1

Physics of extraordinary transmission through subwavelength hole arrays

EVGENY POPOV AND NICOLAS BONOD

1.1 A brief reminder of the history of grating anomalies and plasmon surface waves

The recent history of the research and development around plasmon surface waves that was initiated by the work published in *Nature* in 1998 by Ebbesen *et al.* [1] looks like a ten-fold compressed version of studies initiated more than a century ago by Robert Wood with his discovery of anomalies in the efficiency of metallic diffraction gratings, now known as Wood's anomalies [2]. In 1902, R. Wood wrote: "I was astounded to find that under certain conditions, the drop from maximum illumination to minimum, a drop certainly from 10 to 1, occurred within a range of wavelengths not greater than the distance between the sodium lines," an observation that marked the discovery of grating anomalies.

The first period of the search for their explanation is marked by the attempt of Lord Rayleigh [3, 4] to link Wood's anomalies to the redistribution of the energy due to the passing-off (cut-off) of higher diffraction orders of the grating (transfer from propagating into evanescent type). As pointed out by Maystre [5], his prediction was all the more remarkable as the author first ignored the groove frequency of the grating used by Wood, and thus could not verify this assumption with experimental data.

It took more than 30 years for the second period of experimental and theoretical studies to establish another explanation of Wood's anomalies. In 1941, Fano [6] was the first to distinguish between two types of anomaly: (i) an edge anomaly, with a sharp behavior connected with the passing-off of a higher diffraction order, and (ii) an anomaly, generally consisting of a minimum and a maximum in the efficiency, which appears in a much broader interval. The second type of anomaly was described by Fano as a resonance one, linked with the excitation of a guided (leaky) wave along the grating surface. Hessel and Oliner [7] published a

Structured Surfaces as Optical Metamaterials, ed. A. A. Maradudin. Published by Cambridge University Press.

pioneering paper that shows for the first time, using a theory based on an analysis of electromagnetic scattering from a generic model of a periodic structure yielding a simple closed form solution, that Wood's anomaly resonances are of two types: one due to branch point singularities that correspond physically to the onset of a new propagating spectral order (first indicated by Lord Rayleigh), and the other due to pole singularities that correspond to the condition of resonance for leaky surface waves guided by the structure. In addition, Hessel and Oliner developed a so-called phenomenological approach to the resonant anomalies that permitted describing the anomaly by a very small number of physical parameters: a pole of the scattering matrix, a zero of the diffracted amplitudes, and smoothly varying coefficients.

The third period that continues even today contains studies in three different directions. First are the grating manufacturers and instrumental optics users for whom it is strongly advisable to avoid anomalies because of their devastating effects for spectral instrument performance. Second, the excitation of surface plasmons can lead to a total absorption of the incident light, a phenomenon predicted and observed by Hutley and Maystre [8] in a single polarization for classical gratings with one-dimensional (1D) periodicity, with the magnetic field vector parallel to the groove direction (TM, transverse magnetic, polarization). By the use of crossed gratings with two-dimensional periodicity, it is possible to obtain total light absorption in unpolarized light [9, 10]. In both cases, it was necessary to use gratings with subwavelength periods that can support only the specular (zeroth) reflected order. The total light absorption by metallic gratings evidenced that the coupling between the incident light and metals can be strongly enhanced by the excitation of surface plasmons, and this effect opened the way to many applications based on the strongly enhanced light–matter interaction. Third, as the electromagnetic field is localized in the vicinity of the metallic surface, light absorption leads to very strong optical intensities at the surface. As any surface presents natural roughness that can excite the surface wave, the field enhancement obtained under specific conditions was sufficient to provide a proper physical explanation of the surface enhanced Raman scattering (SERS) effect [11]. The same effect is also used to enhance otherwise weak nonlinear phenomena [12]. Biosensors based on surface plasmons are highly dependent on the refractive index of the surrounding media. Binding or adsorption of molecules on the metallic surface induces a change of the local refractive index of the dielectric medium, so that such biosensors can be called refractometric sensors [13–17].

