2.5 An Efficient and Physically Plausible Real-Time Shading Model

CHRISTIAN SCHÜLER

INTRODUCTION

This article is a contribution to the ongoing race for visual fidelity. We will learn a mathematical model for surface shading that captures a wide range of materials while still being computationally efficient and practical with regard to content creation. The model can reproduce both metals and dielectrics, such as paper, wood, leather, varnish, paint, chrome, gold, and more. It does so with only a few parameters: two colors and a scalar measure of surface roughness. An alpha channel as either surface mask or surface transparency completes the parameterization via two common RGBA textures.

The intended target audience of this article includes authors of shader programs and technical artists. Knowledge of bidirectional reflection distribution function (BRDF) theory is helpful but not required. We will be discussing a *local illumination model* that is about one point on a surface and the light received and emitted from this point in various *directions*. This article assumes that any global questions are already solved; these include questions of occlusion, visibility, and generally what amount of light can be seen from which direction.

The outline is as follows: First, we will review two influential shading models (Blinn-Phong and Cook-Torrance), as they serve as the basis for our development. We will also briefly review some physics of light-surface interaction. This will lead the way to the formulation of the new model. Finally, we will discuss the practical issues of content creation and display gamma.

REVIEW: BLINN-PHONG AND COOK-TORRANCE

One of the earliest and most ubiquitous shading models is due to Bui-Tuong Phong together with its modification by Jim Blinn [Phong75, Blinn77]. It is an empirical model (i.e., "made-up") for a point on a surface lit by a number of discrete point light sources. The spectral intensity of this point, as seen by a viewer, is decomposed into four components. These are, in order of increasing computational cost, the *ambient*, *diffuse*, and *specular* components (ignoring the emissive component; see also Figure 2.5.1). They are characterized as follows:

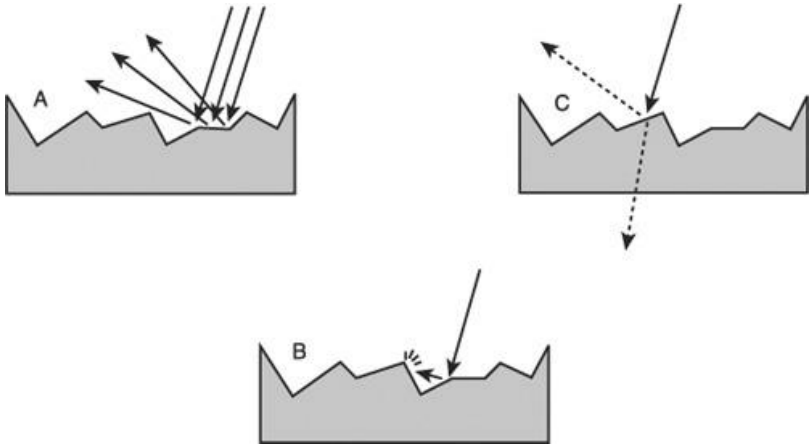
- The ambient component models a uniform field as a crude approximation for the combined effect of all indirect light. Ambient illumination is assumed to have equal intensity from all directions.
- The diffuse component assumes a Lambertian response to direct illumination from a discrete light source. This response is simply proportional to the dot product between surface normal and light direction.
- The specular component models the effect of mirror images (highlights) of discrete light sources via a simple formula. The fuzziness of these mirror images can be adjusted by an additional parameter that controls the appearance of surface roughness. At this point, the models differ in their approach: While the Phong model generates perfect reflections of fuzzed light sources, the Blinn model generates fuzzy reflections of perfect point lights. Both models are idealizations, but in reality, the Blinn model is closer to the observation more often.

Being designed for computer graphics 30 years ago, the advantage of the Blinn-Phong model is simply speed. The numerous disadvantages are both artistic (a trademark plastic look, limited expressiveness, and the uniform ambient term is lacking) and technical (it is not energy conserving and has somewhat arbitrary parameters).

On the other hand is the Cook-Torrance model, developed rigorously from the theory of micro-facets [Cook81]. This model puts the specular component on a physical basis, while the diffuse and ambient components are the same. The full Cook-Torrance model is expensive to compute, but it makes the important contribution of separating the specular component into three factors: a *distribution* factor accounting for surface roughness, a *geometry* factor accounting for self-shadowing and occlusion, and the *Fresnel* term (see Figure 2.5.2). This separation allows searching for approximations to each of these factors independently, which we will do later.

