像中的颜色来执行，直到获得所需的创造性“外观”，然后将相同顺序的操作重新应用于一个镜头或序列中的所有图像，颜色分级从电影传播到游戏，现在被广泛使用[392, 424, 756, 856, 1222]。

Selan[1601]展示了如何从颜色分级或图像编辑应用程序中“烘焙”任意颜色转换到三维颜色查找表（LUT）中，通过使用输入R、G和B值作为x、y和z坐标来应用这些表，用于在表中查找新颜色，因此可以用于从输入到输出颜色的任意映射，直到遇到LUT分辨率的限制，Selan的烘焙过程首先是获取一个标识LUT（将所有输入颜色映射为相同颜色的LUT），并将其“切片”以创建二维图像，然后将这个切片LUT图像加载到颜色分级应用程序中，并对它应用定义所需创造性外观的操作，需要注意的是只对LUT应用颜色操作，避免诸如模糊等空间操作，然后将编辑过的LUT保存出来，“打包”到三维GPU纹理中，并在渲染应用程序中使用，对渲染的像素动态应用相同的颜色转换。Iwanicki[806]提出了一种使用最小二乘法最小化的聪明方法，来减少存储颜色变换到LUT中时的采样误差。

在后来的出版物中，Selan[1602]区分了两种进行颜色分级的方法，一种方法是对显示相关状态的图像数据执行颜色分级，另一种是对场景相关状态的数据执行颜色分级操作，并通过显示变换预览，虽然显示相关状态的颜色分级方法更容易设置，但场景相关状态的数据的分级可以产生更高保真度的结果。

当实时应用程序首次采用颜色分级时，显示相关状态的方法占主导地位[756, 856]，但是，场景相关的方法由于其更高的视觉质量开始获得关注[198, 497, 672]，参见图8.16，将颜色分级应用于场景相关的数据还提供了将色调映射曲线烘焙到分级LUT[672]中以节省一些计算的机会，就像在游戏《神秘海域4》[198]中所做的那样。



**图 8.16. 游戏《神秘海域4》中的一个场景，上面的屏幕截图没有颜色分级，其它两个截图都应用了颜色分级操作，为了说明目的，选择了一种极端的颜色分级操作（乘上高度饱和的青色），在左下方截图中，颜色分级应用于显示相关状态的（色调映射后）的图像，在右下方截图中，它应用于场景相关状态的（色调映射前）的图像。（《神秘海域4：盗贼末路》©/TM 2016 SIE。由Naughty Dog 有限责任公司创造和开发。）**

在LUT查找之前，必须将场景相关的数据重新映射到范围[0, 1][1601]，在寒霜引擎[497]中，感知量化器OETF用于此目的，但也可以使用更简单的曲线，Duiker[392]使用对数曲线，而Hable[635]建议使用一次或两次平方根算子。

Hable[635]很好地概述了常见的颜色分级操作和实现的注意事项。

进一步阅读和资源 Further Reading and Resources

比色学和颜色科学的“圣经”是Wyszecki和Stiles所著的《颜色科学（Color Science）》[1934]，其他好的比色法参考文献包括Hunt所著的《测量颜色（Measuring Colour）》[789]和由Fairchild所著的《颜色外观模型（Color Appearance Models）》[456]。

Selan的白皮书[1602]很好地概述了图像重现和“场景到屏幕”的问题，想进一步了解这一主题的读者可以去看Hunt[788]的《颜色再现（The Reproduction of Colour）》和Giorgianni和Madden的《数字颜色管理（Digital Color Management）》[537]，它们都是很好的参考资料。Ansel Adams摄影系列里的三本书[9, 10, 11]，特别是《底片（The Negative）》，为人们提供了电影摄影艺术和科学如何影响今天的图像重现的理论和实践的理解。最后，Reinhard等人的《彩色成像：基础和应用（Color Imaging: Fundamentals and Applications）》一书[1480]对整个研究领域进行了全面的概述。

第9章 基于物理的着色 Physically Based Shading

“让一个物体的形式成为它可能的样子，光、影和透视总是会使它变得美丽。”

——John Constable

在本章中，我们将介绍基于物理的着色的各个方面，我们在第9.1节以描述光-物质相互作用的物理性质开始，在第9.2节到9.4节我们展示这些性质如何与着色处理过程产生联系，第9.5节到9.7节专门讨论用于构建基于物理着色的模型的构建块，而模型本身——涵盖了各种各样的材质类型——将在第9.8节到9.12节中讨论，最后，在9.13节中，我们描述了如何将材质们混合在一起，并介绍了避免走样和保持表面外观的滤波方法。

9.1 光的物理性质 Physics of Light

光和物质的相互作用构成了基于物理的着色的基础，要理解这些相互作用，就需要对光的本质有一个基本的了解。

在物理光学中，光被建模为一种电磁横波（transverse wave），一种振荡电场和磁场垂直于其传播方向的波，这两个场的振幅是耦合的，磁场和电场矢量相互垂直，且它们的长度之比固定，这个比值等于相速度，我们稍后会讨论它。

在图9.1中我们看到一个简单的光波，事实上，它可能是最简单的——一个完美的正弦函数，这种波只有一个波长（wavelength），用希腊字母λ（lambda）表示。正如我们在第8.1节中看到的，人感知到的光的颜色与它的波长密切相关，因此，单一波长的光被称为单色光（monochromatic light），意思是“单色的”，然而，在实践中遇到的大多数光波是多色的（polychromatic），包含许多不同的波长。

图示

描述已自动生成

**图 9.1. 光，一种电磁横波。电场和磁场矢量彼此之间形成90°并在传播方向上振荡，图中所示的波是最简单的可能光波，它既是单色的（具有单一波长λ），又是线性偏振的（电场和磁场都沿着一条直线振荡）。**

图9.1中的光波从另一个角度来看异常简单，它是线性偏振的（linearly polarized），这意味着对于空间中的一个固定点，电场和磁场都沿着一条线来回移动，相反，在这本书中，我们关注的是非偏振（unpolarized）光，这是更普遍的，在非偏振光中场的振幅均匀分布在垂直于传播轴的所有方向上。不过尽管它们很简单，理解单色线性偏振波的行为是有用的，因为任何光波都可以被分解成这种波的组合。