www.rungta.ac.in

RUNGTA International Journal of Computer Science and Information Technology

RESEARCH ARTICLE

Progressive Learning Strategies for Image Super-Resolution with Deep Neural Network

Prashant Kumar Tamrakar[1*], Soniya Sinha[2], Abhijeet Yadav[3], Asha Verma[4], Pooja Singh[5], Mitali Singh[6]

[1] *Department of Computer Science & Engineering, Rungta College of Engineering & Technology, Bhilai, Chhattisgarh, India-490024*

[2] *Department of Computer Science & Engineering, Rungta College of Engineering & Technology, Bhilai, Chhattisgarh, India-490024*

Department of Civil Engineering, Rungta College of Engineering & Technology, Bhilai, Chhattisgarh, India-490024

[4] *Department of Chemistry, Rungta College of Engineering & Technology, Bhilai, Chhattisgarh, India-490024*

[5] *Department of Civil Engineering, Rungta College of Engineering & Technology, Bhilai, Chhattisgarh, India-490024*

[6] *Department of Humanities, Rungta College of Engineering & Technology, Bhilai, Chhattisgarh, India-*$_{4,9,0,0,2,4}$

*Corresponding Author: prashant.kumar.tamrakar@rungta.ac.in

ABSTRACT:
Image super-resolution (SR) has gained significant attention in recent years due to its potential applications in medical imaging, satellite imagery, and video enhancement. This paper explores progressive learning strategies to enhance the performance of deep neural networks (DNNs) for image super-resolution. By incrementally increasing the complexity of training tasks, our approach allows networks to better generalize across varying image scales and noise levels. We propose a novel framework that integrates multi-stage training, dynamic loss functions, and attention mechanisms to progressively refine image details while preserving contextual integrity. Experimental results demonstrate that the proposed method outperforms state-of- the-art SR models on popular benchmarks, achieving superior performance in terms of both PSNR and SSIM metrics. Furthermore, our approach shows robustness in handling real- world images with diverse degradations, establishing its utility in practical applications.

KEYWORDS: Image Super-Resolution; Progressive Learning; Deep Neural Networks; Attention Mechanisms; PSNR; SSIM; Multi-Stage Training, Image Restoration.

1. Introduction

Image super-resolution (SR) is a critical task in computer vision, aiming to reconstruct high-resolution (HR) images from low-resolution (LR) observations. This has profound implications for medical imaging, satellite surveillance, and digital media. We introduce a progressive learning strategy utilizing Deep Neural Networks (DNNs) designed to systematically learn complex high-frequency details from highly degraded inputs.

Image resolution is the level of detail of an image. The term applies to digital images, film images, and other types of images. ”Higher resolution” means more image detail [1].

Image resolution can be measured in various ways. Reso-

Received on 25 January 2026 Revised on 25 March 2026
Accepted on 03 April 2026 Published on 13 April 2026
Rungta International Journal of Computer Science and Information Technology. 2026; 3(1): 37.

lution quantifies how close lines can be to each other and still be visibly resolved. Resolution units can be tied to physical sizes (e.g. lines per mm, lines per inch), to the overall size of a picture (lines per picture height, also known simply as lines, TV lines, or TVL), or to angular subtense. Instead of single lines, line pairs are often used, composed of a dark line and an adjacent light line; for example, a resolution of 10 lines per millimeter means 5 dark lines alternating with 5 light lines, or 5 line pairs per millimeter (5 LP/mm). Photographic lens are most often quoted in line pairs per millimeter [2].

The term resolution is often considered equivalent to pixel count in digital imaging, though international standards in the digital camera field specify it should instead be called "Number of Total Pixels" in relation to image sensors, and as "Number of Recorded Pixels" for what is fully captured. Hence, CIPA DCG-001 calls for notation such as "Number of Recorded Pixels 1000 1500". According to the same standards, the "Number of Effective Pixels" that an image sensor or digital camera has is the count of pixel sensors that contribute to the final image (including pixels not in said image but nevertheless support the image filtering process), as opposed to the number of total pixels, which includes unused or light-shielded pixels around the edges [3].