FIGURE 2.5.1 Components of the Blinn-Phong model for a single-point light. Top left: ambient component; top right: diffuse component; bottom left: specular component; bottom right: sum of all components.

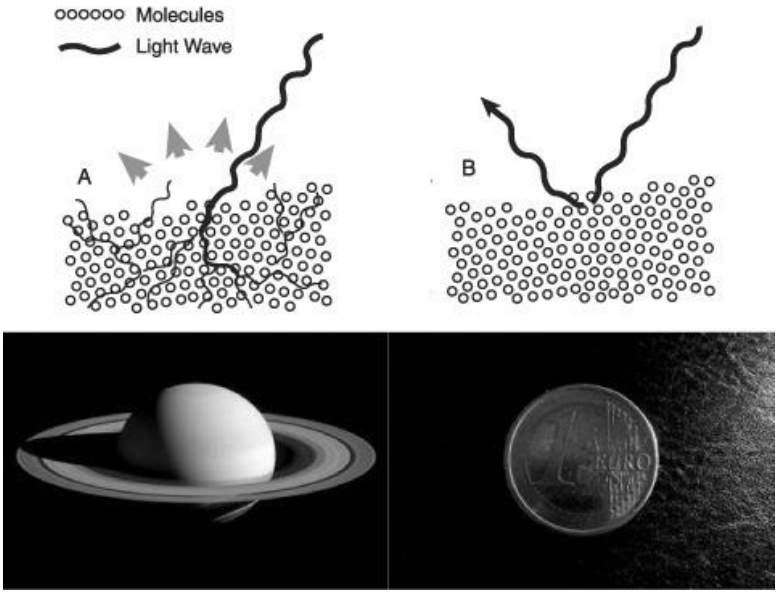
FIGURE 2.5.2 The three factors of the micro-facet model. A: distribution of micro-facets; B: geometric self-occlusion; C: Fresnel reflectance.



SOME PHYSICS OF LIGHT-SURFACE INTERACTION

The situation we are trying to model is that of light (an electromagnetic wave) traveling in a medium (air) and hitting another medium (the solid material). The light wave changes traveling speeds at the boundary, which causes part of it to be reflected back (the specular component according to our model). The remaining part passes the boundary into the second medium. However, the story does not end here because the light that passes can still return after some internal scattering (the diffuse component according to our model). Figure 2.5.3 illustrates these two principles of action.

FIGURE 2.5.3 Idealized physical reflection models. A: purely diffuse (subsurface) reflection; B: purely specular (Fresnel) reflection.



We can draw several conclusions from these observations.

- Specular reflection happens on the surface. This is why the specular component is colorless most of the time: The reflected light never has the opportunity to enter the substrate and take on its color.
- The amount of specular reflection on the surface is connected to the *refractive index* of the underlying material. This is why some metals, such as gold and copper, do have a colored specular reflection: Their refractive indices vary with wavelength.
- Diffuse reflection, in the sense of our model, is a subsurface effect. This is why metals have no diffuse component. An electromagnetic wave cannot penetrate an electrical conductor; it is short-circuited right at the surface. Therefore, the diffuse component of a metal is simply black (that is for atomic metal of course; if a metal is painted, enameled, or rusty, it is a not a metal in the sense of surface shading).

Toward an Improved Shading Model

While experimenting with shading models a few years back, it soon became apparent that a fundamental deviation from the Blinn-Phong model was needed.

As explained in the previous section, materials look most metallic if they have no diffuse component, but artists usually implemented a metallic look with a grayish diffuse texture and some specular added on top of it. When artists were told to leave the diffuse texture black, suddenly the ambient illumination was gone, since virtually all shading models link the ambient illumination to the diffuse texture. A new shading model would at least need to account for the specular component even under pure ambient illumination. In reality, we need individual terms for ambient diffuse and ambient specular reflection for anything more interesting than a uniform ambient field.

Another significant improvement comes from implementing the *Fresnel effect*. Surfaces become more reflective in the limit of a grazing angle. This is again a natural phenomenon connected to wave traveling speed. See Figure 2.5.4 for a real-world observation of the Fresnel effect.

FIGURE 2.5.4 Fresnel effect in the real world. A wooden desk lit by a halogen lamp from different view angles. The specular highlight becomes more intense from left to right.

