The future of atomic resolution electron microscopy for materials science

J.C.H. Spence*
Department of Physics and Astronomy, Arizona State University, Tempe, AZ 85287, USA
Accepted 9 August 1999

Abstract

The field of atomic-resolution transmission electron microscopy and its application to materials science is reviewed. This technique, whose spatial resolution is now about one Angstrom, is valuable wherever nanoscale characterization of materials is needed. The history of the subject is briefly outlined, followed by a discussion of experimental techniques. Resolution-limiting factors are summarized, together with the underlying theory of image formation. Seven promising approaches to super-resolution are reviewed. The statistical principles of quantitative image analysis and defect modelling are outlined for both HREM and STEM. Methods for obtaining defect energies from images are discussed. The review ends with a summary of some recent applications, including such topics as the Fullerenes, nanotubes, dislocation kink imaging, superconductors, atomic-resolution imaging of whole semiconductor devices, the study of atomic defects in mediating first-order phase transitions, collosal magnetoresistance, ceramic interfaces, quasicrystals, imaging of surfaces, glasses, catalysts and magnetic materials.

1. Introduction

Over the past three decades the transmission electron microscope has become the instrument of choice whenever questions of microstructural characterization arise in materials science. As scientists have learnt to control the microstructure of modern materials on an ever finer scale (down to nanometer dimensions), so the technique of atomic resolution transmission electron microscopy has become increasingly important. This technique has become known simply as high resolution electron microscopy (HREM). Increasingly, it has been combined with the many other analytical signals available on the electron microscope, including energy-loss spectroscopy (which gives similar information to soft X-ray absorption spectroscopy), X-ray microanalysis (similar to X-ray fluorescence microscopy, but from thinned samples), microdiffraction and cathodoluminescence. In this review we will concentrate solely on HREM. Just as the atomic force microscope, the scanning tunneling microscope and the atom probe have become the main tools for the study of defect structures on the surface of materials, so HREM has become the method of choice for the study of defect structures in the bulk when atomic resolution is needed. These defects control the properties of most materials. For example, mechanical properties are controlled by the defects responsible for stress concentration, electrical properties are influenced by defects which limit charge-carrier lifetimes in semiconductors, and first-order phase transitions are mediated by moving defects.
‘What materials problems has HREM ever solved?’ is a question sometimes asked by skeptics. The solution of materials problems often involves both diagnosis (characterization) and treatment (modified synthesis). In a few fields, HREM has now become the only possible characterization method; in others it is an essential accompaniment to others. In many fields in which hetrostructure interfaces are important, it has become almost obligatory to include in publications a lattice image of the interface, to confirm that it is atomically abrupt. As one quick example, Fig. 1 shows the kind of whole-device imaging now used in the semiconductor industry — with device sizes approaching a few tens of nanometers (and gate oxides a few nanometers wide), only HREM can reveal the location of flaws in the internal atomic structure of devices on this scale. Similarly, the intergranular phases which can dramatically alter the mechanical properties of fine-grained sintered materials and metal–ceramic interfaces can only be seen by HREM. These phases may be only a few nanometers wide. There are many other examples, from the intergrowth of micro-phases in oxides and minerals to twinning in the oxide superconductors, the identification of polytypes, the study of quasicrystals and the abruptness of metal-semiconductor and multilayer interfaces. HREM may ‘solve’ a materials problem by providing an explanation for properties at the atomic level — for example the recent observation of three-fold dissociation of the core of screw dislocations in Mo [1] explains the high flow stress of BCC metals at low temperatures. Finally, entirely new phenomena, such as bucky-tubes, and the growth of diamond crystals within bucky-balls (Fig. 15) have been discovered by HREM.

The first major discovery which resulted from HREM was an explanation, at the atomic level, for the causes of non-stoichiometry in complex oxides [2]. The first major discovery for materials science due to transmission electron microscopy was the direct observation of dislocations (see [3] for a review). The most recent major discovery was the observation and synthesis of bucky-tubes, which resulted entirely from HREM observation [4].
Theory and simulation play an increasingly large role in modern materials science, and there is a continuous effort underway to relate the atomic structure of defects to bulk behavior. Ambitious computational schemes have recently been launched which attempt, by computer modeling, to span these length scales in an attempt to predict effects such as work hardening, fracture toughness, hardness numbers or embrittlement. Here HREM can provide essential information on the density, type and even (in favorable cases) the atomic structure of the defects responsible. The role of defects in phase transitions is another area where HREM has provided crucial information on transformation mechanisms.

A superb collection of HREM images taken from all areas of materials science has just appeared [5]. The interpretation of the images and their implications for materials properties are also fully discussed. The field of HREM has recently also been reviewed by Smith [6]. An up-to-date account of the state of the art in HREM can be found in vol. 190, parts 1 and 2 of the Journal of Microscopy for 1998, which, together with the text edited by Buseck et al. [7] and the Shindo–Hiraga book provide excellent starting points and overviews of current issues in HREM. Many excellent reviews of the theory of electron scattering and imaging in thin crystals exist [3,7–12].

With the rapid growth of the world-wide web, the nature of review articles (and publishing itself) must change. Most recent papers on any specific topic can now be found rapidly through the on-line library web search services, using topics, keywords and authors’ names. What is difficult to find is material before about 1980, or a means of organizing material relevant to a particular topic. With that in mind, this article aims: (i) to review the basic theoretical ideas behind modern atomic-resolution electron microscopy, (ii) to provide some historical perspective, (iii) to describe selected examples of the highest quality work in applications from most areas of materials science, and (iv) to summarize the most exciting future directions. For details of work on a particular materials system, the reader is referred to the references in the papers cited, to the references in the other review articles cited here, and to the web. The use of Science Citations index (Institute for Scientific Information; http://isi1.med.iacnet.com/ISI/CIW.cgion) on the web is one efficient method of searching either by author or topic. Alternatively, all the papers which cite a given paper may be found using ISI. By following these earlier citations one rapidly builds up a collection of papers on a given topic. Searches may also be performed for the home-page of a particular scientist.

2. History

HREM can be said to begin with publication of the first lattice image of a thin crystal by J. Menter in 1956 [13]. This was a two-beam interference image of copper phthalocyanine, with a spacing of 1.2 nm. Theoretical work had begun much earlier, notably with the important paper of Scherzer in 1949 [14] defining the wave-front aberration function and the resolution of the electron microscope, and with early work on electron holography [15]. The electron microscope itself was invented in 1931 by Ruska and Knoll. Throughout the sixties, improvements in electronic and mechanical stability lead to instruments with a ‘line’ resolution of less than 1 Å before the end of the decade [16], but it was widely held that these few-beam images provided more information about the electron microscope than they did about the sample, and it proved almost impossible to extract useful information from them regarding the sample. They were analyzed using the two-beam dynamical theory [17]. Three breakthroughs occurred in the early seventies which gave renewed confidence to the field — the observation of individual atoms by scanning transmission electron microscopy (STEM) [18], the first observations of two-dimensional lattice images containing useful information...
on crystal chemistry and defects [2,19], and the widespread adoption of Thon’s ‘diffractograms’ to 
measure lens aberrations and focus settings (see [10] for details). The electron microscope point 
resolution has improved steadily from about 0.34 nm (a common spacing in graphite) at that time to 
0.1 nm today, sufficient to resolve individual columns of atoms in most thin crystals.

Over the past decade three crucial advances in instrumentation have produced major advances in 
the field — the use of field emission electron sources, the use of charge-coupled device (CCD) 
detectors, and the introduction of imaging energy-filters. Finally, dramatic increases in computing 
power has made image simulation (including all multiple scattering effects) routine, and the 
automated alignment and control of electron microscopes possible. Together these advances now 
allow much more quantitative analysis of high resolution images, and the more accurate 
measurement of the experimental parameters on which image simulations depend.

Two approaches to the theory of HREM are in use — those intended for the thinnest samples 
and based on the weak phase object approximation [20], and those based on the dynamical theory, 
intended for thicker crystals in which multiple scattering cannot be neglected. Several forms of 
dynamical theory have been developed independently, including those based on the path-integral 
formulation of quantum mechanics, on the Born series and on Green’s function methods (the 
relationship between them, the multislice and Sturkey’s Scattering matrix method is demonstrated in 
[10]). But the multislice formulation [21] has proven most efficient for computational purposes, 
while the original Bloch-wave theory of H. Bethe [22] is also sometimes still used, particularly for 
diffraction calculations. For those interested in writing new multislice programs, reviews of the 
multislice method used for HREM computations can be found in [10,23–25] and [8]. The text by 
Spence and Zuo [26] contains a complete fortran listing, and the new book by Kirkland [27] contains 
an excellent analysis of HREM image simulation. It includes a CD for Mac or PC containing 
multislice programs in C with image outputs in TIFF file format. This book is strongly 
recommended.

3. Sources, stages, filters, detectors, sample preparation

The use of a field-emission source for HREM has two beneficial effects — the energy spread of 
electrons $\Delta E$ in the beam is reduced, thereby improving resolution by reducing chromatic aberration 
(see Eq. (4.10)), while the improved spatial coherence (leading to a coherently filled illumination 
aperture) also improves resolution by a factor $\sqrt{2}$ [28]. The smaller illuminated area is not usually a 
disadvantage for HREM work, while the ability to focus down to a small probe allows 
microdiffraction patterns, energy loss spectra and characteristic X-ray spectra to be collected from 
regions which can be correlated with the HREM image. Cold field emission sources can supply 
currents as high as 1 nA into a probe diameter of about 1 nm$^2$. For the Z-contrast STEM method 
described in Section 8 the field-emission source is essential.

Top-entry stages remain the most stable for the ultimate resolution, however the many other 
advantages of side-entry sample holders, together with improvements in their design, have meant 
that these have largely replaced the top-entry type. In well-equilibrated systems, the sample 
movement due to thermal expansion of the stage may be as low as a few Angstroms per minute. A 
variety of heating, cooling and environmental (controlled atmosphere) stages have become available, 
with the result that near-atomic resolution imaging has become possible at temperatures as high as 
800°C. The use of video recording in this work minimizes the effects of thermal motion of the 
sample. Piezo-driven stages are now appearing which are dramatically effective in counteracting the 
effects of thermal motion (drift). The alternative method of deflecting the final image
electrostatically in such a way as to counteract slow sample motion is limited in range, but otherwise effective also. An image feature is used to fiducialise the feedback system.

The value of imaging energy-filters for HREM work has been debated. One view holds that any sample thick enough to generate significant inelastic scattering will be too thick to provide interpretable (or alignable) HREM images. Thermal diffuse scattering, which involves large angle scattering and very small energy losses, cannot be removed by filtering. It may be shown to make a high resolution contribution to the detail in HREM images [29]. By contrast, the weak background of inelastic scattering due to plasmon and single-electron losses which may be removed by filtering contributes only a low-resolution, out-of-focus background intensity variation to HREM images in thin samples. For most other modes of the TEM, the energy filter has greatly beneficial effects. Elastic energy filtering, for example, greatly improves the contrast of microdiffraction patterns [26] from thicker samples. These may be used to determine the symmetry elements of a thin crystal, which may then be applied to lattice images from thinner regions, resulting in improved signal-to-noise ratio.

Two major advances have occurred in detector technology for HREM in recent years — the CCD (charge-coupled device) detector and the Image Plate system. Direct detection of HREM images by CCD camera and Yttrium–Aluminium–Garnett (YAG) scintillator was first demonstrated in 1988 [30], and a highly successful commercial system soon followed [31]. An analysis of the detective quantum efficiency (DQE), modulation transfer function (MTF) and sensitivity of these cooled, slow-scan devices for TEM applications can be found in [32,33]. A comparison of the image plate and CCD systems can be found in [33,34]. In summary, the CCD system provides for immediate image processing at the microscope and also allows pixel-to-pixel registry between successive images. However, the MTF contains a sharp peak at the origin, corresponding to a broad tail in the impulse response due to light scattered sideways in the scintillator screen. (This is usually a single-crystal of YAG). This causes ‘cross-talk’ between adjacent pixels, significantly reducing the number of independent picture elements. As it involves only low spatial frequencies, its effect may be greatly reduced by numerical deconvolution, without appreciable noise amplification. However, the characterization of a CCD system (measurement of MTF, DQE, sensitivity and dynamic range) has now become a major part of any quantitative HREM project, and each camera from a manufacturer must be characterised separately. Methods for doing this are described in [33]. (The MTF may also be measured using fringes formed by a Fresnel biprism, if one is fitted for electron holography applications. The fringe period must be varied, by varying the magnification, with constant fringe contrast). The image plate (IP) system offers many more pixels, larger dynamic range (in most cases), great flexibility and possibly lower cost than a CCD camera, since a single IP reader can service many microscopes. The flexible ‘electronic film’ used (of standard size) can be re-used many times, and requires no wet processing. The performance of this system is limited by that of the IP reader rather than the IP itself. Our measurements suggest that the IP system has better DQE at low dose than the CCD camera, however, immediate image processing is not possible (plates must all be exposed and read in groups), and pixel-to-pixel registry is not possible. Thus the IP system appears to be ideal for radiation-sensitive materials, for the study of weak diffuse scattering in diffraction patterns, and for holography, where very large numbers of pixels are needed.

Sample preparation has always been the most tedious part of HREM work, however many recently developed techniques have improved matters considerably. The list of chemical and electrochemical thinning methods given in [3] remains the most complete, in addition to the texts by Goodhew [35] and Williams and Carter [36]. Any text on metallography can be a useful source for finding chemicals which polish metals. For most other materials the trend has been toward the use of dimple grinding followed by ion-beam thinning [37] — details of these methods can be obtained
from a web search of the equipment manufacturers, listed, for example in the MSA Bulletin or at
conference exhibitions. A diamond wheel saw (or wire saw) is used to cut out the required sample.
The diamond saw can be fitted with a goniometer head so that orientation settings can be transferred
from an X-ray diffractometer, if single crystals are used. For most ceramics and semiconductors,
preliminary thinning will be needed down to about 30 microns, before ion-beam thinning is
attempted. For this purpose a good soluble glue (e.g. Crystalbond [38]) is needed — the sample is
glued onto a thick glass disc (e.g. optical flat or lens blank) whose thickness has been accurately
measured. The sample can then be ground down by hand quickly using a grinding wheel. For
semiconductors, chemical thinning may still be preferred if smooth surfaces and low temperatures
are needed [39]. One recent paper reports the effect of oxygen-deficient solutions of NH₄F on silicon
(1 1 1) surfaces, which produces large atomically flat terraces as observed by AFM [40]. The special
methods which must be used when preparing cross-section views of semiconductor interfaces are
described in [41].