1.2 Generalities of the surface waves on a single interface

Before discussing in detail the historical development of the studies of enhanced light transmission through arrays of holes in a metallic screen, let us introduce

several notations and basic principles. Let us consider a plane metal–dielectric interface in the $x0z$ plane that separates two nonmagnetic media with relative dielectric permittivities ε_1 and ε_2. In TM (transverse magnetic) polarization and incidence in the $x0y$ plane, the two components of the electric and magnetic field that are parallel to the interface, and thus continuous across it, are:

$$\begin{aligned} \omega\mu_0 H_z &= \exp(ik_x x)\left[\exp(-ik_{1y}y) + r\exp(ik_{1y}y)\right], \\ E_x &= \frac{k_{1y}}{k_0^2\varepsilon_1}\exp(ik_x x)\left[\exp(-ik_{1y}y) - r\exp(ik_{1y}y)\right] \end{aligned} \tag{1.1}$$

in the cladding and

$$\begin{aligned} \omega\mu_0 H_z &= t\exp(ik_x x)\exp(-ik_{2y}y) \\ E_x &= \frac{k_{2y}}{k_0^2\varepsilon_2}t\exp(ik_x x)\exp(-ik_{2y}y) \end{aligned} \tag{1.2}$$

in the substrate. The first terms in the brackets in Eqs. (1.1) correspond to the incident wave, and the second terms correspond to the reflected wave, with r the reflection coefficient for the magnetic field amplitude. The transmission coefficient is denoted by t. Note that k_x is the x-component of the incident wavevector, and that k_{1y} and k_{2y} are the y-components of the wavevectors in the cladding and in the substrate:

$$k_{jy} = \sqrt{k_0^2\varepsilon_j - k_x^2}, \qquad j = 1, 2, \tag{1.3}$$

with k_0 the free-space wavenumber. The Fresnel reflection coefficients depend on the polarization and have the following form for transverse electric (TE) polarization:

$$r_{TE} = \frac{k_{1y} - k_{2y}}{k_{1y} + k_{2y}}, \tag{1.4}$$

and for TM polarization:

$$r_{TM} = \frac{k_{1y}/\varepsilon_1 - k_{2y}/\varepsilon_2}{k_{1y}/\varepsilon_1 + k_{2y}/\varepsilon_2}. \tag{1.5}$$

It is well-known that r_{TE} has neither a pole nor a zero. In contrast, when both ε_1 and ε_2 are real and positive, there is a zero of r_{TM} called the Brewster effect. There also exists a pole (a zero of the denominator) if one of the media is a dielectric and the other a metal, a pole that corresponds to a surface wave that can propagate along the interface. When expressed in terms of the wavevector component parallel to the interface, the solution has the same form for the Brewster effect and the pole,

$$k_x = k_0\sqrt{\frac{\varepsilon_1\varepsilon_2}{\varepsilon_1 + \varepsilon_2}}, \tag{1.6}$$

due to the ambiguity of the choice in Eq. (1.3) of the sign of the square root for complex arguments. Indeed, combining Eqs. (1.3) and (1.6), we obtain the classical form of the Brewster angle in the incident medium: $\tan\theta_1 = k_x/k_{1y} = \sqrt{\varepsilon_2/\varepsilon_1}$. When the second medium is a metal with the real part of ε_2 negative and smaller than $-\varepsilon_1$, the real part of k_x in Eq. (1.6) is greater than the wavenumber $k_0\sqrt{\varepsilon_1}$ in the upper medium; i.e., the wave is evanescent in the cladding (and inside the metal), with increasing distance from the interface, representing a surface wave with a propagation constant equal to k_x, a solution that we shall note as k_g, the index g standing for "guided," and its normalized propagation constant will be denoted as $\alpha_g = k_g/k_0$. As ε_2 always has a non-zero imaginary part due to the absorption losses in the metal, the surface wave decays as it propagates. Quite often the negative permittivity is due to the collective oscillations of the free electron plasma in the metal, which gives the names surface plasmon or plasmon surface wave (PSW) to these surface waves. As an incident electric field creates polarization states of the plasma, some authors call this wave a surface plasmon polariton (SPP). In the case of a polar crystal/vacuum interface, the corresponding surface waves represent surface phonon polaritons. They all have common properties from an electromagnetic point of view, although the background solid state physics can be quite different. As they represent a zero of the denominator of r_{TM}, they are solutions of the homogeneous problem – a scattered field with zero incident field – and thus represent proper (eigen) modes of the system.

When considering an idealized presentation of a perfectly conducting metal with $\varepsilon_2 \to -\infty$, the propagation constant in Eq. (1.6) becomes equal to the vacuum wavenumber, and thus the solution represents a plane wave propagating parallel to the interface inside the cladding, with its electric field vector perpendicular to the surface; i.e., the solution is not localized to the surface.