The last important aspect is *energy conservation*. A glossy surface concentrates the energy of reflected light in a small range of directions. A rough surface spreads the energy of reflected light over a large range of directions. The glossiness of a surface therefore determines both

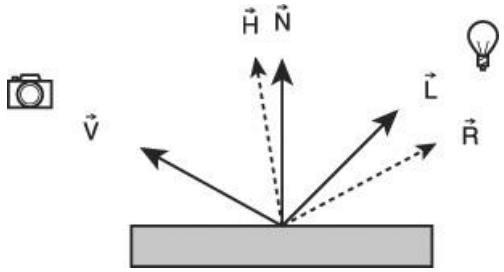
the width and scale factors for the specular reflection image. Failing to account for energy normalization results in the material parameters not being well separated.

Mathematical Formulation

Conventions

The following direction vectors are used in the lighting calculations: The surface normal $\vec{N}$, the direction pointing toward the viewer $\vec{V}$, the direction toward the light source $\vec{L}$, the reflection of the view direction $\vec{R}$, and finally the vector $\vec{H}$ as a bisector between view and light directions. All vectors are assumed to be of unit length. See also Figure 2.5.5.

FIGURE 2.5.5 Direction vectors used in lighting calculations.



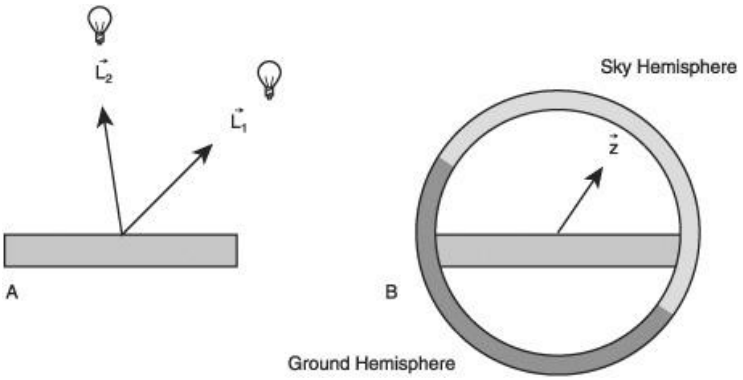
On a high level, the shading model calculates radiance from a point on a surface with normal $\vec{N}$, observed by a viewer into direction $\vec{V}$ and being lit by irradiance *E*. We choose *I* (for intensity) as the symbol for radiance, to prevent confusion with the light direction. Total radiance is the sum of the diffuse and specular parts.

$$\mathbf{I = I_{diffuse} + I_{specular}}$$

The ambient term no longer appears as a first-class component. Instead, we offer two general illumination models that can be mixed freely (see Figure 2.5.6).

- Illumination by a discrete set of light sources: This model has *n* lights, with the *i*th light into direction $\vec{L}^i$ and irradiance E^i_{light}.
- Illumination by a continuous two-hemisphere (sky and ground): This model has differently colored lower and upper hemispheres with irradiances $\mathbf{E_{lower}}$ and $\mathbf{E_{upper}}$ with the up direction assumed as $\vec{Z}$.

FIGURE 2.5.6 Illumination models. A: discrete light sources; B: continuous two-hemisphere model.



Diffuse Component

The diffuse component is straightforward, assuming a Lambertian reflectance. The material constant $\mathbf{K_{diffuse}}$ accounts for the average subsurface absorption (diffuse color). The formulae for different illumination models are then given as

$$I_{diffusediscrete} = k_{diffuse} \sum_{lights} \vec{N} \cdot \vec{L}^i \, E^i_{light},$$

$$I_{diffusehemisphere} = k_{diffuse} \mathrm{lerp}\left\{E_{lower}, E_{upper}, \frac{1 + \vec{N} \cdot \vec{z}}{2}\right\}.$$

Here, the function lerp(*a*,*b*,*t*) = *a* + *t*(*b*−*a*) is the linear interpolation between *a* and *b* for $\mathbf{0 \le t \le 1}$, an intrinsic function of the HLSL language. The shading model assumes that any global effects, for example, visibility and distance attenuation, are already factored into the available *E*. Visibility for the hemispheres can be precomputed as ambient occlusion. The two-hemisphere model is the solution of a simple integral (see the appendix at the end of this article).