Two recent development have had a large impact in TEM sample preparation — the tripod
thinner and the plasma cleaner. A few minutes in a radio-frequency excited plasma of argon and
oxygen gasses is found to dramatically reduce contamination, following well established ‘ash-
removal’ practice in the coating industry. When combined with the very high currents available from
the field-emission gun, the result is, that if a focussed probe is left on a thin area of sample for a few
minutes, instead of generating a mound of contamination, holes are drilled by the beam in the sample
[42]. It appears that the plasma cleaning does not remove oxide layers or smoothen sample surfaces,
but the additional surface ‘Fresnel noise’ due to contamination is reduced, thereby improving HREM
contrast. The tripod thinner (available from several companies) has proven invaluable for making
samples of semiconductor devices (such as Fig. 1, which was obtained in this way), where very large
thin areas are needed in pre-defined regions. However, it requires skill and practice to use well.

The focussed-ion beam milling machine (FIB) is being evaluated in many labs for use as a TEM
sample preparation device [43]. This imaging device allows thinning to be performed on selected
regions, however, implantation of the Gallium ions used has caused difficulties, and an amorphous
layer may be formed on crystalline material.

The simplest way to handle sample fragments is to use a vacuum-tweezer nozzle (similar to a
hypodermic syringe) or glass pipette connected to a plastic tube. By sucking gently on the tube, it
will be found that fragments which can only be seen under an optical microscope can be picked up,
positioned and released more gently than by the use of tweezers. A vacuum pump is thus not needed.
An important recent development for biological microscopy has been the preparation of sample grids
by new lithographic methods — these ‘Quantifoil’ grids contain a periodic array of 2 µm diameter
holes in carbon films (available from Ted Pella). Many ceramic samples can be obtained for HREM
study just by crushing them under alcohol in a mortar and pestle, then collecting fragments on a grid.
The time-honoured method of collecting burning magnesium metal smoke on a grid produces some
of the best samples in the form of MgO cubes and platelets. These are an excellent test sample,
which can be made in a few minutes. They possess the great advantage for HREM that, since they
form perfect cubes, a cube viewed along [1 1 0] displays a projection in which the wedge angle and
hence the thickness of the sample are precisely known at every point in the atomic-resolution image.

4. Theoretical background and resolution limits

The purpose of this section is to identify the factors which limit the resolution of modern HREM
instruments, using the simplest theory. A comprehensive review of resolution-limiting factors can be
Many approximations are required to establish the following results, which are those now in common use. A collimated electron beam with an energy of a few hundred kilovolts is assumed to traverse a thin sample, whose thickness is generally less than about 20 nm. This sample may be crystalline, in which case Bragg beams are generated, and these beams are then imaged by an electron lens. For samples sufficiently thin that multiple scattering may be neglected, and beam energies sufficiently high that excitation error effects can be neglected, the electron wavefunction across the exit face of the sample may be written in the first Born approximation as

$$
\psi_e(x, y) = \exp(-i\sigma \phi_p(x, y)) \approx 1 - i\sigma \phi_p(x, y)
$$

(4.1)

where

$$
\phi_p(x, y) = \int_{-t/2}^{t/2} \phi(x, y) \, dz
$$

is the electrostatic potential projected in the beam direction. Here $\sigma = 2\pi me\lambda/\hbar^2$ is the positive interaction constant, with relativistically corrected values of the electron wavelength $\lambda$ and mass $m$. The Ewald sphere has been approximated by a plane, i.e. all the excitation errors in the zero-order Laue Zone have been set to zero. The second term on the right-hand side of Eq. (4.1) is thus the phase shift introduced by the sample, which we treat as a pure phase object. Due to the unavoidable introduction of lens aberrations, the final image is convolved or blurred by a ‘point spread function’ given by the last term in Eq. (4.2) (The aberrations of the first lens, the objective lens, is most critical, since here the scattering angles are largest). The image intensity distribution recorded on the detector is then

$$
I(x, y) \approx 1 + 2\sigma \bar{\phi}_p(-x, -y) \otimes \text{Im}\{\sin(\chi(u))P(u)\}
$$

(4.2)

Here Im denotes Fourier transform, $\otimes$ denotes convolution, and the spatial frequencies transferred by the lens are $1/d = |u| = \Theta/\lambda = u$, where $\Theta$ is the scattering angle (equal to twice the Bragg angle $\Theta_B$ for crystalline samples). We assume, as is often the case in practice, that no objective aperture is present to limit resolution. The last term in Eq. (4.2) (the impulse response) is a negative, peaked function, near the optimum (Scherzer) focus setting, so the bright-field image consists of dark peaks in regions of high projected electrostatic potential, superimposed on a constant background (the first term). The defocus $\Delta f$ is taken negative for a underfocus, a lens weakened from Gaussian focus. For a simple one-dimensional phase grating of period $d$, since the scattering angles are all small, $d = \lambda/\Theta = \lambda/2\Theta_B$. The dominant aberrations of the electron lens are described by an aberration function

$$
\chi(u) = \pi \Delta f \lambda u^2 + \frac{\pi C_s \lambda^3 u^4}{2}
$$

(4.3a)

where $\Delta f$ is the defocus (negative for a lens weakened from Gaussian focus) and $C_s$ is the (positive) spherical aberration coefficient. Thus the effect of the lens and any objective aperture can be represented by a ‘transfer function’

$$
A(u) = P(u) \exp(i\chi(u))
$$

(4.3b)

where $P(u)$ describes the objective aperture, being equal to unity within it and zero elsewhere. We may now seek the choice of focus setting $\Delta f_0$ which makes the last term in Eq. (4.2) (the impulse
response of the lens) as narrow as possible for best resolution. This problem was first studied by O. Scherzer [14], who defined focal settings

$$\Delta f_n = \left[ \frac{C_s \lambda (8n + 3)}{2} \right]^{1/2}$$

(4.4)

for which the slope of $\chi(u)$ is small over extended regions, called passbands. The case $n = 0$ has become known as the Scherzer focus and is used for most HREM work — this defines the lowest order passband which extends from $u = 0$. The Scherzer focus is thus

$$\Delta f_0 = -1.2(C_s \lambda)^{1/2}$$

(4.5)

Fig. 2 shows the function $\sin \chi(u)$ (together with some damping functions described later) at the Scherzer focus for several accelerating voltages. This is the transform of the impulse response. Defocus is conveniently measured in Scherzer units, $D = \Delta f / (C_s \lambda)^{1/2}$, which simplifies Eq. (4.4). An objective aperture could then be matched to the outer cutoff of the $n = 0$ passband, subtending a semi-angle

$$\theta_{ap} = 1.5(C_s \lambda)^{1/4}$$

(4.6)

This combination provides the highest resolution possible (the ‘point’ resolution)

$$d_p = 0.66C_s^{1/4} \lambda^{3/4}$$

(4.7)

which is useful if a straightforward interpretation of the image is required, since the impulse response for $n = 0$ remains a fairly simple peaked function. The use of higher order passbands, in conjunction with image processing, is discussed in later sections.

The resolution limiting effects of partial spatial and chromatic coherence are described by a function

$$B(u) = \exp \left( \frac{-\pi^2 \Delta^2 \lambda^2 u^4}{4} \right) \exp(-\pi^2 u^2 g)$$

(4.8)

Fig. 2. Phase contrast transfer functions (CTFs) for a mid-voltage HRTEM (curve 1), and two high-voltage HRTEMs (curves 2 and 3). Operating conditions are listed above the plots; the spatial frequency scale and CTF cross-overs are given in Å units (from [48] with permission).
which multiplies the transfer function. The first term describes the effect of chromatic aberration, while the second describes the limited spatial coherence which results from using an extended source. Here \( q = (C_s \lambda^3 u^3 + \Delta f \lambda u)^2 \). The semiangle subtended at the sample by an incoherently filled Gaussian effective source is \( \Theta_c = \lambda u_0 (\ln 2)^{1/2} \). The quantity \( \Delta = C_s Q \) is the spread in focus values which results from fluctuations in accelerating voltage, lens current and the thermal energy spread of electrons leaving the filament. Thus

\[
Q = \left( \frac{\sigma^2(V_0)}{V_0^2} + \frac{4\sigma^2(I)}{I_0^2} + \frac{\sigma^2(E_0)}{E_0^2} \right)^{1/2} \tag{4.9}
\]

More precisely, \( \sigma^2(V_0) \) and \( \sigma^2(I) \) are the variance in the statistically independent fluctuations of accelerating voltage \( V_0 \) and objective lens current \( I_0 \), respectively. This factor describes the effect of chromatic aberration on the images [45,46]. The full width at half maximum height of the energy distribution of electrons leaving the filament is

\[
\Delta E = 2.345\sigma(E_0) = 2.345[\sigma^2(E_0)]^{1/2} \tag{4.10}
\]

Using these simple expressions we can understand most of the resolution limiting factors affecting modern HREM instruments, and also use them as the basis for understanding many image processing and super-resolution schemes.

We first define the ‘information resolution limit’, due to the first term on the right of Eq. (4.8). This term describes a real exponential damping of spatial frequencies \( u \), more severe than Gaussian, with a width

\[
u_i(\Delta) = \left[ \frac{2}{\pi \lambda \Delta} \right]^{1/2} \tag{4.11}\]

The corresponding resolution limit \( d_i = u_i(\Delta)^{-1} \) (for which fringe contrast has fallen to 13%) results from electronic instabilities in the high voltage and lens current power supplies, and from the spread of energies in electrons leaving the electron source. For the imaging of non-periodic structures it is a fundamental limit, which is present for all focus settings and partial spatial coherence conditions (illumination conditions). (See Section 5 for recent successful efforts to exceed this limit in cases where the sample is known to be periodic). For the state-of-the-art JEOL ARM1250 electron microscope, measurements [47] give \( \Delta = 8 + 2 \) nm, so that, taking \( \lambda = 0.000735 \) nm at 1.25 MeV, we find that negligible information will be transmitted by the ARM1250 lens for spatial frequencies finer than \( d_i = 0.096 \) nm. Several less severe limits have also been used to define \( u_i(\Delta) \). These include one based on comparisons of computed and experimental diffractograms, or on images of thin crystals of known structure with large unit cells. These give slightly different measurements of the information resolution limit. A term may also be introduced to describe mechanical vibration of the sample [48].

Secondly, the ‘point resolution’ \( d_p \) defined in Eq. (4.7) represents the limit of directly interpretable detail in images of weak phase objects, and is the most important quantity for HREM instruments. Detail in the range between \( d_i \) and \( d_p \) may be transferred with incorrect contrast, and may (noise permitting) be extracted by image processing. All of the detail coarser than \( d_p \) (except at the very lowest spatial frequencies shown in Fig. 1) contributes faithfully to the image. If an objective aperture is present which is smaller than that defined by Eq. (4.6), that may also impose its own, more severe resolution limit.
Finally, we consider the second term in Eq. (4.8), which defines a third type of resolution limit, imposed by the partial spatial coherence of the beam. This arises because the electron source is not an ideal point emitter of electrons. For ideal plane-wave illumination (providing perfect coherence), $\Theta_c = u_0 = 0$ and this term is unity, resulting in no loss of contrast (but very little intensity in the image!). As the illumination semi-angle $\Theta_c$ is increased (for an incoherently filled illumination aperture), the transfer of spatial frequencies becomes increasingly severely attenuated in regions where the slope of $\chi(u)$ is large. Conversely, within the passbands defined by Eq. (4.4) where the slope is small, little attenuation occurs as the source is enlarged. This has been the basis of many image processing schemes for improved resolution. When field-emission electron guns are used, it may happen that the illumination aperture is coherently filled. Then the transfer theory outlined above must be modified, and resolution is somewhat improved for two reasons [28,49]. Firstly, the coherently filled illumination aperture provides a smaller effective source width in the revised theory, and secondly, because the reduced energy spread in the beam reduces $d_\lambda$. However, Gang et al. [50] reach different conclusions.

Fig. 2 shows plots of $\sin \chi(u) B(u)$ for several sets of experimental parameters (from [48]). Good contrast is obtained from regions where $\sin \chi(u) \approx -1$, so that the phase shifts introduced by defocus and spherical aberration combine to give $-\pi/2$ over this limited range of spatial frequencies.

Recently, improvements in resolution to the 0.1 nm level have meant that a new aberration has been exposed — three-fold astigmatism. An example of its measurement and removal (by imposing a hexapole field on the objective stigmator) can be found in Wang et al. [51].

For samples too thick for the approximation of Eq. (4.1) to be made, full multiple scattering calculations must be made for the images, and it is no longer possible to use the simple concept of a transfer function or even to define resolution simply, because it then becomes a property of the sample and the microscope, rather than of the microscope alone. Nevertheless, although an intuitive interpretation of the image in terms of sample structure is lost, it is still possible to compare simulated images for model structures with experimental images, and to seek the best-fitting model structure [52]. This topic is treated more fully in Section 8.

Fig. 3. Image of NiAl recorded at a resolution of about 0.1 nm, projected down (1 1 0). [208].
The preceding discussion of resolution assumes that one has no a-priori information about the sample. In some cases one may know a-priori that a single intensity maximum corresponds to a single atomic column, seen in projection. Then the position of the center of that column may be found, by modeling, with an accuracy which exceeds \( d \) by an order of magnitude or more [53].

Fig. 3 gives an example of a state-of-the-art HREM image, demonstrating a point resolution of very close to 1 Å.

5. Super-resolution schemes

A variety of methods have been developed to improve the resolution limit in HREM to include detail finer than the point resolution \( d \). Currently \( d \) is equal to about 0.1 nm. Finer detail is needed, for example, to distinguish unresolved atoms which occur in staggered positions along an atomic column, seen in projection, and to assist in the reconstruction of three-dimensional detail from defects, such as interfaces. The imaging of amorphous materials and glasses, in order to determine the smallest structural unit, also requires sub-angstrom resolution imaging. The contrast of atomic resolution images increases with resolution, since more intensity is then concentrated into a narrower peak.