When α_g is greater than n_1, such a wave cannot be excited with an incident plane propagating wave. The excitation of the surface wave is possible through the Kretschmann configuration [13]: the surface plasmon is excited on the lower surface of a metallic layer having on its upper surface a prism with refractive index higher than the index of the substrate in order to match the horizontal component of the incident wavevector to the real part of the PSW wavenumber on the lower interface. The surface plasmon is then coupled to the incident light by tunneling through the metallic layer. A surface plasmon can also be excited in a prism coupler in the Otto configuration. In that case, the metallic film is coated on a glass substrate. The strength of excitation depends on the distance between the prism and the metallic layer. In both cases, the reflection of light is strongly attenuated when the surface plasmon is coupled to the incident wave, and the angle of incidence where the absorption is maximum depends on the refractive index of the dielectric medium surrounding the metallic layer.

Much more efficient coupling occurs when gratings are used with a periodicity that serves as a generator of wavevectors parallel to the surface. An incident plane wave generates an infinite number of diffraction orders. In the case of 1D periodicity in the x-direction with period d, the wavevector component of each diffraction order m is given by the grating equation:

$$k_{1x,m} = k_{1x,0} + mK, \qquad K = \frac{2\pi}{d}, \tag{1.7}$$

where $k_{1x,0}$ is equal to the x-component of the incident wavevector and m is an integer. If, for a certain value of m, $k_{1x,m}$ is close to k_g, a surface wave can be excited. A simplified notation leads to the condition of excitation:

$$Re(\alpha_g) = \sin\theta_i + m\frac{\lambda}{d} \tag{1.8}$$

if the upper interface is air (more precisely, a vacuum).

The coupling of the incident wave to the surface wave (mode) is reciprocal; i.e., the surface wave can be radiated into propagating diffraction orders in the cladding following Eq. (1.7). This phenomenon is called leakage and the surface wave becomes a leaky one, which leads to an increase of the imaginary part of α_g. Another important feature that is not obvious from Eq. (1.8) is that the real part of the propagation constant, as well as the electromagnetic field distribution of the surface wave characteristics, are modified by the presence of surface corrugation. Another possibility to excite a PSW realizes itself in SERS, where the surface roughness scatters the incident plane wave into waves with different k_x, and, in particular, with $k_x > k_0$, with part of the incident energy coupled to the PSW.

1.3 Extraordinary transmission and its first explanations

Just as Robert Wood 95 years earlier was astounded by his experimental observation, Thomas Ebbesen and his collaborators found it quite surprising to observe that when light tries to pass through an array of holes of subwavelength dimensions in an optically thick (opaque) metallic sheet (Fig. 1.1(a)), and whose entire area is much smaller than the total illuminated surface, there are spectral regions with anomalously high transmission (Fig. 1.1(b)) compared with the predictions of classical diffraction theory. The surprise was so great that it prevented the publication of the results for almost ten years from their first observation [18] of the effect in the NEC laboratories.

As with Wood's anomalies, it is possible to separate the studies on this extraordinary transmission into three much shorter and more dynamic periods. In 1998, in contrast to the situation at the start of the twentieth century, electromagnetic theories