Specular Component

The specular component is more involved; the results are displayed first:

$$I_{speculardiscrete} = k_{specular} \sum_{lights} \frac{1 + e}{8(\vec{L}^i \cdot \vec{H}^i)^3} (\vec{N} \cdot \vec{H}^i)^e \, \vec{N} \cdot \vec{L}^i \, E^i_{light}$$

$$I_{specularhemisphere} = F(g\vec{N} \cdot \vec{V}) \mathrm{lerp}\left\{E_{lower}, E_{upper}, \mathrm{clamp}\left[\frac{\vec{R} \cdot \vec{z}}{1 - g}, -1, 1\right]\right\}$$

$$\mathbf{e = 2^{12g},}$$

$$\mathbf{F(t) = lerp\{K_{specular}, min(60K_{specular}, 1), (1-t)^4\}.}$$

Here, the function clamp(*x*,*a*,*b*) = min(max(*x*,*a*),*b*) also denotes an intrinsic HLSL function, which limits a value *x* between lower and upper bounds. The discrete model is motivated from Cook-Torrance, but is dramatically simplified with many factors lumped together (see the appendix at the end of this article for a complete derivation).

The result only differs by an additional factor $\frac{(1 + e)\vec{N} \cdot \vec{L}}{8(\vec{L} \cdot \vec{H})^3}$ from the original Blinn-Phong model. It includes the specular exponent *e*, which is derived from a surface glossiness parameter *g* in the range 0–1 (rough to smooth).

The two-hemisphere model was developed empirically. The idea was to "look up" the hemisphere via the reflection vector $\vec{R}$ instead of the surface normal $\vec{N}$, and to scale the transition zone between the two hemispheres with the surface roughness. It includes *F*(…), which is a simplified Fresnel term, scaled such that the highest values are only reached for smooth surfaces.

Display Gamma

The response of a common display device is not linear with respect to the contents of the framebuffer, but rather follows a power law. The inverse exponent of this function is called *gamma*. This non-linearity will affect our shading calculations in two major ways, namely:

- Unfaithful representation of intensity (premature darkening)
- Unfaithful representation of color sums

Premature darkening is observed for intensities that are already low. For instance, a 2% reflection of the sky on a dark water surface would be perfectly visible, but after the display has raised this to some power, it is gone. Distance attenuation is also affected, since the attenuation law is raised to the power of the gamma value, making lights fade unnaturally fast if left uncorrected.

The second effect is observed, because the display of a sum (e.g., a gray pixel) is no longer

proportional to the sum of displays (e.g., a checkered pattern of black and white pixels). This affects all color mixtures, from anti-aliasing, texture filtering, and light sums to accumulation of shading terms. Adding the specular contribution on top of a diffuse one easily results in clipping and washout if left uncorrected.

The benefit in having a power law is the numerical compression of dynamic range. The sRGB standard has settled on a gamma value of 2.2, which is followed by virtually any current display technology. The 2.2 display gamma expands the dynamic range of an image with 8-bit precision to the equivalent of 17.6 bits (or a contrast of about 1:200k).

A practical shading model must account for display gamma on two fronts: texture input and shading output. Textures typically have at most 8 bits of precision and are authored via their display on a screen, so they have display gamma inherent. Texture colors must be raised to the power of 2.2 before they enter the shading calculation as intensities. There exists a feature on most graphics hardware to perform this as part of the texture sampler with approximate precision; it is activated either as a sampler state (DirectX 9) or as a surface format (DirectX 10 and OpenGL). The conversion can also be performed explicitly in shader code, which results in higher fidelity, especially for lower values, but makes the shader code make assumptions about the texture format.

To account for display gamma on the output side, there are multiple scenarios.

- **Low dynamic range framebuffer with display gamma (traditional case).** In this case the shader must raise the output color to a power of 0.45 (inverse display gamma) before it is written into the framebuffer. This makes sure the intensity that was calculated is actually displayed. There also exists a hardware feature for this; it is activated as a render state (DirectX 9 and OpenGL) or via the surface format of the render target (DirectX 10). In DirectX 10 and OpenGL, this state also affects framebuffer blending, which means a power of 2.2 is applied before the blending operation and a power of 0.45 afterward. If the hardware features are not used, it may be practical to assume a gamma of 2.0, so a simple squaring and square root can do the conversions.