The state-of-the-art in microscopy generally, for attainment of the highest interpretable resolution by any technique, (including all scanned probe microscopies and the atom probe) appears to be the images recorded on the JEOL ARM1250 instruments, operating at about 1 MeV. Using one of these, Ichinose, for example, has clearly resolved the individual columns of Silicon and Carbon in SiC with 0.109 nm separation [54]. The nitrogen atom columns in a thin sample of GaN have also been resolved.

To improve on this, a variety of super-resolution schemes have been developed. The situation in optics as regards such schemes which provide resolution beyond the diffraction limit has recently been reviewed [55]. In general, noise is the fundamental limit on resolution, not the Rayleigh criterion. In the absence of noise, the complete object spectrum may be reconstructed (beyond the lens cutoff) by analytic continuation, if the object is small, at the expense of field-of-view. Incoherent imaging theory (to which the Rayleigh criterion applies) may, under certain approximations, be applied to \( Z \)-contrast STEM imaging, and has the desirable property that resolution is then independent of sample. It should be noted, however, that the resolution in phase contrast imaging is significantly better for points scattering in anti-phase than for incoherent imaging.

Near-field methods allow resolution to be obtained which is smaller than the wavelength of the radiation, as first pointed out by Synge [56]. An instructive analysis, presumably with reference to the cavity magnetron, was given by Bethe [57]. The principle can be tested experimentally using microwaves. It will be found that microwaves of 1 cm wavelength incident on a metal screen containing a 1 mm hole will generate a static bubble of intensity on the far side of the screen. By scanning this evanescent wavefield past an object and detecting scattered radiation, an image can be obtained whose resolution is limited by the hole size, not the wavelength of the radiation. The uncertainty principle is not violated, since the wavefield is not propagating. Similar principles explain the attainment of 0.2 nm resolution using 1 V electrons in STM. When combined with fluorescence imaging of living cells, near-field optical techniques hold great promise for biology, in view of their non-invasive, chemically specific and high resolution character (See Ultramicroscopy, vol. 61 for a review of the field). Images have been obtained with visible light showing variation of
intensity on a 20 nm scale, however, image interpretation is complicated and, as in STM and AFM, requires a detailed specification of the tip shape.

From Eq. (4.7) we see that, for HREM, resolution may be improved by reducing either the spherical aberration constant or the electron wavelength. The second approach has led to the recent installation of several HREM machines operating at energies above 1 MeV in Japan, and one in Europe. Disadvantages of this approach include high cost and the introduction of radiation damage due to the ballistic 'knock-on' effect. (According to the Bethe law, ionization damage actually decreases (inversely as the square of the electron velocity) with increasing accelerating voltage).

The following methods possess various strengths and weaknesses. For many, the key issues are the alignment of images recorded under slightly different optical conditions in the presence of noise, and the accurate measurement of the electron-optical parameters needed for image analysis. For off-axis holography, the ability to record extremely fine carrier fringes can be very demanding. In the past, the problem of aligning aberration-corrector devices by some systematic procedure has proven insurmountable, however, fast computers and new detectors for the automated analysis of aberration figures are addressing this problem. The seven main approaches to improved resolution in HREM are listed in the subsections given below.

5.1. Through-focus methods

The earliest attempts to improve resolution during the 1960s were based on the collection of a series of images for various values of \( n \) in Eq. (4.4), each covering a different passband of spatial frequencies, but all subject to the resolution limit \( d_i < d_p \). Within the approximation that the imaging is linear in complex amplitude, Eq. (4.2), these images can be Fourier analysed (to give a function \( T(u) \)), and the effects of the transfer function modulations 'removed' by deconvolution, or, if \( \phi_p(x, y) \) in Eq. (4.1) is complex, by multiplication by the conjugate of the wave aberration function \( \exp(-i\chi(u)) \). The results for all the images may then be added together, then inverse transformed. In this way, problems with the zero crossings in the transfer function in this deconvolution procedure are avoided, since these occur at different points for each image. A filter \( r(u) \) should be introduced to control noise [58,59]. The use of an equally spaced focal series has also been proposed [60]. Assuming that the focus and spherical aberration constants are known, one may thus form the sum

\[
\psi(u) = \frac{1}{N} \sum_n T_n(u) \exp(i\chi(\Delta f_n, u)) r_n(u)
\]  

(5.1)

The transform of this function will provide an image whose resolution extends, in principle, to \( d_i \). The main problems which arise in the implementation of this approach include the registration (alignment) of the images, the optimum choice of noise filter, and the correct incorporation of partial coherence effects [61]. An example of such a use of through focus series can be found in [53]. Eq. (5.1) does have the remarkable property that the unwanted conjugate image and the non-linear terms (neglected in Eq. (4.2)) are both reduced by a factor \( N^{1/2} \), since we may consider the non-linear term to vary randomly with focus. This method is a development of in-line holographic reconstruction, an in-line hologram being simply an out-of-focus image. The general principle of this method — ‘coherent detection’ — may be extended to parameters other than defocus in the transfer function, provided they can be accurately measured and that they do influence resolution. Inclination of the beam direction, one such alternative method, is discussed next.
5.2. Tilt series, aperture synthesis

From the earliest days of HREM, it was realized that resolution could be approximately doubled, if the incident beam direction (and the sample) were inclined to the optic axis by, say, the Bragg angle $\Theta_B$, so that the optic axis then bisects the angle $\Theta = 2\Theta_B$ between the incident beam and a first-order Bragg beam [62]. An instrument whose resolution limit $d_i$ was equal to only half this Bragg spacing $d = \lambda/\Theta = \lambda/2\Theta_B$ would then nevertheless be capable of imaging this spacing, since the two beams would fall at opposite sides of the transfer function. The two beams then define the diameter of an achromatic circle, on which all even-order aberrations cancel, since the aberration function and its conjugate are then multiplied together in the final image intensity expression. Obviously only one side of the diffraction pattern contributes to such an image, and so, for a non-periodic sample, it becomes necessary to record at least four images, each with the incident beam tilted into a different direction. This has been the basis of many ‘aberration-free’ imaging schemes based on aperture synthesis. The modern implementation depends on computer control of microscopes for beam tilting and image acquisition, and is described, for example, in [63]. In that work, images of gold particles on amorphous germanium were recorded in five orientations (symmetric and four orthogonal tilts surrounding). Tilt angles of about $(\lambda/C_s)^{1/4}$ are useful. By Fourier analysis and deconvolution of the aberration function for the tilted images, these images could be re-assembled into a composite image in which the resolution had improved from 0.23 to 0.14 nm. To align the images, a modified form of the cross-correlation method was used, and a type of Wiener noise filter applied. The method is claimed to require much less accurate specification of optical parameters than the through-focus method and to achieve higher resolution. It assumes, however, that the transmission function of the sample is independent of tilt (unless the sample is also tilted with the beam), so that very thin samples must be used, and the noise level is poor at high spatial frequencies. Alignment of the images in the presence of noise is difficult. A recent application of the method can be found in [64].

5.3. Aberration correctors

If $C_s$ could be reduced by electron-optical methods, a straightforward improvement in resolution could be obtained according to Eq. (4.7). This has been a goal of electron-optical designers for several decades, and it appears to have been finally reached within the last year. As the aberration coefficients of rotationally symmetric electron lenses are always positive, there was thought to be little hope of canceling out aberrations using round lenses. Scherzer, in 1947 was the first to suggest the use of multipole lenses for this purpose [65]. Pioneering work by Seeliger, Mollensted, Deltrap, Crewe, Rose, Koops and Hely described various schemes for reducing both spherical and chromatic aberration. Over the past 2 years the first experimental results have been published showing useful improvements in resolution for both STEM [66] and TEM [67] instruments. It should be pointed out, however, that earlier workers were frequently able to improve the performance of sufficiently poor instruments by correctors — the real challenge has been to exceed the performance of the very best uncorrected instrument using a corrector. One of the chief difficulties has always been the determination of a systematic procedure for the alignment of the large number of electrostatic and magnetic optical elements. The first successful modern results were obtained recently using an SEM instrument — here resolution was improved from 5 to 1.8 nm using a corrector [68].

For HREM work, a hexapole corrector [69] which corrects for third-order spherical aberration $C_s = C_3$ has been fitted to a 200 kV field-emission TEM [67]. Resolution remains limited by the
Due to electronic instabilities, and by chromatic aberration, which is not corrected ($C_c = 1.3 \text{ mm}$). Improvements in $d_i$ would expose a resolution limit of 0.05 nm, due to fifth-order aberrations, with $C_5 = 4 \text{ mm}$ for this instrument. The corrector uses two hexapoles to introduce negative spherical aberration which just cancels that of the round objective lens. The result is an improvement in point resolution from 0.24 nm to about 0.13 nm.

Fig. 4 shows a corrected image of a thin GaN crystal obtained from this instrument, projected in the [0 1 1] direction. The pairs of closely spaced atomic columns ('dumbells') are clearly resolved. The researchers comment that the improvement in image sharpness and contrast (due to the concentration of energy into a taller, narrower lens impulse response) is more striking than the improvement in resolution. Images are recorded by CCD camera, analysed to produce diffractograms for measurement of aberrations, and the results fed back to the correctors for adjustment of their currents and voltages.

Aberrations correction can be expected to bring new challenges and solutions to the field of HREM. In addition to improved resolution, the size of the region which can be studied by selected area diffraction is usefully reduced, since $C_s$ is reduced. The much wider passbands and oscillation-free transfer function eliminate the damping effects of spatial coherence on the transfer function (Eq. (4.8) with $C_s \approx 0$). Tilting the incident beam by up to 30 mrad was found to introduce negligible image shift, defocus, astigmatism or coma, so that the use of hollow-cone illumination is greatly facilitated, and, for the thinnest samples, a more intense lattice image may be obtained by using a larger illumination angle. A larger pole-piece gap may be possible, allowing greater access for X-ray detectors and other exotic sample holders for in situ work.

The interpretation of phase-contrast images from $C_s$-corrected HREM instruments requires renewed analysis, since the conventional method of balancing-off the spherical aberration phase shift against the defocus phase shift to give the 90° shift required for phase contrast imaging cannot be used. Since the resolution limit $d_i$ is known, one approach may be to re-introduce some spherical aberration, consistent with this limit, in order to obtain phase contrast. Alternatively, we note that the projected charge density approximation (PCD) assumes that $C_s = 0$, and so is ideally suited to this situation [70]. It predicts an image $I(x, y)$ from thin crystals whose contrast reveals directly the projected charge-density $\rho_p(x, y)$ (rather than the electrostatic potential) in the sample, and is
proportional to defocus (for small defocus). The result is

\[ I(x, y) = 1 + \frac{(\lambda A f \sigma)}{2\pi \epsilon_0 \epsilon} \rho_p(x, y) \]  

(5.2)

Note that the images are symmetrical in focus, and depend on the total projected charge density, including the nuclear contribution (unlike X-ray diffraction). Note also that the derivation of Eq. (5.2) [10] does not require the expansion of the exponential as used in Eq. (4.1). Hence the PCD approximation takes better account of multiple scattering than the weak-phase object approximation.

Finally, we note a challenge for all super-resolution schemes — that of aligning thin crystals sufficiently accurately with the beam and the optic axis. ‘Resolution’ rapidly becomes limited by imperfect alignment for crystalline samples thicker than a few nanometers.

In summary, the Heidelberg experimental aberration-corrected TEM, operating at 200 keV, has now almost achieved the level of resolution previously obtainable only from high-voltage machines operating at 1 MeV or above, with obvious advantages in cost and radiation damage reduction. The information resolution limit can be improved to the 0.1 nm level of the HVEM instruments by decreasing the energy spread of electrons from the source to about 0.2 eV. This may require the fitting of an electron monochrometer, similar to those already in use for energy-loss spectroscopy. Commercial development of aberration correctors for both SEM, TEM and STEM instruments seems assured in the near future, and the prospects for an aberration corrected STEM [66] are even more exciting than those for TEM, in view of the benefits of correctors for analysis by ELS and EDX. Two aberration corrected STEMs are under construction in the USA. Similar correctors have also been incorporated recently into photo-emission microscopes.

5.4. Ptychography, Ronchigrams and shadow imaging

All of the methods discussed above improve resolution for general non-periodic objects. The method of ptychography uses a scanning transmission electron microscope (STEM), and the sample must be assumed to be periodic over lateral dimensions comparable with the STEM probe. This may be as small as a few nanometers in width. Ptychography is unique amongst resolution improvement schemes in being the only method capable of improving resolution beyond the information limit \( d_i \). The idea originated in optics, where it was realized many years ago that, if coherent, convergent, radiation (with semi-angle \( \Theta_c \)) is used to illuminate a transmission grating, the resulting far-field diffraction pattern will consist of a set of disks, instead of spots. If \( \Theta_c > \Theta_B \), these discs will overlap and interfere, and the region of interference can be used to determine the phase difference between adjacent Bragg beams. Thus the ‘phase problem’ may be solved, and the complex wavefunction at the grating may be found. The first experiments in electron microscopy were due to Hoppe [71], and an elegant formulation of the theory has been given in a series of papers by Rodenberg, Bates and others (see [72] for references, and for the extension of this method to non-periodic samples). Normally, the resolution of a shadow image is about equal to the size of the (incoherent) illuminating source, however in ptychography it is not limited in this way. In the electron microscope, the corresponding geometry is that of coherent convergent beam diffraction (see [26] for a review), and the usual experimental conditions involve strong multiple scattering. This may prevent such a simple kinematic interpretation. The errors in phase determination due to multiple scattering have been estimated when using this method in electron diffraction [73].

The theory of this effect makes close connection with the theory of Fourier or Talbot self-imaging, which uses the same geometry [11]. Before describing the ptychography experiments, we
show by a simple ray diagram that the intensity distribution in the region of overlapping orders consists of a magnified shadow lattice-image of the grating or crystal. If the illumination angle is not limited, the effects of spherical aberration become dominant, and the pattern is known as a rochigram. Ronchigrams have been recorded from field-emission STEM instruments which clearly show directly the crystal lattice planes [74]. They may be used to measure the aberration constants of STEM instruments [75].