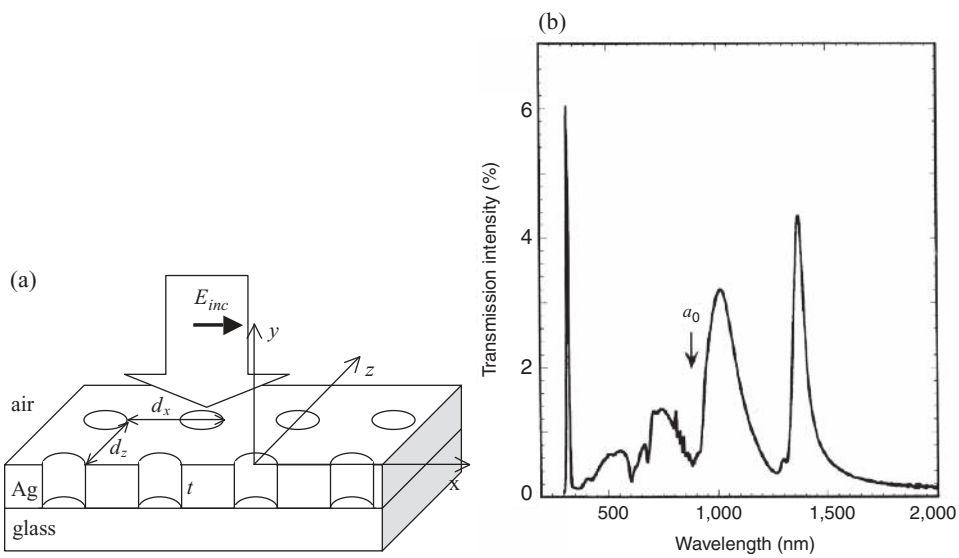

Figure 1.1. (a) Schematic representation and notations of a two-dimensional hole array perforated in a metallic screen deposited on a glass substrate and illuminated from above with a linearly polarized incident wave. (b) Spectral dependence of the transmission of the structure presented in (a), with $d_x = d_z = 0.9\ \mu m$, $t = 0.2\ \mu m$, and a hole diameter of $0.2\ \mu m$ [1]. Reprinted with permission from Macmillan Publishers Ltd. © 1998.

of gratings (1D or 2D) were largely developed. The understanding of the role of the PSW (and surface waves in general) in grating anomalies, field enhancement, etc., was much deeper, and the number of scientists working in the field incomparably larger. The end of the Cold War moved large human resources from defense microelectronics, solid state and high energy physics into optics, creating neologisms like photonic crystals, photonics, metamaterials, etc., causing, for instance, the change of *Optics News* into *Optics and Photonic News*. In addition, production resources such as optical photolithography, focused ion-beam and laser-beam writing and etching, became more available in optics laboratories, which permitted structuring metals and dielectric media at the nanometer scale and developing photonic devices able to control light at the subwavelength scale. Already in the original paper [1], the authors clearly indicated that the transmission enhancement appears at spectral positions closely given by PSW excitation by a bi-periodic structure, but they were not satisfied by this qualitative explication. The publication of these results by Ebbesen and coworkers in *Nature* immediately strongly impacted the newly enlarged optical community that was closely interested in photonics, to find a new but similar interest in plasmon surface waves, giving birth to another neologism, plasmonics.

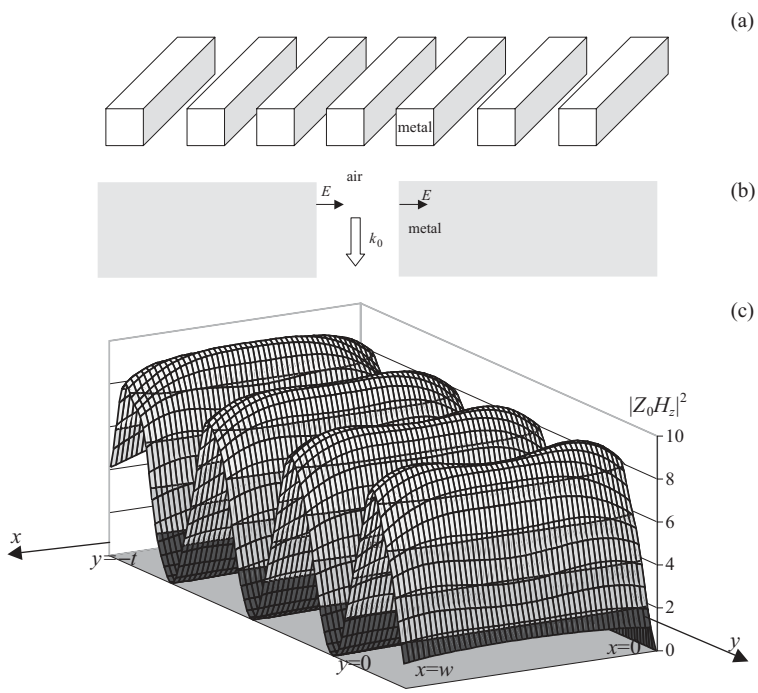

Figure 1.2. (a) Slit grating having a one-dimensional periodicity and characterized by vertical straight channels. (b) Propagation of a TM electromagnetic field inside the slit with perfectly conducting walls. The electric field vector perpendicular to the slit walls satisfies the boundary conditions there and the vertical wave propagates as in free space. (c) Propagating character of the electromagnetic field inside a narrow slit ($w = 40$ nm). Reprinted with permission from ref. [22]. © 2000, American Physical Society.