- **Low dynamic range framebuffer, no display gamma.** The display gamma is already accounted for by a lookup table (“gamma ramp”), so the shader can output calculation results verbatim. This is not recommended since an LDR framebuffer does not have enough precision.

- **High dynamic range framebuffer.** An HDR framebuffer is usually not displayed directly but converted into an LDR framebuffer for display via tone mapping. The tone map operator needs to account for the display gamma. The shader can output calculation results verbatim into the HDR framebuffer.

Authoring Model Parameters

The shading model discussed so far has three parameters that determine material appearance. These are $\mathbf{K}_{\mathbf{diffuse}}$, the diffuse reflectance; $\mathbf{K}_{\mathbf{specular}}$, the specular reflectance for normal incidence (face-on); and *g*, a measure of surface roughness. Together with a surface mask (the [insert **α**-channel) these are four parameters, two colors and two scalars, which fit nicely in two RGBA textures. In this way, all parameters of the shading model are spatially varying, which allows us to batch a large collection of different appearances onto a single texture atlas.

The “specular color” on a texture, after gamma conversion, becomes $\mathbf{K}_{\mathbf{specular}}$, which is connected to the refractive index and as such describes a physical property. The shading model then varies this reflectance based on view angle and surface roughness. An artist should therefore use the *g* parameter to create variety, but not vary the $\mathbf{K}_{\mathbf{specular}}$ for a given material. For instance, the varying specular effect on the parts of a concrete floor is due to variance in glossiness; the $\mathbf{K}_{\mathbf{specular}}$ of the rock minerals is the same for the entire surface.

Table 2.5.1 lists as examples the theoretical values for some materials, derived numerically from spectral data [LuxPop]. As a rule of thumb, all metals have near unit reflectance ($\mathbf{K}_{\mathbf{specular}}$ is close to white), and all dielectrics have reflectances in the single-digit percent range.

Applications and Extensions

Applications of the discussed shading model are shown in scenes from the game *Velvet Assassin*, courtesy of Replay Studios GmbH. The main contribution is the unification of different materials into the same parameter set, allowing for large texture atlases. Color plate number 1 shows an interior scene with many materials (including cloth, wood, glass, and brass), with a magnification of a syringe on a plate in the inlet. Color plate number 2 shows how glass vs. stone can be achieved on the same texture, with glass shards in the inlet. Color plate

number 3 shows how wet vs. dry can be achieved on the same texture, with an opposite view in the inlet.

Possible extensions to the shading model are (a) the inclusion of an environment map for both diffuse and specular components with varying glossiness and (b) the correct handling of transparency with respect to Fresnel law. These topics weren't discussed in this article for reasons of scope, but the scenes shown in the color plates have both of these effects.

TABLE 2.5.1 Reflectivity and corresponding specular color by gamma conversion of selected materials

	Reflectance normal	Specular gray level or color
Water	0.02	44
Glass	0.03	56
Polystyrene	0.05	66
Calcite	0.06	72
Alumina	0.08	80
Diamond	0.17–0.18	114;115;117
Silicon	0.34–0.49	157;163;187
Copper	0.49–0.92	245;214;184
Gold	0.39–0.96	251;233;166
Silver	0.94–0.99	252;250;239

Note: The value for polystyrene can be used as typical for plastics and organic materials. The values for calcite and alumina can be used as typical for rock minerals (sapphires and rubies are also made from alumina). The metal examples are for pure and clean metals.

APPENDIX

DIFFUSE REFLECTION OF THE TWO-HEMISPHERE

A point on the surface is illuminated by complementary parts of a lower and an upper hemisphere. The result is a linear interpolation between irradiances $\mathbf{E_{lower}}$ and $\mathbf{E_{upper}}$ by some interpolation parameter *t*. With an angle $\boldsymbol{\alpha}$ defined as the angle between surface normal $\vec{N}$ and up-direction $\vec{Z}$, we can formulate *t* as the fraction of how much the upper hemisphere contributes to total irradiance, weighted by a Lambertian factor:

$$\mathbf{I_{diffusehemishere} = K_{diffuse} lerp(E_{lower}, E_{upper}, t),}$$

$$E_{upper}(\alpha) = \int_{\varphi=0}^{\pi}\int_{\theta=\alpha}^{\pi} \sin^2\varphi\sin\theta d\varphi d\theta = \frac{\pi + \pi\cos\alpha}{2}$$