This geometrical-optics interpretation of a coherent CBED pattern with overlapping orders is illustrated in Fig. 5. The asymptotic extension of diffracted CBED cones toward the source defines virtual sources $S'$, necessarily coherent with $S$ (since they are images of the same source). By geometric construction it will be found that these virtual sources lie on one plane in the reciprocal lattice, and so form a virtual point diffraction pattern. The underfocus case (not shown) corresponds to the arrangement used to form ‘Tanaka’ large-angle CBED patterns with the beam selecting aperture removed. Taken alone, two of these sources $S$ and $S'$ will produce Young’s fringes on a distant screen with angular period $\Delta \alpha = \lambda / d_s$, where $d_s$ is the separation of the sources. Now $d_s \approx 2 \Theta_B \Delta f$, so the period of the fringes is

$$\Delta \alpha = \frac{\lambda}{2 \Theta_B \Delta f} \quad (5.3)$$

where $\Delta \alpha$ is the angle subtended at the sample by one fringe on the plane of the CBED detector. The
fringes thus become finer with increasing defocus, as observed. This expression is a good approximation if \( C_s \) is small for low order reflections and large defocus. The fringes may thus be interpreted as a point-projection shadow lattice image, with magnification \( M = L/\Delta f \). We can see this simply by noting that the period of the fringes on the screen is approximately \( X = L/\alpha = M(\lambda/2\Theta_B) = Ma \), with \( a \) the lattice spacing. This is in accordance with the theory of Fourier imaging, and such shadow lattice images have now been observed experimentally by several groups using field-emission STEM instruments [74,76].

The remarkable thing about this geometry is that, although interference is needed only between adjacent orders, this suffices (by an obvious ‘stepping out’ process) to determine the phases of very high order reflections. Thus the resolution of the method is not limited by the size of the electron source (or its image, the probe), which need only be small enough to provide coherence over an angular range as large as twice the Bragg angle. Unlike HREM imaging, the high order reflections at the resolution limit never need interfere with the zero-order beam. Resolution in ptychography is ultimately limited by noise and by the Debye–Waller factor, and excitation error effects, which attenuate high order Bragg beams, rather than by any electron-optical parameters or diffraction limits. Once the phases (and amplitudes) of all the orders have been determined (if single scattering conditions can be assumed), a map of the crystal potential may be found by Fourier synthesis.

If the illumination angle is made much larger than the Bragg angle, so that gross overlap of orders occurs, we have just the geometry proposed by Gabor for in-line holography. The patterns can be interpreted as holograms if the direct, central disc is much more intense than the others, providing a reference beam. Reconstruction must deal with the twin image problem, however, the recent Transport of Intensity method appear to provide a solution to this problem [77].

A lattice image may also be formed by collecting the intensities at the mid-points where each pair of discs overlap and interfere, as the probe is scanned over the sample. This intensity (at one such point) is used to form a STEM lattice image [78]. At the precise mid-point, the intensity has the useful property of being independent of all even-order aberrations, including focus and spherical aberration (but not probe position) [78]. By collecting the intensity at many such mid-points simultaneously, Nellist and co-workers have recently been successful in exceeding the information limit \( d_i^{-1} \). [79] Intensities were collected from a thin silicon crystal using Vacuum Generators HB501 STEM at 100 kV with a point resolution of 0.42 nm. The synthesized image clearly shows the silicon ‘dumbells’ projected along \([0 1 1]\) with a resolution of 0.136 nm. This is finer that the information limit of \( d_i = 0.33 \) nm.

It must be remembered that this ability to beat the information resolution limit is made possible by the assumption of crystallinity — in that respect ptychography has more in common with crystallographic diffraction methods than with imaging, and cannot normally be used to image defects (but see [72]). However, it can be shown that the method remains valid for probe sizes as small as a few unit cells (repeat distances) in width, so that microcrystals can be analyzed.

5.5. Electron holography

While many forms of electron holography exist, the in-line geometry shown in Fig. 5 and the off-axis geometry are the most common. The off-axis mode uses a beam splitter to pass a coherent portion of the beam around the sample, so that it can interfere at the detector with the beam which has passed through the sample. Holography was originally proposed by Gabor [80] in the in-line geometry as a means for overcoming the aberrations of electron microscopes. Very recently, this aim has finally been achieved [81], using, however, the off-axis geometry. The development of optical
holography in the commonly used off-axis geometry came many years after Gabor’s original work, greatly facilitated by the latter invention of the laser.

In-line holography allows lensless imaging, since the use of diverging radiation from a small source produces an image magnified by \( M = L/\Delta f \) (Fig. 5). In practice, the ‘source’ may be defined by a focussed probe (as for STEM), and the resolution is then given approximately by this source size, which may be limited by the aberrations of the probe-forming lens. Gabor’s in-line proposal failed in its original form, because of the twin image problem — it was found impossible to separate the real and virtual images in the reconstruction. Several solutions to this problem have since been proposed — a summary can be found in the chapter by Spence in [82]. The most useful solutions include the Fraunhoffer (small object) method [83], integration over lateral source position, defocus or wavelength (as in photoemission holography), and finally the recent Transport of Intensity method [77]. In-line electron holography has been steadily developed using STEM instruments, first in Chicago [84], and more extensively at Arizona State University [85]. However, to date, no applications of in-line holography have resulted in resolution improvements, except perhaps the methods of Section 5.1 in conventional bright-field HREM. If the source-to-object distance in Fig. 5 is made very large, so that the illumination becomes collimated, and the source to detector distance \( L \) is made small, we have just the arrangement used for conventional HREM, when lenses are added below the sample. Thus the through-focus methods of Section 5.1 may also be considered to be methods for reconstructing in-line electron holograms. A bright-field, out-of-focus HREM image is an electron hologram.

The most extensively developed mode of electron holography is the off-axis geometry (as in optics). The reference wave is obtained using an electrostatic biprism [86], which crosses the wavefield passing through the sample with a set of ‘carrier fringes’, the second term on the right of the first equation below. The reconstruction is then simple, and no twin-image problem is encountered. The intensity on a plane just below the sample is

\[
I(x) = |q(x) \otimes t(x) + \exp(-2\pi i u_0 x)|^2 = 1 + |q(x) \otimes t(x)|^2 + \exp(-2\pi i u_0 x)(q(x) \otimes t(x)) + \exp(+2\pi i u_0 x)(q^*(x) \otimes t(x)^*)
\]  

where \( t(x) \) is the Fourier transform of \( \exp(i\chi(u)) \) (describing propagation of the image over the defocus distance), \( \otimes \) denotes convolution, and \( q(x) \) is the transmission function of the sample. This intensity is Fourier transformed for reconstruction, resulting in distributions centered at \( u = 0 \) and \( u = \pm u_0 \). By isolating one of these satellite spectra (say the last term), multiplying it by \( \exp(-i\chi(u)) \), and inverse transforming, the required transmission function (the unaberrated image) may be recovered. Thus the holographic reconstruction method gives access to the complex wavefunction at the specimen exit face, and so allows correction of aberrations. The normal method of implementing this in TEM uses a biprism placed below the sample, so that the wavefield leaving the sample is magnified before it is crossed by the reference wave. This greatly relaxes the stability requirements needed to record the carrier fringes.

The chief difficulty in off-axis holography arises from the fact that, in order to separate the satellite spectra, the carrier fringe spacing must be about three times finer than the resolution limit required in the reconstruction. Thus, stable 0.03 nm fringes are required in order to reconstruct images at 0.1 nm resolution. (As aberrations may be removed by holography, this 0.1 nm figure may correspond to the information resolution limit rather than the point resolution limit). Mechanical stability (and the provision of sufficient magnification) then become the limiting factors. The Tubingen group [81] have recently obtained reconstructed images of silicon crystals showing 0.104 nm resolution using an electron microscope whose point resolution was 0.198 nm, thus finally
achieving Gabor’s original aim for holography. Reviews of this work can be found in [87]. The applications of electron holography to magnetic materials and superconductors is reviewed in [88].

Many other geometries for lensless electron holographic imaging are possible — a recent survey listed 20 modes [89]. The effects of aberrations enter in different ways in these geometries, allowing new possibilities for their removal.

5.6. Combining images and diffraction patterns: direct inversion

From the earliest days of electron diffraction there has been a continuous effort to solve crystal structures using only electron diffraction data (see [90] for a review of early work). With a few important exceptions, this effort has been largely unsuccessful. At first, it was hoped that a kinematic analysis might serve, and a certain number of layer structures which could be prepared as very thin films were solved using methods modified from X-ray crystallography practice. (The books [91–93] review this work — see also summaries of more recent Russian work (in English) by Avilov, and other work in the two volumes [94], which are devoted to electron diffraction rather than imaging. In the modern revival of interest in electron diffraction, this early Japanese, Russian and Australian work has been largely overlooked). The agreement between a model trial structure and experimental data is conventionally made using the crystallographer’s ‘R factor’, which is similar to the goodness-of-fit term defined in Eq. (9.2) below. Generally, where comparisons could be made [95] it was found that low R factors (e.g. less than 5%) could only be obtained if dynamical effects were included in calculations, since unrealistic sample thicknesses of less than about 4 nm were needed to obtain kinematic conditions for light elements at 100 kV. These results are in the process of being re-discovered by current researchers. The validity domain of the kinematic approximation was investigated in detail for oxides, for example in 1975 [70] in a paper that will repay close study.

By the 1960s it was widely appreciated that the phase problem did not exist in the presence of multiple scattering, since dynamical intensities depend on structure factor phases. (Thus, structure factor phases in non-centrosymmetric crystals, for example, may be measured with high accuracy using quantitative CBED [96], much more accurately than is possibly using many-beam X-ray methods). This raised the hope that structures could be solved by taking advantage of dynamical effects. In fact, this has proven possible in very few cases [97,98] — there are usually too many adjustable parameters (three-dimensional atom positions, absorption parameters and crystal thickness etc.). But the development of the CBED method in the 1960s and 1970s (see [26] for a review), together with the automated refinement of HOLZ line positions, has meant that one can now usually determine both the space group and the Bravais lattice [99] for an unknown crystal, a most important first step for structure analysis which heavily constrains the atom-position determination. This analysis must be performed on a region of crystal thicker than that used for HREM or kinematic diffraction.

An important breakthrough occurred in 1975 [100], when the structure of purple membrane was solved by a combination of electron diffraction and imaging, and this work has served as the model for much work since. For this class of organic two-dimensional crystals, the DNA responsible for its growth ensures that the film is exactly one ‘unit cell’ thick (5.17 nm). The extension of these methods to inorganic samples of varying thickness is therefore not straightforward. (For a comprehensive review of methods for solving organic crystal structures by electron diffraction, including direct methods for solving the phase problem, see [101]). The purple membrane work was important because the samples were unstained, because low-dose methods were used which kept the dose per cell sub-critical (while taking advantage of periodic redundancy to accumulate statistically
significant data), and because of the way in which images (deconvoluted for transfer function effects) were combined with diffraction data to solve the phase problem.

Several groups have now applied this method, and variants of it, to inorganic crystals (see [102] for a review). If HREM images and spot diffraction patterns can be obtained from the same region of crystal, thin enough for kinematic conditions to be assumed, then, using the Fourier transform of Eq. (4.2) (corrected for transfer function effects as described in Section 4), the amplitudes and phases of all the structure factors which lie within the information resolution limit of the microscope may be found. The amplitudes of higher order reflections can be obtained from the spot diffraction patterns, and the direct methods of X-ray crystallography [103] then used to predict (with high probability) the phases of these high orders, based on the known low order amplitudes and phases. If the symmetry of the crystal is known (perhaps as a result of CBED analysis on thicker regions), then these symmetry operations may be applied to the HREM images to improve statistics, and, if regions of constant thickness can be found, the images of many unit cells may be added together for the same purpose. The maximum entropy method has also been used to phase high order reflections.

The differences between X-ray and electron diffraction must be understood before software intended for X-ray crystallography can be used with kinematic electron diffraction data. X-ray rocking curves cover a tiny fraction of the Bragg angle (unlike CBED patterns), so that X-ray data is usually ‘angle-integrated’ across the entire rocking curve. TED spot patterns, however, provide the intensity at one point in the rocking curve. This difference affects the thickness dependence of the intensities. In the two-beam theory at the Bragg condition, the electron diffraction intensity (not angle-integrated) varies sinusoidal with thickness — for small thickness the thickness variation is therefore quadratic in the kinematic limit. The X-ray situation for angle-integrated data is quite different [104]. There is also no analogy for polarization effects in electron diffraction.

Using combinations of diffraction and imaging data some spectacular successes have recently been claimed. For example, atom positions in Ti$_{11}$Se$_4$ have recently been determined to within ±0.002 nm [105]. Using the maximum entropy algorithm and the Sayre equation for phase extension [101], the positions of the oxygen atoms in a mixed potassium–niobium oxide were determined [106]. The HREM images showed only the metal atom columns in projection. A third group have extended the resolution of images of a radiation-sensitive organic film from 0.32 to 0.10 nm by applying the maximum entropy algorithm to diffraction pattern intensities [107]. An impressive application of the use of spot diffraction patterns combined with HREM images has been the solution of the structure of a crystalline magnesium–silicon particle in an aluminium matrix [108]. More recently, this group has also used direct least-squares refinement, combined with the multislice algorithm, to analyse spot patterns, so that multiple scattering effects are included in the analysis [109]. This is similar to the approach used for automated refinement of CBED patterns, however, with large unit cell dimensions the CBED method is difficult to use, since there is little angular variation in intensity within the small Bragg angle. ([110] describes the automated CBED refinement algorithm in detail). Larger unit cells also mean more atom parameters to refine. Typical refinement parameters in such an analysis include Debye–Waller factors, atomic coordinates (three for each atom), scaling factors, thickness, crystal misalignment and absorption parameters. Most optimization methods, when combined with a multiple scattering algorithm, become prohibitively slow when dealing with more than about eight parameters, and the problem of local ‘false’ minima must be faced (A vast literature can be found dealing with this problem of global optimization [111]).