The first numerical results were so close to the experimental enhancement that doubts were generated whether the transmission increase was so extraordinary. The main characteristic of this first period (see, for example, refs. [19] and [20]) was the use by theoreticians of gratings with 1D periodicity made of periodic slits in a metallic screen (Fig. 1.2(a)). The results were quite nice, with a clearly visible flow of the electromagnetic field inside the slits, so that some authors started to indicate the decisive role of another wave – a vertical plasmon wave that propagates inside the slits of the metal–dielectric interface – that is responsible for the enhanced transmission, acting simultaneously with the grating-induced resonances of the horizontally propagating PSW, excited at spectral positions given by Eq. (1.8). Figure 1.2(b) represents such a vertical wave which propagates inside the slits as in free space for perfectly conducting walls. For metals with a finite conductivity, this wave represents a hybrid wave that can propagate in the vertical direction formed by two coupled plasmons on the slit walls [21].

The finite conductivity of the metal changes the boundary conditions on the vertical walls, but even for very narrow slits and a silver grating, the TM electromagnetic field preserves its propagating nature, as seen in Fig. 1.2(c).

These idyllic conclusions were common for the first period of about two to three years following 1998. The main problem was that they were not applicable to the geometry involved in the initial experiment made by Ebbesen *et al.*, where the slits were replaced by small holes. In fact, the enhanced transmission through slit metallic gratings (or gratings having similar grooves) in TM polarization had been known for quite a long time and resulted in commercially available wire-grating polarizers (see, for example, [10]). Such gratings (as represented in Fig. 1.2(a)) with subwavelength periods small enough to support just the specular reflected and transmitted orders, reflect incident light of TE polarization almost completely, while light of TM polarization can be transmitted almost totally for a proper choice of grating parameters. The reason lies in the existence of a waveguide mode inside each slit, which in TM polarization has no cut-off wavelength. Let us consider a slit, neglecting the absorption losses inside the metal walls. As in the case of the plane horizontal interface between a lossless metal and a dielectric in TM polarization, a vertical interface also supports a wave of plane-wave type inside the dielectric propagating parallel to the interface with a magnetic field vector parallel to the interface (Fig. 1.2(b)). The same wave can propagate inside slits with lossless walls, as it satisfies the boundary conditions on both walls, whatever the width of the slit. The mode is characterized by a real propagation constant in the vertical direction (neglecting absorption losses, as assumed). On the upper and lower interfaces (Fig. 1.2(a)), the mode is reflected backwards in the slit, and is partially transferred into propagating waves in the cladding and in the substrate, representing a Fabry–Perot resonator. If the system is symmetrical (the same optical index of the substrate and the cladding, as in wire polarizers), the Fabry–Perot resonance maxima can reach 100% in transmission. In contrast, in TE polarization the corresponding slit mode has a cut-off, because the electric field vector is parallel to the slit walls. The electric field vanishes on both walls and satisfies the following relation (Fig. 1.3(a)):

$$E_z \sim \sin\left(\frac{\pi}{w}x\right), \tag{1.9}$$

so that the y-component of its wavevector, given by

$$q_y = \sqrt{k_0^2 - \frac{\pi^2}{w^2}} = \pi\sqrt{\frac{4}{\lambda^2} - \frac{1}{w^2}}, \tag{1.10}$$

becomes imaginary for small widths w; i.e., the mode is evanescent in the vertical y-direction if the slit width is smaller than $\lambda/2$. Thesmaller the width, the faster the