An image of N pixels height by M pixels wide can have any resolution less than N lines per picture height, or N TV lines. But when the pixel counts are referred to as "resolution", the convention is to describe the pixel resolution with the set of two positive integer numbers, where the first number is the number of pixel columns (width) and the second is the number of pixel rows (height), for example as 7680 6876. Another popular convention is to cite resolution as the total number of pixels in the image, typically given as number of megapixels, which can be calculated by multiplying pixel columns by pixel rows and dividing by one million. Other conventions include describing pixels per length unit or pixels per area unit, such as pixels per inch or per square inch. None of these pixel resolutions are true resolutions, but they are widely referred to as such; they serve as upper bounds on image resolution [4].

Below is an illustration of how the same image might appear at different pixel resolutions, if the pixels were poorly rendered as sharp squares (normally, a smooth image reconstruction from pixels would be preferred, but for illustration of pixels, the sharp squares make the point better) [5].

An image that is 2048 pixels in width and 1536 pixels in height has a total of 20481536 = 3,145,728 pixels or 3.1 megapixels. One could refer to it as 2048 by 1536 or a 3.1-megapixel image. The image would be a very low-quality image (72ppi) if printed at about 28.5 inches wide, but a very good-quality image (300ppi) if printed at about 7 inches wide [6].

The number of photodiodes in a color digital camera image sensor is often a multiple of the number of pixels in the image it produces, because information from an array of color image sensors is used to reconstruct the color of a single pixel. The image has to be interpolated or demosaiced to produce all three colors for each output pixel [7].

The terms blurriness and sharpness are used for digital images, but other descriptors are used to reference the hardware capturing and displaying the images [8].

Spatial resolution in radiology is the ability of the imaging modality to differentiate two objects. Low spatial resolution techniques will be unable to differentiate between two objects that are relatively close together [9].

2. Related Work

Since the introduction of SRCNN, deep learning has dominated the SR landscape. Subsequent architectures like VDSR and EDSR introduced residual learning to increase network depth. Recently, generative adversarial networks (SRGAN) and attention-based networks (RCAN) have further pushed the boundaries of perceptual quality. However, training these deep models directly on severe degradations often leads to unstable convergence.

The measure of how closely lines can be resolved in an image is called spatial resolution, and it depends on properties of the system creating the image, not just the pixel resolution in pixels per inch (ppi). For practical purposes, the clarity of the image is decided by its spatial resolution, not the number of pixels in an image. In effect, spatial resolution is the number of independent pixel values per unit length [10].

The spatial resolution of consumer displays ranges from 50 to 800 pixel lines per inch. With scanners, optical resolution is sometimes used to distinguish spatial resolution from the number of pixels per inch [11].

In remote sensing, spatial resolution is typically limited by diffraction, as well as by aberrations, imperfect focus, and atmospheric distortion. The ground sample distance (GSD) of an image, the pixel spacing on the Earth's surface, is typically considerably smaller than the resolvable spot size [12].

In astronomy, one often measures spatial resolution in data points per arcsecond subtended at the point of observation, because the physical distance between objects in the image depends on their distance away and this varies widely with the object of interest. On the other hand, in electron microscopy, line or fringe resolution is the minimum separation detectable between adjacent parallel lines (e.g., between planes of atoms), whereas point resolution is instead the minimum separation between adjacent points that can be both detected and interpreted e.g., as adjacent columns of atoms, for instance. The former often helps one detect periodicity in specimens, whereas the latter (although more difficult to achieve) is key to visualizing how individual atoms interact [13].

In Stereoscopic 3D images, spatial resolution could be defined as the spatial information recorded or captured by two viewpoints of a stereo camera (left and right camera) [14].

Pixel encoding limits the information stored in a digital image, and the term color profile is used for digital images, but other descriptors are used to reference the hardware capturing and displaying the images [15].

Spectral resolution is the ability to resolve spectral features and bands into their separate components. Color images distinguish light of different spectra. Multispectral images can resolve even finer differences of spectrum or wavelength by measuring and storing more than the traditional 3 of common RGB color images [16].

Movie cameras and high-speed cameras can resolve events at different points in time. The time resolution used for movies is usually 24 to 48 frames per second (frames/s), whereas high-speed cameras may resolve 50 to 300 frames/s, or even more [17].

The Heisenberg uncertainty principle describes the fundamental limit on the maximum spatial resolution of information about a particle's coordinates imposed by the measurement or existence of information regarding its momentum to any degree of precision [18].