$$t = \frac{E_{upper}(\alpha)}{E_{upper}(0)} = \frac{1 + \cos\alpha}{2}$$

NORMALIZING THE BLINN-PHONG SPECULAR HIGHLIGHT

The specular highlight of the Blinn-Phong model reflects less energy with increasing specular exponent *e*, because the shape of the highlight gets smaller without it getting brighter. A normalization factor must scale the term $(\vec{N}.\vec{H})^e$ such that it integrates to one over all directions. An integral over all possible $\vec{H}$ evaluates to

$$I(e) = \int_{\vec{H}} (\vec{N} \cdot \vec{H})^e d\omega^* = \int_{\varphi=0}^{2\pi} \int_{\theta=0}^{\pi/2} \cos\theta \sin^e\theta d\varphi d\theta = \frac{2\pi}{1+e}.$$

(*the differential solid angle)

An additional factor of $^1/2$ is introduced when changing variables from $\vec{H}$ to $\vec{V}$ because of the different angular distance covered. The complete normalization factor is therefore

$$\frac{1}{2}\frac{I(0)}{I(e)} = \frac{1+e}{2}$$

Simplifying the Cook-Torrance Model

The micro-facet model has a distribution factor *D*, a Fresnel factor *F*, and the geometric self-occlusion factor *G* in the form

$$I_{speculardiscrete,DFG} = \sum_{lights} \frac{D^i F^i G^i}{4\vec{N} \cdot \vec{V}} E^i_{light}$$

If *D* is the Beckmann distribution and *F* is the full Fresnel equations, this is the original Cook-Torrance model. This model is available as a technique "CookTorranceFull" in the accompanying HLSL effect file on the DVD-ROM. In the first step, *D* is replaced by the normalized Blinn-Phong distribution (see above), and *F* becomes a simplified Fresnel term. This model is available as "CookTorranceSimplified" in the effect file with the factors

$$D = \frac{1+e}{2}(\vec{N} \cdot \vec{H})^e,$$

$$F = \text{lerp}\left\{k_{specular}, \min(60k_{specular}, 1), (1 - \vec{L} \cdot \vec{H})^4\right\},$$

$$G = \min\left\{\frac{2\vec{N} \cdot \vec{H} \min(\vec{N} \cdot \vec{V}, \vec{N} \cdot \vec{L})}{\vec{L} \cdot \vec{H}}, 1\right\}.$$

The next optimization draws from work in [Kelemen01]. It is observed that the geometry factor divided by $(\vec{N}.\vec{V})$ could safely be replaced by a much simpler term, making *G* redundant. This implementation is available as "KelemenSzirmayKalos" in the effect file on the DVD-ROM. The simplified formula is:

$$I_{speculardiscrete,DF} = \sum_{lights} \frac{D^i F^i}{4(\vec{L}^i \cdot \vec{H}^i)^2} \vec{N} \cdot \vec{L}^i E^i_{light}$$

The final aggressive optimization exploits the fact that both the Fresnel factor and the remains of *G* are functions of $.\vec{L}.\vec{H}$ A simpler formula can be found that roughly exhibits a comparable combined effect, giving up Fresnel color shift. This version is linear in $\mathbf{K_{specular}}$ again, allowing it to be factored out. An implementation is available as a technique named "Optimized" in the effect file. The formula is as previously shown:

$$I_{speculardiscrete} = k_{specular} \sum_{lights} \frac{1+e}{8(\vec{L}^i \cdot \vec{H}^i)^3} (\vec{N} \cdot \vec{H}^i)^e \vec{N} \cdot \vec{L}^i E^i_{light}$$

References

[Blinn77] James F. Blinn, “Models of light reflection for computer synthesized pictures,” *Proc. 4th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 192–198, 1977.

[Cook81] Robert L. Cook and Kenneth E. Torrance, “A reflectance model for computer graphics,” *Computer Graphics*, vol. 15(3), pp. 307–316, August 1981.

[Kelemen01] Csaba Kelemen and László Szirmay-Kalos, “A microfacet based coupled specular-matte BRDF model with importance sampling,” *Eurographics* 2001 / N.N. short presenation.

[LuxPop] Thin film and bulk index of refraction and photonics calculations; available online at http://www.luxpop.com

[Phong75] Bui-Tuong Phong, “Illumination for Computer Generated Pictures,” *Communications of the ACM*, vol. 18(6), pp. 311–317, June 1975.