While there are many difficulties to be resolved with these approaches, it is, in the opinion of this reviewer, the most promising approach to super-resolution for crystalline samples. The most important problems include the following: 1. Thickness variation of the sample over the region used to form the periodically-averaged HREM image and diffraction patterns. 2. Obtaining
microdiffraction patterns from the same regions used for HREM analysis. This would be greatly facilitated by the use of an aberration corrected lens, which can give selected area patterns from much smaller regions. 3. The neglect of multiple scattering will lead to errors in atom positions. Similarly, phase factors are introduced by curvature of the Ewald sphere, which is often neglected, but can be included in a kinematic analysis, since it is known. 4. Errors in the a-priori symmetries applied to the HREM images before averaging will result in incorrect structure determinations. 5. Some technical problems include the design of detectors (such as YAG/CCD systems) with sufficient dynamic range to include both the central beam and weak diffracted orders, and the development of cold stages sufficiently stable to allow recording of both HREM images and diffraction patterns at low temperature. Methods for correcting diffraction patterns for the various distortions which occur are given in [102].

It must be remembered that all these statistical approaches, such as direct methods and maximum entropy, give probabilistic results. There may also be uniqueness problems. We can always obtain a result by using these algorithms — the difficulty is to know whether our result is true. $R$ factors are an important but not infallible guide, and it is significant that one recent study of Ce$_5$Cu$_{19}$P$_{12}$ found $R$ factors ranging (for various thicknesses) between 1 and 4% if based on multiple scattering calculations, but between 20 and 60% when using kinematic theory [109]. More generally, $R$ factors are found to fall from about 15% to about 5% when multiple scattering is included. A useful set of curves, showing intensity against thickness for typical oxides, can be found in [70]. Here the validity domains of the kinematic and other approximations are indicated. The validity of the kinematic approximation for HOLZ lines is discussed in [112].

Clearly the trend is toward the incorporation of all the a-priori knowledge we have about a structure, such as its symmetry and Bravais lattice (from CBED analysis) and composition (from EDX and ELS analysis). The density of a microcrystal may be very difficult to estimate, and this information is needed by most X-ray programs. Information on the statistical distribution of interatomic bond-lengths which occur in nature has not been used in the past, and is available. Several schemes for direct inversion of dynamical diffraction pattern intensities to complex structure factors have recently appeared [113–119]. If successful, these methods could provide much higher resolution than imaging, while at the same time, as in X-ray crystallography, they are capable of identifying the species present in the unit cell. However, they are restricted to crystalline samples and cannot be used to analyse the atomic structure of defects. As one example, the bonds between atoms have recently been seen directly by quantitative CBED analysis.

5.7. Atom lenses

The aberrations of electron lenses scale with their size. It follows that a lens of sub-nanometer dimensions can be expected to have extremely small aberration coefficients. The focussing properties of a column of heavy atoms have been calculated, and found to have a focal length of about 2 nm, with a negative spherical aberration coefficient of similar magnitude. Multiple electron scattering calculations confirm that a 0.2 nm diameter STEM probe, focussed over a column of atoms, is re-focussed beyond the column and demagnified to about 0.04 nm diameter [120].

A more practical scheme is suggested in Fig. 6, appropriate to a TEM instrument. Here the Fourier images formed beyond a thin focusing crystal are used as fine probes, which are scanned over the surface of the sample by tilting the incident beam. The focusing crystal consists of widely spaced heavy-atom columns. The Fourier lattice images form narrow waists in the electron stream, which repeat periodically with distance downstream. Calculations show that the array of Fourier ‘probes’ incident on the second crystal (or on a general sample) may be much finer than 0.05 nm for
favorable thicknesses of the focusing crystal. The wavefield exiting the sample is then imaged by a conventional TEM lens, whose resolution need only be sufficient to resolve the individual atomic columns in the focusing crystal. The resulting TEM images are recorded for each incident beam tilt, and this array of images then contains enough information to assemble an image of the sample with 0.05 nm resolution. Multislice calculations show that these schemes work well, in addition to several others related by reciprocity. The Fourier images may be focused onto the sample by variation of electron wavelength, since their period $Z_0$ along the axis is $2d^2/\lambda$, where $d$ is the lateral period of the focusing crystal [10]. Since the function of the TEM lens is simply to integrate over the response of the sample to the incident probe, the final image is independent of the aberrations and instabilities of this lens, provided it can resolve the atomic columns of the focusing crystal. Sample structures might be made to test this method, using focused ion beam (FIB) and ion-implant stop-etch methods. In common with the ‘ALCHEMI’ method, this method relies on the very fine structure of the dynamical wavefield inside a crystal to obtain improved spatial resolution. It is essential that the separation of the heavy atom columns in the focusing xtal be greater than the resolution of the TEM lens. Advantages of the method, in addition to improved resolution, include the small depth of field (about 1 nm at 100 kV), which would allow three-dimensional imaging by optical sectioning, and microanalysis with higher spatial resolution.

Since the wavefield beyond the focussing crystal is periodic along the axis, a sample located distance $nZ_0$ beyond the focuser may be moved back against the focuser without altering the imaging conditions. The arrangement in Fig. 6 thus produces a Moire image, and we see that the intensity in this image, recorded as a function of incident beam tilt, is equivalent to the lateral intensity distribution in a Moire image.

6. Quantitative HREM image analysis — general comments

Several popular computer programs now exist which provide dynamical image simulations of HREM images for given crystal structures, thickness and electron optical parameters ($C_s$, $\Delta f$, $\Delta$, $\Theta_c$) (see, for example [121]). These solve the one-electron Schroedinger equation for the scattering of the beam electron by a thin slab of crystal, represented by its Coulomb electrostatic potential. Other corrections to the potential seen by the beam electron can normally be neglected [122]. X-ray scattering factors for ions or atoms [123,124] are converted to Fourier coefficients of this potential.
using the Mott–Bethe formula (see [26] for examples. The electron structure factors may also be refined using experimental CBED patterns [125] which provide an independent check on ionicity). For HREM image analysis the multi-slice algorithm is commonly used to solve the Schroedinger equation — the relationship of this theory to others (including the Bloch wave method, Feynman’s path integrals and the Howie–Whelan equations) is reviewed in [10,23]. A complete analysis of five-beam imaging in the presence of aberrations and misalignment can be found in [126] — this paper should be studied in detail by beginning microscopists.

Some excellent matches between computed and experimental images have been published over the past two decades [127–129] for crystals of known structure — here images were matched over a range of beam-divergence, focus, and thickness. As a result, the use of tungsten oxide samples became popular as a test sample, since the large unit cells produce many Bragg reflections which finely sample the transfer function, while specimen preparation (by mechanical grinding) is simple. The development of faster computers and the use of CCD detectors [30] has greatly facilitated quantitative image analysis, producing a recent flood of literature on this topic, and exposing many new sources of error (see, for example, [127,128], [129] for good examples of modern work). It has been suggested that cleaved wedges of semiconductors (or MgO smoke cubes viewed down [1 1 0]) form the ideal test sample [10], since then the sample thickness is accurately known at each point in the image if the semiconductor cleavage angle is known. The most troublesome variable is thereby eliminated from the analysis. For the case of MgO smoke cubes viewed along [1 1 0], one then has the following parameters to adjust for best fit:

1. The electron optical parameters (\(C_s, \Delta f, \Delta, \Theta_c\))
2. The ionic state of atoms. ([125] provides measured values for the ions).
3. Absorption parameters (in thicker regions).
4. Alignment of optical axis, beam direction and crystal axis.

An example of an attempt to do this can be found in [130]. As sample preparation is so simple for MgO smoke cubes, this exercise is recommended for all beginning HREM microscopists.

In much of the early attempts to match dynamical calculations to images, the background intensity and overall contrast were adjusted as a free parameters, and a subjective criteria used for best fit. Fig. 7 shows a typical set of images of WO, comparing computed and experimental images in this way. Fits of this quality have now been possible for about 25 years — going beyond this level of agreement using automated quantitative refinement has proven surprisingly difficult, partly for the reasons listed below.

To improve on this it has become common to use least-squares and other automated optimization routines together with the multislice to fit images. If atomic coordinates are included as adjustable parameters, the number of parameters becomes very large, and the problem of local minima is encountered. It has also been discovered that most HREM images show a contrast level about three times less than the computed images. In a recent survey of the causes of systematic error in image matching [131], the influence of the following factors has been evaluated:

1. Errors in parameterized scattering factors, ionicity.
2. The contribution to the image from electrons which excite phonons in the sample. As the resulting energy losses are small (but the scattering angles large), these electrons cannot be excluded from the images by energy filtering.
3. Contributions to the images from other energy loss processes which contribute high resolution detail. (For the small thicknesses used for HREM work, these energy loss processes are weak).
4. Contaminating surface layers of amorphous material or oxide.
5. Fringing fields.
6. Electron beam damage.
7. Errors in measurement of experimental parameters.
8. Stray scattering, such as X-rays generated beyond the film, and secondary electrons.

It was concluded from this study that none of the above factors, acting alone, could account for the typical factor of three difference in contrast between computed and experimental images. The excellent agreement between computed and elastically-filtered experimental CBED patterns [132] suggests that the error does not lie in the diffraction computations or scattering factors, but rather in the imaging process. There is an urgent need for more work on this problem, using a crystal of known structure, which preferably possesses both a large unit cell and predictable cleavage properties so that thickness is known at each point in the image.

The possibility of reconstructing the three-dimensional potential within an inorganic crystal using HREM images recorded in different projections is demonstrated in [133].

7. Species identification in HREM

An outstanding problem in HREM is the identification of atomic species in images. In certain cases this is possible using energy-loss spectroscopy combined with STEM lattice imaging, as discussed in the next section [134,135]. It may also be possible using HREM images, if the matching with calculations can be made sufficiently accurate, however the problem of distinguishing a small change in thickness (due, for example, to an atomically rough surface) from a change in composition...
has never really been solved. Since we always see a projection along atomic columns in HREM images, there seems little hope that this problem will be solved unless we can be assured that samples have atomically smooth surfaces, as in ultra-high vacuum transmission electron microscopy. However, the heat treatments needed to produce such smooth surfaces will usually cause diffusion of impurities to surfaces. (The ALCHEMI technique [136], which uses the effect of channelled electrons on characteristic X-ray emission spectra in TEM is another solution to this problem of species identification with atomic spatial resolution).

One approach to this problem, based on the quantitative analysis of HREM images, is the QUANTITEM method of Ourmazd et al. [137]. In many small-unit cell crystals of high symmetry (such as semiconductors) it is common for rather few strong Bloch waves to be excited (despite the excitation of many beams). This is an indication that the irreducible representation of the group of scattering matrices allows the dispersion matrix to be put into block-diagonal form, with few, small blocks along the diagonal [138–141], see [10] for a worked example of this ‘beam-reduction’ method). Then all beams will oscillate with the same periodicity in thickness, and the appearance of lattice images will also be almost periodic in thickness [142]. Then a single parameter, the extinction distance, will describe the thickness dependence of the images. Certain image features, averaged over one unit cell, may now be extracted, such as the background, the interference between the central beam and first-order beams, and that amongst the diffracted beams, and the amount of these present represented by distance along three orthogonal axes. It will then be found that a vector representing a single unit cell in the image moves on an ellipse with changing thickness. (This result is exact only if there are only two Bloch waves excited). The eccentricity of the ellipse depends on defocus. By comparing the trajectory of this image vector within crystalline material of known composition with that in an unknown region of identically varying thickness, the composition of the unknown region may be extracted. In most cases the samples are very high quality MBE grown

![Fig. 8. Quantitem analysis of Si/GeSi/Si multilayer image, showing composition variation across the interfaces. The height of the figure gives the sample thickness in areas of constant composition. This may be converted into a map of composition variation alone under certain assumptions [137].](image)
semiconductor multilayers, wedge shaped, with the interfaces in the plane of the beam. Then the thickness variation along lines parallel to the edge of the wedge can be assumed constant, and image analysis performed along these lines to give maps of composition. The composition in two layers sandwiching a third may be assumed known. In practice, each pixel in a unit cell defines an axis, and reflections which depend on the difference in scattering factors between two elements present (such as the (0 0 2) in GaAs) may give further sensitivity. Experimentally, the method depends crucially on sample preparation, and the elliptical model can be expected to hold only over a limited range of thickness (less than one two-beam extinction distance) due to defocus changes over the field of view. Dynamical calculations are used to test the elliptical model for each system. The method cannot be expected to be useful for crystals of lower symmetry with a large unit cell, such as the oxide superconductors, where beam-reduction has limited application. A review of applications to semiconductor multilayers, buried interfaces, the Si/SiO₂ interface, and the GeSi/Si system are given in [137]. The resolution of the method is approximately equal to the lateral unit cell dimension, and the method fails if there are large compositional modulations along the beam direction. Fig. 8 shows an application of this method.

8. STEM and Z-contrast

Atoms were first seen directly in electron microscopy using scanning transmission electron microscopes (STEM) [18], rather than TEM instruments. An annular detector was used, which collected most of the high-angle scattering, and the beam energy was 30 kV. Probe widths in modern STEM instruments, which operate between 100 and 300 kV, may now be as small as 0.2 nm or less. Recording times in STEM depend on the number of image pixels required, unlike TEM, and the STEM geometry favors the simultaneous collection of energy loss spectra for analysis at the atomic level [134]. Depending on assumptions made about the phase shifts introduced on scattering, the STEM dark field mode may give higher resolution than the TEM geometry. In practice, spacings of 0.13 nm are routinely observed on STEM instruments with a bright-field point resolution of 0.19 nm. The early STEM atomic-resolution images were interpreted in terms of scattering cross-sections, rather than the theory of phase contrast, and so it was expected (in view of the large scattering angles involved and correspondingly small impact parameters) that the image contrast would depend strongly on atomic number \( Z \) [143], perhaps as \( Z^2 \). This dependence could be enhanced by dividing the high angle elastic scattering by the direct beam. Reciprocity arguments [144] show that STEM with an annular detector is equivalent to HREM imaging using conical-beam illumination, while STEM using a small axial detector is equivalent to conventional HREM. The theory of STEM lattice imaging with a small axial detector then shows that the lattice fringes arise from interference between overlapping orders of the coherent convergent-beam pattern [78]. For equivalent aperturing, these STEM images may therefore be interpreted, via reciprocity, according to the phase-contrast HREM imaging theory given in Section 4, with the illumination aperture (the STEM objective aperture) playing the role of the TEM objective aperture, and the STEM detector replacing the TEM condenser aperture in its effect. All the phase contrast effects normally seen in HREM (Fresnel fringes, contrast reversals with focus, Scherzer focus images etc. are then seen in STEM if a small axial detector is used.