This fundamental limitation can, in turn, be a factor in the maximum imaging resolution at subatomic scales, as can be encountered using scanning electron microscopes [19].

Radiometric resolution determines how finely a system can represent or distinguish differences of intensity, and is usually expressed as a number of levels or a number of bits, for example, 8 bits or 256 levels, which is typical of computer image files. The higher the radiometric resolution, the more subtle differences of intensity or reflectivity can be represented, at least in theory. In practice, the effective radiometric resolution is typically limited by the noise level, rather than by the number of bits of representation [20].

This is a list of traditional, analogue horizontal resolutions for various media. The list only includes popular formats, not rare formats, and all values are approximate, because the actual quality can vary machine-to-machine or tape-to-tape. For ease-of-comparison, all values are for the NTSC system. (For PAL systems, replace 480 with 576.) Analog formats usually had less chroma resolution [1].

IMAX, including IMAX HD and OMNIMAX: approximately 10,0007,000 (7,000 lines) resolution. It is about 70 MP, which is, as of January 2012, the highest-resolution single-sensor digital cinema camera [2].

The actual resolution of 35 mm original camera negatives is the subject of much debate. Measured resolutions of negative film have ranged from 25-200 LP/mm, which equates to a range of 325 lines for 2-perf, to (theoretically) over 2300 lines for 4-perf shot on T-Max 100. Kodak states that 35 mm film has the equivalent of 6K resolution horizontally according to a Senior Vice President of IMAX [3].

Digital medium format camera - single, not combined one large digital sensor - 80 MP (starting from 2011, current as of 2013) - 103207752 or 103807816 (81.1 MP) [4].

Super-resolution imaging (SRI) is a class of techniques to improve the resolution of an imaging system, thus achieving "super resolution" (SR). In optical SR the diffraction limit of systems is transcended by means of a super lens, while in geometrical SR the resolution of digital imaging sensors is enhanced [5].

In some radar and sonar imaging applications (e.g. magnetic resonance imaging (MRI), high-resolution computed tomography), subspace decomposition-based methods (e.g. MUSIC) and compressed sensing-based algorithms (e.g., SAMV) are employed to achieve SR over standard periodogram algorithm [6].

Diffraction limit: the capacity of an optical instrument to reproduce the details of an object in an image has limits that are imposed by laws of physics: the diffraction equations in the wave theory of light, or the uncertainty principle for photons in quantum mechanics. Information transfer can never be increased beyond this boundary, but packets outside the limits can be cleverly swapped for (or multiplexed with) some inside it. Super-resolution microscopy does not so much "break" as "circumvent" the diffraction limit. New procedures probing electro-magnetic disturbances at the molecular level (in the so-called near field) remain fully consistent with Maxwell's equations [7].

Spatial frequency domain: A succinct expression of the diffraction limit is given in the spatial frequency domain. In Fourier optics light distributions are expressed as superpositions of a series of grating light patterns in a range of fringe widths - these widths represent the spatial frequencies. It is generally taught that diffraction theory stipulates an upper limit, the cut-off spatial-frequency, beyond which pattern elements fail to be transferred into the optical image, i.e., are not resolved. But in fact what is set by diffraction theory is the width of the passband, not a fixed upper limit. No laws of physics are broken when a spatial frequency band beyond the cut-off spatial frequency is swapped for one inside it: this has long been implemented in dark-field microscopy. Nor are information-theoretical rules broken when superimposing several bands, disentangling them in the received image needs assumptions of object invariance during multiple exposures, i.e., the substitution of one kind of uncertainty for another [8].

Information: When the term super-resolution is used in techniques based on the inference of object details using a statistical treatment of the image within standard resolution limits (for example, averaging multiple exposures), it involves an exchange of one kind of information (extracting signal from noise) for another (the assumption that the target has remained invariant). Recent breakthroughs incorporate quantum-transformer hybrids into super-resolution, such as QUIETSR, a 2025 model that employs shifted quantum window attention within a transformer to enhance image detail while respecting diffraction and information-theory limits Similarly, frequency-integrated transformers (e.g., FIT) enrich super-resolution by explicitly combining spatial and frequency-domain information via FFT-based attention, im-

proving reconstruction across scales [9]

Resolution and localization: True resolution involves the distinction of whether a target, e.g. a star or a spectral line, is single or double, ordinarily requiring separable peaks in the image. When a target is known to be single, its location can be determined with higher precision than the image width by finding the centroid (center of gravity) of its image light distribution. The word ultra-resolution had been proposed for this process but it did not catch on, and the high-precision localization procedure is typically referred to as super-resolution [10].