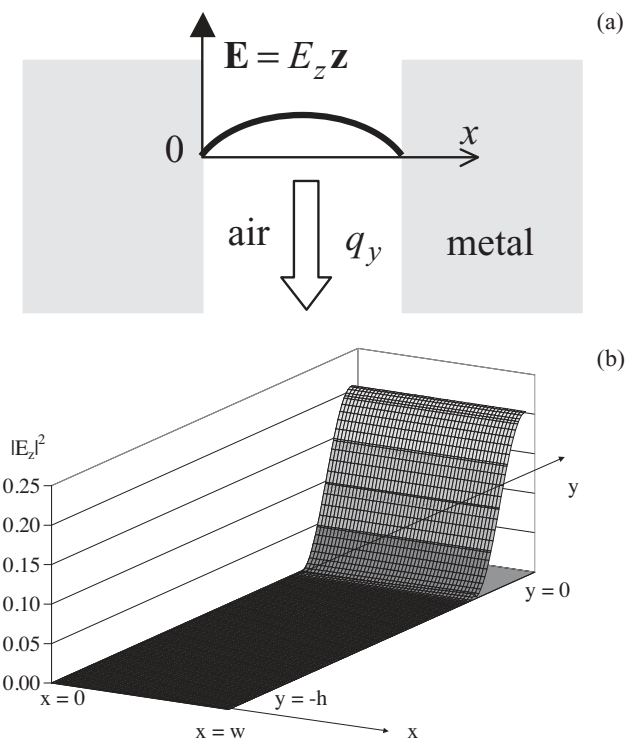

Figure 1.3. (a) TE mode inside a slit with perfectly conducting walls. The x-dependence of E_z is given by the thick line, and it has to vanish on the walls. (b) A map of the evanescent TE mode inside a subwavelength slit ($w = 40$ nm) with silver walls. The slit is so small compared to the wavelength that the sinusoidal dependence that appears for perfectly conducting walls (a) cannot be distinguished in the x-dependence. Reprinted with permission from ref. [22]. © 2000, American Physical Society.

exponential decrease of the mode amplitude inside the slit depth. The narrower the slit and the thicker the metal layer, the smaller the amount of transmitted energy, and thus the polarizing properties of the device.

When finite conductivity is taken into account, the cut-off width is slightly smaller for finitely conducting walls than the $\lambda/2$ value obtained from Eq. (1.10). In addition, very narrow slits absorb an electromagnetic field, and the transverse variation of the field becomes very weak, as can be observed in Fig. 1.3(b) for silver walls and a 40 nm slit width.

In contrast to what happens with slits, holes with a finite width in both directions of their cross-section do not support modes without cut-off; i.e., below a given width of the hole, the field of the modes inside is always evanescently decreasing. Although obvious, these considerations were not taken into account in the first modelizations, when the hole array was replaced by periodic slits. However, these

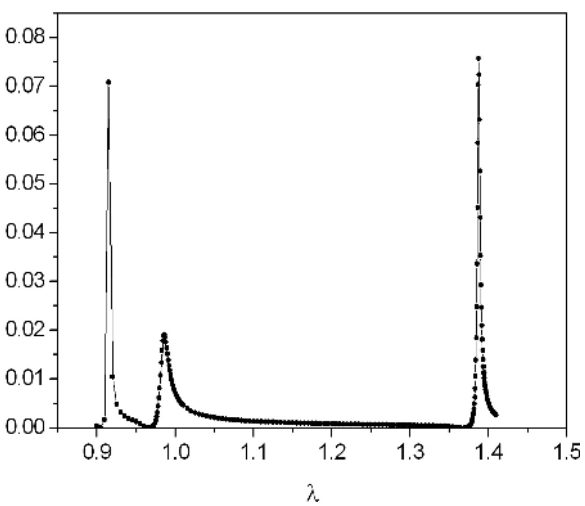

Figure 1.4. Computed spectral dependence of the transmission intensity of a square-hole array in a 200 nm thick silver screen deposited on a glass substrate. Reprinted with permission from ref. [22]. © 2000, American Physical Society.

first works were relatively easily carried out from theoretical and computational points of view, and the large number of studies based on this assumption attracted substantial interest to metallic gratings and surface plasmons.

1.4 The role of the evanescent mode

The second period of the studies of extraordinary transmission started with the first rigorous electromagnetic modeling of the array of holes with finite subwavelength cross-section dimensions [22]. The numerical results are similar to the experimental observations (Fig. 1.4).

The grating period in both directions is the same and equal to 0.9 μm, the cladding is air, and the substrate is glass. The metal is silver 0.2 μm thick, and the holes have a square cross-section with a width of 0.25 μm, much below the cut-off dimensions for the spectral interval under study. Two peaks are clearly distinguished, the shorter-wavelength one, lying around 1 μm, corresponds to the excitation of PSW on the upper air–silver interface. The long-wavelength peak is due to the excitation of PSW on the lower glass–silver interface.

For an infinitely conducting metal, the fundamental TE mode of the hollow square waveguide formed inside each hole has a propagation constant of the order of $q_y \approx i11\ \mu\text{m}^{-1}$, which corresponds to a decay constant in the y-direction, $\gamma = Im(q_y)/k_0 \approx 2.5$. When the finite conductivity of the metal is taken into account,