Although the signal is much weaker than for the axial detector case, the use of an annular detector in STEM has important advantages, including elemental discrimination (‘Z-contrast’) and, in most cases, much simpler image interpretation. Since the detection process in STEM is incoherent, the reciprocity-related HREM configuration uses an incoherently filled annular...
illumination aperture. It is therefore useful to make a simplified comparison of ideally incoherent imaging and ideally coherent imaging in TEM, in order to compare the two cases [10]. The incoherent TEM case (annular illumination) will correspond to the STEM case (annular detector). We have seen that the complex image amplitude due to a single incident plane-wave $K_0$ is, in general,

$$\Psi_i(r, K_0) = \int \psi_c(r_0, K_0) \hat{A}(r_1 - r_0) \, dr_0$$  \hspace{1cm} (7.1)$$

where $\hat{A}(r)$ is the Fourier transform of $A(u)$ in Eq. (4.3b), the transfer function of the HREM objective lens, or, by reciprocity, of the STEM probe-forming lens. Then the image intensity for either bright-field HREM or STEM with a small axial detector is given by the convolution of complex amplitudes:

$$I(r) = \left| \int \psi_c(r_0, 0) \hat{A}(r_1 - r_0) \, dr_0 \right|^2$$  \hspace{1cm} (7.2)$$

For the incoherent case (conical illumination TEM or annular detector dark-field STEM), Eq. (7.1) must be integrated over all the illumination angles (for TEM), and it is then the intensity of the image which is convoluted with the intensity of the lens impulse response:

$$I(r) = \int |\psi_c(r_0, 0)|^2 |\hat{A}(r_1 - r_0)|^2 \, dr_0$$  \hspace{1cm} (7.3)$$

Eq. (7.3) provides a very simplified description of ‘Z-contrast’ STEM imaging theory [145,146] and is similar to the expression used to describe incoherent imaging (e.g. by a 35 mm slide projector) in optics. The wave-function squared must be interpreted as the scattering strength of each atom, while the second term on the right is simply the probe intensity distribution. The image intensity still depends on the focus setting, but in a much simpler way, since the transfer function now takes the form of a simple monotonically decaying function, without zero-crossings and hence without contrast reversals with focus. In addition, deconvolution of the impulse response now becomes much simpler using Eq. (7.3) rather than Eq. (7.2), which involves either the phase or twin-image problem. Eq. (7.3) was the basis of the interpretation of the earliest STEM atomic resolution images, with atomic cross-sections used for the scattering strength. It results in an image intensity proportional to the square of atomic number, if high angle scattering dominates the image formation process. Reconciling this interpretation with the phase-contrast theory of dark-field STEM imaging (since atoms are in fact phase objects) is not entirely straightforward, and depends on many simplifying assumptions, such as the shape of the detector [147]. The most important approximations which are needed to derive Eq. (7.3) are:

1. That the scattered radiation (unlike the direct beam) falling within the central hole in the detector can be neglected.
2. Three-dimensional diffraction effects along the beam path are neglected — the sample must be thin. But not too thin, since a true phase object is invisible when imaged incoherently. (Use the unexpanded form of Eq. (4.1) in Eq. (7.3)). If high-order Laue Zone (HOLZ) lines become strong, additional complications can be expected [148].
3. Multiple scattering effects are neglected. For example, if the images are generated (at high angles) by thermal diffuse scattering or Rutherford scattering from the nuclei, then the required integration over thickness will wash out the Pendellosung (thickness-fringe) effects which
influence TEM images. The low angle contribution may not be averaged out in this way. If the high-angle scattering arises from deeply bound Bloch wave states (in an independent Bloch-wave model) it is found that Eq. (7.3) remains a good approximation.

4. Coherence effects along the atomic columns are neglected [149]. This may complicate a simple interpretation in terms of thickness, defocus and atomic number. The effects of strain on these images has also been controversial.

5. STEM Z-contrast imaging is referred to as ‘incoherent’ because of the detection process. The probe formation process is completely coherent, and hence lattice resolution is only possible if the coherence semi-angle across the illumination aperture exceeds the Bragg angle, so that interference occurs between the direct beam and the first-order Bragg beam. This condition ensures that the probe width is roughly equal to the first-order d-spacing [78]. In other words, a small, coherent probe is still needed for Z-contrast, despite the ‘incoherent’ nature of the imaging. Several researchers (e.g. [150,151]) have shown how columns of heavy atoms focus the STEM probe — a form of channeling — if it lies over a column (see Section 5.7). Electron channeling is the tendency of charged particles to run along paths of low potential energy, so that electron beams tend to run along atomic columns in a semi-classical treatment (positive ions run between the atomic columns).

For detailed dynamical simulations testing these approximations, see [150]. The maximum entropy algorithm is commonly used to sharpen the images. In summary, we may say that Z-contrast STEM images are much less sensitive to both instrumental parameters and thickness variation than HREM images.

Using either the Vacuum generators HB6 series of instruments [152], or, more recently, field-emission TEM/STEM machines [153], many studies of the atomic structure of line and planar defects have been reported in the literature (see [145,154] for reviews). These have included determinations of the atomic structure of grain boundaries and interfaces in ceramics, superconductors and semiconductors, and of dislocation core structures in semiconductors (see Section 10 for references). When combined with energy loss spectroscopy the method becomes particularly

Fig. 9. STEM Z-contrast image of a Mn-doped grain boundary in SrTiO$_3$ [277]. The numbers represent position of the probe at which the energy-loss spectra shown in Fig. 10 were recorded.
powerful, as shown in Figs. 9 and 10. Here ELS spectra recorded from individual columns of atoms show differences between spectra recorded in the bulk and at the boundary. An ELNES analysis [155] of these spectra is also possible.

9. Defect modeling and imaging

When the first high resolution images were published in the late 1960s, the hope immediately arose that the three-dimensional atomic structure of point, line and planar defects could be determined, once atomic resolution had been achieved. In general this has not proved to be the case, although a qualitative analysis of defect images may certainly be used to rule out many possible candidate defect structures, and can strongly suggest others. The greatest successes in quantitative analysis have been with planar defects, for which, in a few recent cases, the atomic structure at an interface has been determined in projection. By repeating the analysis for different projections, in principle a three-dimensional model could be built up in favorable cases. Images of dislocation cores have been used as the basis for electronic structure calculations, but the direct measurement of atomic column positions around a core is a formidable challenge. (Under conditions of plane strain, a projection along a straight dislocation line is periodic along the projection direction). There has been almost no useful information obtained on point defect structure by direct HREM imaging — the study of the elastically-filtered diffuse scattering at low temperature may prove more useful for this purpose.

Early work aimed at determining the atomic structure of defects from HREM images used qualitative comparisons of images and calculations (e.g. [156–158] describe perhaps the first determinations by HREM of dislocation core, Schottky barrier semiconductor interfaces and metal-ceramic interfaces respectively). Modern work, which we now review, is based on standard
optimization methods [111]. A figure-of-merit (FOM) is defined which measures the difference between experimental and computed images for a given structural model. The FOM is then minimized with respect to both electron optical and structural parameters (e.g. atom positions) by varying these parameters. In view of the large number of parameters, and of the inevitable presence of systematic errors and errors in ‘given’ parameters, this is a formidable undertaking, which has so far achieved only limited success. The underlying theory, usually based on Baysian statistics, becomes rather complicated, and no agreement has so far been reached on the best FOM for HREM. Unfortunately the structural models which result from the analyses depend on the choice of FOM. Here we review the most straightforward approach and indicate its limitations. Two excellent texts which give more background are [159] (on the method of maximum likelihood) and [160] (on data analysis in crystallography — the HREM analysis closely parallels the Reitvelt method used in Neutron diffraction).

Consider a set of \( N \) independent experimental image intensities \( I^e_i \) (each with error standard deviation \( \sigma_i \)) in the presence of Gaussian, additive, uncorrelated noise. These may be the intensities recorded at each pixel within one projected unit cell of a thin crystal image using a CCD camera, or some average of many cell images believed to occur at the same crystal thickness. If the symmetry of the crystal is known (e.g. from CBED analysis), the images may be further averaged after application of known plane-group symmetry operations. We assume that these experimental data were generated from a certain theoretical model with the addition of noise, so that \( I^e_i = I^t_i + e_i \), where \( e_i \) is the normally distributed noise and \( I^t_i(a_i) \) is the theoretically predicted intensity. This depends on the model parameters \( a_i \). These might include the positions of atomic columns, the sample thickness and electron optical coefficients etc. The likelihood \( P \) of observing, by chance, the particular set of experimental intensities (within a small range) that we did actually observe is then just the product of all the Gaussian distributions, centered around each theoretically predicted intensity with the given noise. The likelihood function is thus given by

\[
P = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(\frac{-s^2}{2}\right)
\]

(9.1)

where

\[
\chi^2 = \frac{s^2}{N-p} = \frac{1}{N-p} \sum_{i=1}^{N} \frac{1}{\sigma^2} [I^e_i - I^t_i(a_i)]^2
\]

(9.2)

Here the number of degrees of freedom is \( N-p \) (there are \( p \) constraints). We then assume that the values of the parameters \( a_i \) which maximize \( P \) are a good estimate of the true values of the parameters, so that the likelihood \( P = P(D|X) \) that the given model \( X \) generated the observed data \( D \) is a maximum if \( \chi^2 \) is a minimum (according to this ‘method of maximum likelihood’), so that \( P \) is a maximum. Scaled variants of \( \chi^2 \) may also be interpreted as a metric (Euclidian distance), which measures the distance between members of two sets — each point in an \( N \)-dimensional set then defines one image. This makes connection with the theory of projection onto complex sets, a powerful tool for image analysis and inversion problems [118]. Computer codes (such as the simplex algorithm, and many others) adjust the model parameters \( a_i \) (e.g. atom positions) until the minimum in \( \chi^2 \) is found. However there may be many local (false) minima, and the problem of finding the true global minimum is fundamental. Some algorithms address this by starting again with a different starting set of parameters — a minimum found to be independent of starting conditions is likely to be global. Others (such as simulated annealing) allow for the possibility of jumping out of a shallow
local minimum to find a deeper one. Bayes’ theorem may then by used to obtain the probability \( P(X|D) \) the posterior inference — that the model is correct, given the data, as

\[
P(X|D) = \frac{P(D|X)P(X)}{P(D)}
\]  

(9.3)

Here the ‘prior’ \( P(X) \) allows for the introduction of a-priori information which limits the probabilities, and \( P(D) \) is a constant. The use of maximum entropy analysis constitutes one possible form of prior \( P(X) \).

Although widely used, the ‘Chi-squared’ goodness of fit index in Eq. (9.2) has several disadvantages. Note that, for \( \chi^2 = 1 \), this equation reduces to the definition of the variance. In particular, a good fit can be obtained (small \( \chi^2 \)) either if the noise is large (large \( \sigma^2 \)) but the model poor, or if the model is good and the noise small. Care is therefore needed in comparing \( \chi^2 \) between different experiments with different noise levels. More importantly, the presence of systematic errors (e.g. wrong value of thickness, absorption, Debye–Waller factors, point defects, bending of the sample, incorrect values of coherence factors, presence of amorphous layers, thermal diffuse scattering etc.) can cause serious problems. In summary, the \( \chi^2 \) optimization analysis works well only under the assumptions of Gaussian, uncorrelated additive noise in the absence of systematic errors. Often, however, the noise distribution is not known, and a different variance may be required for each point. (For the simplest case of Poisson distributed noise, the variance is equal to the mean).

It will always be subject in practice to the problem of avoiding becoming trapped in local minima, and even the fastest computers today can rarely handle minimization with more than half a dozen parameters using HREM data. An evaluation of the method of simulated evolution for finding global minima in HREM refinement can be found in [161]. Independent measurement of thickness, defocus and spherical aberration coefficients is therefore very helpful. Saxton [162] has pointed out that if differences are taken in Eq. (9.2) between the Fourier transforms of the image intensities, then the noise on this data is uncorrelated if the image contrast is not too high.

A procedure is said to be ‘robust’ if it gives parameter estimates close to the minimum for a wide range of error distributions — for HREM the form of this distribution may not be known. The preceding method is not robust in this sense. A method is ‘resistant’ if it is insensitive to ‘outlier’ data points, due, for example, to systematic error. A variety of shaping functions can be introduced into Eq. (9.2) to make it more robust and resistant [160]. Six alternative definitions of goodness-of-fit indices which have been proposed for HREM are summarized in [163]. Different definitions of the Euclidian distance, the cross-correlation function between computed and experimental images, the weighted least squares [164], mean relative distance, and the traditional X-ray \( R \) factor have all been tried. My own view is that the situation will only improve when one can work with uncontaminated samples with atomically smooth surfaces containing defects consisting of columns of atoms perfectly aligned with the beam. The factors listed at the end of Section 6 must also be understood. It is significant that the CBED diffraction calculations, which are performed for much thicker samples, give an excellent fit to experimental diffraction data at the 0.1% level. This work has produced the most accurate qualification of any TEM data, and should be used as a model for HREM qualification (see [26]).

9.1. Line, point and planar defects

The atomic mechanisms involved in ductility have been studied by TEM for many years. For most materials — even silicon — we do not yet know which atomic process limits dislocation
mobility — possibilities include the kink pair formation energy, the kink migration energy or localized defects (or impurity atoms) along the core. Hence for semiconductors at least, it is not clear whether the Hirth–Lothe theory [165] or the obstacle theory of Celli et al. [166] should be used to describe the of ductile deformation. In addition, it is not even clear if kinks collide and annihilate, or become pinned first[167]. Many of these question could in principle be answered by dynamic HREM studies — in practice the problems of radiation-induced glide and the inevitable roughness of the sample surfaces on an atomic scale have prevented this so far. In the future it can be expected that the use of low-dose methods and ultra-high vacuum electron microscopes will overcome both these problems. Images of stationary dislocation cores have been recorded at near-atomic resolution both in the end-on geometry (electron beam parallel to dislocation line) and in the side-view orientation for many years. A review of HREM work on dislocation cores can be found in [168] while reviews of theories and all types of experiments aimed at dislocation core structure determination can be found in [169,170], and in special issues of journals devoted to dislocations, such as J. Phys. III France 7 (1997) and references therein. As it is such a well characterised material, much of this work is devoted to silicon. Single-crystal silicon is the only material for which the dependence of dislocation velocity on applied stress, temperature and doping is reasonably well understood at the atomic level. The effects of the lattice friction due to core structures on materials properties are reviewed for a wide range of materials in [171]. In silicon, the dissociated 60° dislocation can be of two types — glide or shuffle, depending on where the additional half-plane terminates. This question has been studied extensively by HREM ([156,172] and references therein). The first of these papers describes a rule, based on a-priori knowledge of the symmetry of an intrinsic stacking fault, for determining whether atoms appear dark or bright in an HREM image. Evidence that dislocations remain dissociated while in motion is given in [173]. Amongst the first end-on HREM studies was that of [174] for the 30° partial dislocation in silicon. End-on views at 0.38 nm point resolution were compared with calculated images for a range of core models for both the 90 and 30° partials. Electronic structure calculations, based on the pseudo-potential method were then applied to the best-fitting model to calculate the band-structure along the core of the unreconstructed 30° partial. These showed the dispersion of the deep states in the gap created by the dangling bonds in the core. A consensus has subsequently developed (based mainly on ab-inito total-energy calculations) that the core is probably reconstructed via a Peierls-like dimerization of bonds along the core. This clears the gap of deep states [175]. Recently, attempts have been made to image this ‘period-doubling’ along the core of partials in silicon [176]. Similar calculations have been performed for the 90° dislocation in silicon [177,178]. A review of recent electronic structure calculations for dislocations in semiconductors can be found in [169,179]. Recent end-on imaging of dislocation cores in metals is described in [1]. Here the non-planar, three-fold dissociation of [1 1 1]/2 screw dislocation cores in Mo proposed by in 1960 by Hirsch is actually observed directly, thereby explaining the high flow stress of b.c.c. metals at low temperatures. Misfit dislocation cores have also been studied extensively — where hard materials epitax with soft they are found to lie slightly recessed back into the softer material [158], with implications for the fracture toughness of the interface. A Burgers vector analysis of misfit dislocations at Niobium–Sapphire interfaces (formed by MBE) is given in [180] based on HREM, which is shown to provide information unobtainable by conventional TEM Burgers’ vector analysis — this paper contains references to many earlier papers on the same topic. A particularly favorable case is the faulted dislocation dipole, which may be imaged end-on by HREM [181]. Since the two Schockley partial dislocations and two stair rod dislocations which constitute the dipole occur at the intersection of slip planes (containing stacking faults), we have a rare case in which the dislocation lines are constrained to be perfectly straight, and can be aligned with the beam (In general, partials tend to relax in thin TEM foils and rotate toward screw orientation.
in order to minimize their self energy as they approach a surface). Recently, a faulted dipole in GaAs has been imaged at 0.17 nm resolution and the positions of the atomic columns around the cores and stacking faults determined quantitatively [182] can be mapped out on an atomic scale from an HREM image. Lattice images of dislocation cores in oxide superconductors, in Ni$_3$(Al,Ti) (L12 structure, superlattice dislocation showing cross-slip), and in CoTi (B2 structure, screws dissociate on a non-slip plane, increasing yield stress) are discussed in [5] The variety of dislocation cores which can be found in AlN and GaN are imaged and analysed in [183,184]. For the III–V compounds, dislocation mobility has been studied extensively by HREM and TEM — this work is reviewed in [185].

As they cannot easily be seen, a lively controversy has surrounded the discussion of the various kinds of localized defects which can occur along the core of a partial dislocation. These `one-dimensional’ defects include the anti-phase soliton-like defect [186], dislocation kinks [187,188] and jogs. (Bulatov’s paper contains an exhaustive list of the different (e.g. left and right) kink structures which can occur on partials in silicon). Electronic structure calculations for kinks, including estimates of formation and migration energy, are reported in [189,190]. An atomic-resolution image of a jog on a dislocation line in Fe$_3$Al (seen from the side) is discussed in [5]. The first HREM images of dislocation kinks were published by [191], and more recently analysed in [192,176] (see also [193,194]). No direct evidence by HREM for the existence of anti-phase defects (a single dangling bond on an otherwise reconstructed core) exists, however the possibility of observing core reconstruction by HREM observation from the side is outlined in [176]. Dislocation mobility in Si–Ge alloys is analysed in terms of kink motion in [195]. The imaging of jogs in the wurtzite structure is considered in [196], and computed lattice images of jogs in silicon are analysed in [197].

HREM imaging of interfaces has produced the largest literature of all — some recent highlights from various materials systems are listed in Section 10. Here we concentrate on the problem of determining the atomic positions at an interface from HREM images. The interface is assumed to contain the beam direction. A three-dimensional determination would clearly then require a set of images with the sample tilted about an axis normal to the interface — this has been attempted in very few cases [198] (see also [133] for three-dimensional HREM of inorganic crystals). Section 9 outlines many of the principles of quantitative HREM image analysis — for an interface between perfect crystals one may require only the translation vector across the interface [199], a determination of the type of atomic plane at the interface (stacking sequence termination in each crystal) and a measure of any relaxation of the terminating layer [127]. Then the number of parameters in the minimization process is greatly reduced and images may be matched to the known structure of the crystals on either side of the interface, far from it, in order to determine electron-optical parameters and thickness. Misfit dislocations must be well separated — the misfit does however allow the possibility of a unique match within the focus-thickness tableau of computed images, since the Fourier-image periods are slightly different for each crystal [127]. The best sequence in which to fit parameters has been found to be:

(A) Optical parameters based on regions of crystal with known structure.
1. Focus, thickness, Deby–Waller factors.
2. Beam tilt — alignment of beam optic axis.
3. Spatial and temporal coherence factors — Eqs. (4.9) and (4.11). Mechanical vibration.
4. Astigmatism, alignment of crystal axis with beam.

(B) Structural parameters from defect of interest.
1. Translation vector.
2. Relaxation of monolayer.
These parameters are found by minimizing one of the goodness-of-fit quantities discussed in Section 9. The method has been used to determine the structure of Niobium–sapphire interfaces by the Rühle group in a series of papers (see, for example [200]), and from this work a series of rules have been derived for the build-up of the interface. For example, Nb atoms always occupy Al sites, and Nb atoms of the second layer lie as close as possible to the Al lattice sites of a continued lattice of the Al sites in the sapphire. An application of similar image-matching methods to the determination of grain boundary structures in aluminium, using non-linear least-squares optimization, can be found in [128], where the results are compared with atomistic simulations of the structure.

9.2. Defect energies from HREM images

Many properties of solids are controlled by the formation and motion of defects. Examples include interface reactions, dislocation motion, first-order phase transitions, crystal growth and diffusion. The ab-initio total energy method combined with lattice dynamics (quantum molecular dynamics in QMD) has proven powerful for understanding these processes. As one example, we have recently used this method to find the energy barrier to crack tip motion (and subsequent surface reconstruction) in silicon [201]. It was thus possible to predict fracture toughness from first principles in good agreement with experiment. The first step in QMD is to determine the atomic structure of the defect. However, in most cases there are simply too many possible metastable defect structures. Quantitative HREM analysis of the defect can often help, although we have seen that this method is usually restricted to defects consisting of laterally disordered columns of atoms. The combination of these techniques with energy-loss spectroscopy is proving to be even more powerful [134], especially in the study of interfaces [202,203] where embedded atom methods (combined with HREM) are efficient. The ELS spectra, obtained from sub-nanometer regions, can give information on the chemical state of atoms at defects.

In addition to defect structures, it is also important to know the formation energy of defects, and the energy barriers to their motion. For thermally activated processes, these can be determined if atomic resolution movies can be made of the defect motion at known temperature, using an Arrhenius analysis. Thus, [204] (and references therein) give fine examples of the application of this method to semiconductor interface motion. This analysis will give an estimate of the migration energy. The ideal experiment would be one in which it was possible to quench-in the saddle-point structure of the defect, to use HREM images to solve this and then use this as the basis for computation. Again there are often too many possibilities for ab-initio methods to be useful — an extreme example is the difficult case of atomic diffusion. Entropy terms are also difficult to estimate accurately. The concentration of quenched-in defects will often give the formation energy. There are three main problems with the associated electron microscopy: 1. Beam induced effects on defect motion, 2. The effects of surface roughness, and 3. The projection problem.

For the case of kinks on the $30^\circ$ partial dislocation in silicon, we have used ab-initio pseudo-potential methods in the local density approximation to estimate the energy barrier to kink motion [189], and to study the electronic band structure and charge density as the kink moves. This theoretical work was complemented by HREM lattice imaging of kink motion [192], and appears to represent the first direct observation of dislocation kinks. The ‘forbidden’ (4 2 2)/3 reflections generated by a stacking-fault (lying normal to the beam) between partial dislocations were used alone to form the lattice image, as shown in Fig. 11. Dynamical calculations for kink images show that their structure can not be determined from these images [197] because of projection difficulties, but their location can be found to within about 0.3 nm. Video recordings at 600°C showed kink
motion (and pinning) at the edges of the lattice image of the stacking fault. To address the problem of surface roughness on the atomic scale, difference images were formed, from which only details which change could be extracted. The effects of beam induced motion were minimized by turning the beam off during dislocation motion (except during the pinning studies), and using low-dose techniques in conjunction with image plates and CCD recording. From this work the kink unpinning energy could be found, together with the single kink formation energy (0.73 eV) and the migration energy (1.24 eV). The formation energy was obtained by applying the Hirth–Lothe nucleation-and-growth rate-equations to the observed concentration of saddle-point kink pairs under high stress conditions. Thus we find that, unlike metals, kink mobility rather than formation is the rate-limiting step for dislocation motion in the absence of obstacles. For segments longer than the kink mean free path, the sum of these values gives an activation energy for 30$^\circ$C partial dislocation motion of 1.97 eV, in excellent agreement with experiment. Due to the need to use very high beam intensities for the movies (which undoubtedly affects our unpinning energies), we cannot determine where the Hirth–Lothe theory or the obstacle theory of dislocation motion applies to silicon.

10. Applications of HREM in materials science

The greatest successes of HREM in materials science over the past 25 years have generally resulted from its ability to answer two questions in the characterization of microstructure: What micro-phases are present? and How sharp is an interface? Most recently, the method has also been

Fig. 11. Lattice image taken from Si down (1 1 1) using only the inner ‘forbidden’ Bragg reflections generated by a ribbon of stacking fault on (1 1 1). The white line runs along the core of a 90$^\circ$ partial dislocation containing kinks. The 0.33 nm separation of the Peierls valleys is shown. Only within the SF (about 3.6 nm wide) lattice fringes are seen.
used in a few cases to determine the atomic structure of line and planar defects and interfaces, as discussed in Section 9. And in some cases it has been possible to determine the likely atomic mechanism by which microphases transform — this is important in many fields including geology, solid-state chemistry and materials science. The general finding has been that real materials contain a much wider range of microstructures than was previously thought, with important implications for mechanical and electronic properties, and for our understanding of the origins of non-stoichiometry. To cite two examples amongst many: the origin of non-stoichiometry in the complex oxides was thought to arise entirely from point defects. A glance at the first HREM images (without any quantitative image interpretation) showed immediately that planar faults were also involved. Secondly, the importance of intergranular phases in ceramics and at metal-ceramic interfaces has been realized — the presence of even a few unit cells of a third phase at an interface can have dramatic effects on fracture, and this can only be identified by HREM. Similarly, carrier lifetime in semiconductors depends on the abruptness on an atomic scale of interfaces — this can only be determined locally by cross-section HREM. For phase identification, the use of CBED together with HREM should not be overlooked, since an indexed microdiffraction pattern provides a powerful form of microanalysis. To some extent, the digital Fourier transforms of small regions of HREM images can be treated as micro (or nano) diffraction patterns for this purpose, but multiple scattering and recording non-linearities may lead to misleading results.

10.1. Review articles

Most materials scientists will require reviews indexed according to particular materials systems — this is most easily accomplished using library-based web search engines such as ‘INSPEC’ and keyword searches. Two comprehensive reviews of the applications of HREM to materials science have recently appeared in [5,6], and these are probably the best starting point as guides to the recent literature. A briefer review, covering oxide superconductors, colossal magnetoresistance, ionic conduction in the Aurivillius Bi₂O₃ based oxides and Josephson junctions grown on SrTiO₃ can be found in [205]. Earlier reviews include [206,207] and a book, devoted entirely to HREM applications [7] which is also an excellent starting point, covering many fields, including mineralogy and glasses. A review, indexed by type of material (intercalates, ceramics, quasicrystals, semiconductors, organic crystals, mineralogical applications, oxides, surface science, metals, glasses, clusters etc.) for work prior to 1988 can be found in [10]. Particular volumes of Ultramicroscopy have been devoted entirely to certain materials topics — for example vol. 14, No. 1/2 (1984) on interfaces, vol. 8, No. 1/2 (1982) on HREM in solid state chemistry, and vol. 56 to HREM generally. The topic indices for that journal and for J. Micros. both provide an invaluable source of information. J. Micros. vol. 190, No. 1/2 (1998) is entirely devoted to HREM image analysis methods. Applications of one of the highest resolution instruments in existence to GaP, fullerenes, grain-boundaries in copper and NiAl are described in [208] This instrument operates at 1.25 MeV with a point resolution of about 0.1 nm. A web-based catalog of HREM images has been established at http://asma7.iamp.tohoku.ac.jp/EMILIA. A new text devoted to HREM has recently appeared [209]. HREM applications in materials science are reviewed in the book [210]. An index by materials of all systems studied by convergent-beam electron diffraction until 1992 can be found in [26]. This method allows the local measurement of strains in thin crystals and space-group determination for phase identification — it can therefore provide a powerful complimentary technique for HREM studies, often using the same instrument.

Section 10.2 provides a list of references to a small sample of recent work on various materials systems, while Figs. 12–19 show some highlights from this work. This covers applications in
semiconductors, superconductors, ceramics, metals, intermetallics, quasicrystals, Fullerenes, in situ movies and surface imaging. Earlier work can be traced through the references in these papers, in the above reviews, and on the web. The various national and international conferences on electron microscopy are also an invaluable source of information (e.g. the Inst. of Phys. 'EMAG' series in the UK, the MSA conferences in the USA, the EUREM conferences in Europe and the International

Fig. 12. Niobium–sapphire interface (running across the page, half-way up the figure) recorded on the JEOL ARM at LBL, Berkeley at 800 kV by J. Mayer. 'A' shows the unprocessed image. 'B' shows the Fourier filtered image, 'C' the optimized simulation and 'D' the atomic coordinates obtained from the best match [127].

Fig. 13. Piston alloy, showing strengthening precipitates. The matrix (top and bottom) is aluminium, while the central region is silicon. The outer precipitates were identified as Al$_5$Cu$_2$Mg$_8$Si$_3$. (Radmilovic, Dahmen and Mishra, Met and Mat Trans, 1999, in press). Recorded on the same instrument as Fig. 12.
Congress on Electron Microscopy). A series of volumes devoted entirely to TEM of semiconductors can be found, for example, in [211].

10.2. Examples of the application of HREM in materials science

This section contains some very brief comments on recent highlights in atomic-resolution electron microscopy from various areas of materials science.

Fig. 14. Before their significance was fully realised, images of Fullerine onions had appeared in the literature. This image was recorded by S. Iijima in 1980 [249].

Fig. 15. Fullerine onion, consisting of graphitic shells, containing a diamond crystallite in the core. The fringe spacing in the core is the 0.206 nm (1 1 1) spacing of diamond [250].
10.2.1. Semiconductors

The distribution of Indium atoms has been measured in a GaN quantum well by HREM [212]. The conditions for atomic resolution in Si and Ge have been analysed [213]. A comprehensive analysis of imaging conditions for semiconductors has appeared [214]. The observation of dislocation kinks in silicon and their motion is described [192]. Composition variation across Si/SiO₂ and Si/GeSi/Si interfaces are shown in [137]. The structure of a Sigma 5 (1 3 0) tilt boundary in Si

![Fig. 16.](image1)

**Fig. 16.** (a) Two- and (b) seven-sheet tubes of graphitic carbon [4]. Iijima has also obtained images of bucky-tubes’ consisting of a cylindrical layer of one graphite sheet.

10.2.1. Semiconductors

The distribution of Indium atoms has been measured in a GaN quantum well by HREM [212]. The conditions for atomic resolution in Si and Ge have been analysed [213]. A comprehensive analysis of imaging conditions for semiconductors has appeared [214]. The observation of dislocation kinks in silicon and their motion is described [192]. Composition variation across Si/SiO₂ and Si/GeSi/Si interfaces are shown in [137]. The structure of a Sigma 5 (1 3 0) tilt boundary in Si

![Fig. 17.](image2)

**Fig. 17.** Still images from a movie of small gold particles about 2 nm in diameter. The fringe spacing is 0.235 nm. (a), (b) and (c) show single twins. Single cuboctahedral crystals are shown in (e), (f) and (i). (b) and (h) show these particles after transformation to multiply twinned icosahedral particles. They contain about 459 atoms [253].
has been solved in three-dimensions by HREM [198]. The CdTe (0 0 1)/GaAs (0 0 1) interface is analysed by HREM [215]. The NiSi₂/Si (1 1 1) interface is studied and the terminating plane determined [157]. CoSi₂/Si interfaces are studied in detail [216]. The important effects of elastic have been relaxation in thinned multilayers have been analyzed [217]. Dislocation core structure is analysed in Si in [156]. The electronic structure of dislocations is determined, based on HREM images in [174]. Ordering in GaInP alloys has been observed directly by HREM [218]. Dislocation core structures in CdTe/GaAs are analyzed in [152]. The atomic structure of the Cobalt silicide/Si interface is determined in [219] by STEM. Misfit defects are found to form at ripple troughs on InGaAs on GaAs [220].

10.2.2. Superconductors

Oxygen ordering has been detected in YBaCuO [221]. A 90° twin in YBaCuO was matched with image simulations [222]. The atomic structure of planar faults in Sr–Ca superconductor is analysed in [223]. A review of work on electron microscopy of superconductors can be found in [224]. HREM work on interfaces in YBCO/PrBCO/YBCO are detailed in [225]. Grain boundaries in YBCO on MgO are described in [226]. YBCO on Si (0 0 1) is discussed in [227]. Z-contrast STEM HREM images of tilt boundaries in YBaCuO are compared with critical currents in [228].

10.2.3. Ceramics and oxides

SrTiO₃/BaTiO₃ interfaces are analysed by HREM in [229]. A summary of the methods used to determine Nb–sapphire interfaces has appeared [127]. The Nb–sapphire interface is analysed, and the misfit dislocation stand-off distance is detected in [158] (see Fig. 12). A review of work on
metal–ceramic interfaces can be found in [230]. MBE-grown Nb–sapphire interface structures are determined by HREM in [180]. The atomic structure of (1 1 1) grain boundaries in SrTiO$_3$ is discussed in [231]. HREM and ELS are applied to Cu–sapphire interfaces in [202]. The structure of nitrogen platelets in diamond are analysed. [232]. Faulting in beta silicon nitrides observed [233]. Ti–TiB$_2$ interfaces are analysed in [234]. A review of structural defects in YBCO can be found in [235]. Structural modulations in BiSrCaCuO are seen in [236]. The microstructure of CMR

Fig. 19. HREM image along (0 0 1) of in situ oxidation product of Nb$_3$W$_{13}$O$_{47}$, oxidised in an environmental cell under 20 mbar of oxygen. Domains of different structure appear as a result of the oxidation. Microdomains of WO$_3$, tetragonal tungsten bronze and disordered bronze structures can be identified. (Modified JEOL 4000EX) [272].
Manganites is analysed in [237]. The MgO/Cu interface and its electronic states are analysed by STEM ELS in [238].

10.2.4. Metals, intermetallics, quasicrystals

Short range order in Cu–Pd alloys is discussed in [239]. Tetrahedrally close-packed phases in superalloys are comprehensively reviewed in [240]. Ordering and defects in Au–Cd alloys is analysed in [241]. GP zones in Al–Cu alloys are observed in [242]. The core structure of dislocation in Ti are determined in [243]. An authoritative review of ordering in alloys, as studied by HREM, can be found in [244]. Al–Pd–Mn quasicrystal are studied, and a structural model obtained by HREM [245] Facetted Pb inclusions in Al — magic numbers are found to control size [246]. B embrittlement in polycrystalline Ni₃Al can be understood from a STEM and ELS analysis [247]. For a grain boundary in Nb, the structure has been refined by quantitative HREM [248]. Figure 13 shows recent work on an aluminium piston alloy.

10.2.5. Fullerenes, nanotubes

The first HREM image of a bucky-ball was actually obtained in 1980! (see Fig. 14) [249]. Diamond crystals are seen growing inside a bucky-ball, and the transformation analysed (Fig. 15) [250]. The discovery of nanotubes by HREM is shown in Fig. 16 [4].

10.2.6. Movies: dynamic observations at atomic resolution

HREM of an atomically sharp gold tip approaching a surface [251] and HREM observations of operating STM are shown in [252]. Twinning is observed in 2 nm gold particles in [253] (see Fig. 17). In situ observations of a phase transition using high temperature lattice images [254] and the motion of a grain boundary in MgO have been observed at atomic resolution [255] Individual W atoms are seen diffusing on MgO at 218°C [256] ‘Clouds’ of gold atoms are seen forming near surfaces in [257]. Direct observations of Ag atoms diffusing on surfaces are seen in [258]. A UHV FEG TEM has been used to make and image gold nanowires [259]. High temperature HREM imaging has been used to study precipitate interface growth mechanisms [260]. Dislocation kinks are observed in motion in silicon in [192].

10.2.7. Imaging surfaces by TEM at atomic resolution

An HREM plan-view of Si (1 1 1)-7×7 shows sub-surface atoms [261] in Fig. 18. A comprehensive review of UHV HREM work [262], a review of HREM imaging of surfaces, especially ‘profile imaging’ [263], and a review of electron microscopy of surfaces [264] have been given. The Si (1 1 1)-(7 × 7) DAS structure was solved by UHV TED [265].

10.2.8. Amorphous materials, glasses

A review of electron microscopy of amorphous materials has recently appeared [266]. Composition variations in amorphous Si–Ge are studied by HREM in [267]. A new variable coherence method has also been developed to detect high order correlations [268].

10.2.9. Catalysts, controlled atmosphere imaging

Low dose imaging methods for zeolites are described in [269]. The STEM Z-contrast method, which minimizes diffraction effects from a substrate, is described for imaging heavy-metal catalysts [270]. An environmental cell for HREM, providing high temperatures and pressures is described in [271]. In situ observation of the reduction of oxides at atomic resolution is analysed [272] (see Fig. 19).
10.2.10. Magnetic materials

The microstructure of samarium–cobalt magnets, and its effect on coercivity is discussed in [273]. An entire volume of Ultramicroscopy has been recently devoted to the microscopy of magnetic materials [274]. Reference [205] provides an entry into the vast literature on colossal magnetoresistance manganates, of which much more HREM work can be expected.

11. Summary — what has materials science learnt from HREM?

This review has concerned only atomic-resolution electron microscopy, not TEM in general or the many closely related techniques such as EDX, ELS and cathodoluminescence. The contribution of lower-resolution TEM work to materials science has also been immense, in characterizing microstructure, and in understanding the importance of elastic strain fields around inclusions, misfit dislocations and other defects and defect interactions.

In the most general terms, since about 1970, HREM has taught materials scientists that real materials — from minerals to magnetic ceramics and quasicrystals — are far less perfect on the atomic scale than was previously believed. A host of microphases has been discovered by HREM, and the identification of polytypes and microphases has filled a large portion of the HREM literature. At the same time, the direct observation of the atomic structure of defects has led naturally to an interest in the atomic mechanisms by which defects control the properties of solids. The net effect of all these HREM observations has been to give theoreticians confidence in their atomic models for defects. While the resolution of the instrument, and the limitations of the projection approximation, have in the past prevented determination of the atomic structure of defects, the images were used to rule out many possible defect structures, and to suggest others. Increasingly, however, defect structures are now being determined in detail by quantitative methods. HREM has also ‘solved’ materials problems by providing an explanation for bulk properties at the atomic level — for example the recent observation of three-fold dissociation of the core of screw dislocations in Molybdenum [1] explains the high flow stress of b.c.c. metals at low temperatures. As a second example from many, the reduction in critical current in YBa$_2$Cu$_3$O$_{7-x}$ superconductors which occurs at [0 0 1] tilt grain boundaries has recently been explained at the atomic level using Z-contrast STEM [228]. A non-superconducting zone is found at the boundary, whose width increases with misorientation angle. Three other trends can also be discerned — the study of artificially formed structures (especially semiconductor and magnetic multilayers) in order to confirm their perfection at the atomic level has developed rapidly. Secondly, efforts to make and study new structures and processes in the microscope itself are just beginning — for example STM instruments have been observed, during operation, by HREM at lattice image resolution [252], and the atomic bonding process has been observed at atomic resolution, as two gold tips approach and touch each other [251]. The opportunities for experiments in low-temperature mesoscopic physics (in connection with entirely new types of electronic devices, beyond those based on doped silicon) by HREM and electron holography are particularly exciting. Finally, the use of real-time image recording over a range of controlled temperatures has revealed new atomic mechanisms in processes as diverse as sintering, nanoparticle growth and grain-boundary diffusion. (It has been said that the deduction of an atomic mechanism from a few still images is as difficult as trying to determine the rules of a football game from still photographs, rather than from a movie).

The outlook for higher resolution in HREM may be summarized as follows. Ten years ago, a report on possible routes to higher resolution in TEM identified higher voltage machines (with the attendant penalty of radiation damage), off-axis holography (which requires mechanical vibration to
allow recording of carrier fringes about three times finer that the resolution required in the final images), in-line holography (in the form of bright-field HREM through-focus series), and mini-lenses (with very small samples) to be the most promising possibilities. As so often happens in science, possibilities which were then entirely unforeseen now seem more promising — the use of Z-contrast in STEM, the use of aberration correctors in HREM, and possibly the use of ‘atom lenses’, consisting of a crystalline atomic focuser. (For a comparison of these approaches to ultra-high resolution, see [275]). The study of glassy materials and the three-dimensional structure of interfaces are two areas which will benefit from higher resolution.

In summary, the fullest understanding of the structural aspects of atomic mechanisms, kinetics and thermodynamic pathways in solids can now best be obtained by combining theoretical computations, based on the latest ideas in condensed matter theory, with HREM observations and image simulations. The theoretical predictions are limited by the need to guess the saddle-point structure which occurs during atomic motion, since there are usually too many possibilities for even the largest computers to enumerate. Many-electron quantum corrections to interatomic forces also pose fundamental difficulties to the theory. HREM observations can help understand these effects, but the experimental images are needed at the highest recording speed possible, and preferably under conditions of controlled temperature. Radiation-induced effects, surface roughness on an atomic scale and the projection problem present the main difficulties. For these reasons, future HREM instruments are likely to diverge increasingly into those intended for the characterization of real-world, plasma-cleaned engineering materials, and those UHV machines with in situ sample preparation facilities intended for the study of fundamental atomic processes. The later type of machine makes possible the preparation of samples with atomically smooth surfaces, however the heat treatment needed to do this usually anneals out any defects of interest (but see [276]). Aberration correctors and monochromators will become common for both STEM and HREM instruments, as will imaging energy filters for HREM machines. (Note, however, that life-time broadening effects on electron-loss near edge structure (ELNES) contributes between 50–200 mV to linewidths as one progresses from low to high energy inner shell losses, so that monochromation beyond these limits is pointless). The determination of saddle-point structures during atomic motion, and the observation of point defect structures remain the outstanding challenges for the field.

Acknowledgements

This work was supported by NSF award DMR9814055 and the Miller Foundation at UC Berkeley.

Note added in proof


References

[38] Aremco Products, PO Box 429, Ossining, NY 10562-0429, USA.
[56] E. Synge, Phil. Mag. 6 (1928) 356.