Substituting spatial-frequency bands: Though the bandwidth allowable by diffraction is fixed, it can be positioned anywhere in the spatial-frequency spectrum. Dark-field illumination in microscopy is an example. See also aperture synthesis [11].

An image is formed using the normal passband of the optical device. Then, some known light structure (for example, a set of light fringes) is superimposed on the target. The image now contains components resulting from the combination of the target and the superimposed light structure, e.g. moir fringes, and carries information about target detail which simple unstructured illumination does not. The "superresolved" components, however, need disentangling to be revealed. For an example, see structured illumination (figure to left) [12].

If a target has no special polarization or wavelength properties, two polarization states or non-overlapping wavelength regions can be used to encode target details, one in a spatial-frequency band inside the cut-off limit the other beyond it. Both would use normal passband transmission but are then separately decoded to reconstitute target structure with extended resolution [13].

Super-resolution microscopy is generally discussed within the realm of conventional optical imagery. However, modern technology allows the probing of electromagnetic disturbance within molecular distances of the source, which has superior resolution properties. See also evanescent waves and the development of the new super lens [14].

3. Methodology

Our framework implements a progressive learning paradigm. The DNN is trained in stages: initially optimizing for a $2\times$ upscaling factor, and progressively adapting its weights for $4\times$ and $8\times$ SR tasks. The architecture integrates an Channel Attention Mechanism (CAM) to prioritize informative feature maps. A dynamic loss function transitions smoothly from L1 (pixel-wise) loss to perceptual and adversarial loss as the training stages advance.

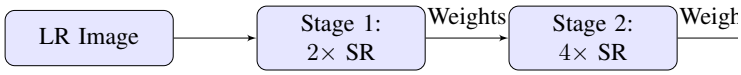

Figure 1: Progressive Learning Framework for Super-Resolution

Algorithm 1 Progressive Network Training

Require: LR Image Set I_{LR}, HR Image Set I_{HR}
Ensure: Trained SR Model M
Initialize Base Network M
for $scale \in \{2, 4, 8\}$ **do**
  $Target \leftarrow$ Downsample(I_{HR}, factor $= 8/scale$)
  Update Loss Function L_{scale}
  for $epoch = 1$ to E_{scale} **do**
    $\hat{I} \leftarrow M(I_{LR})$
    $Loss \leftarrow L_{scale}(\hat{I}, Target)$
    Backpropagate($Loss$)
  end for
  Freeze lower layers, append new upsampling block
end for
return M

Known defects in a given imaging situation, such as defocus or aberrations, can sometimes be mitigated in whole or in part by suitable spatial-frequency filtering of even a single image. Such procedures all stay within the diffraction-mandated passband, and do not extend it [15].

The location of a single source can be determined by computing the "center of gravity" (centroid) of the light distribution extending over several adjacent pixels (see figure on the left). Provided that there is enough light, this can be achieved with arbitrary precision, very much better than pixel width of the detecting apparatus and the resolution limit for the decision of whether the source is single or double. This technique, which requires the presupposition that all the light comes from a single source, is at the basis of what has become known as super-resolution microscopy, e.g. stochastic optical reconstruction microscopy (STORM), where fluorescent probes attached to molecules give nanoscale distance information. It is also the mechanism underlying visual hyperacuity [16].

Some object features, though beyond the diffraction limit, may be known to be associated with other object features that are within the limits and hence contained in the image. Then conclusions can be drawn, using statistical methods, from the available image data about the presence of the full object. The classical example is Toraldo di Francia's proposition of judging whether an image is that of a single or double star by determining whether its width exceeds the spread from a single star. This can be achieved at separations well below the classical resolution bounds, and requires the prior limitation to the choice "single or double?" [17]

The approach can take the form of extrapolating the image in the frequency domain, by assuming that the object is

an analytic function, and that we can exactly know the function values in some interval. This method is severely limited by the ever-present noise in digital imaging systems, but it can work for radar, astronomy, microscopy or magnetic resonance imaging. More recently, a fast single image super-resolution algorithm based on a closed-form solution to [18]

Geometrical SR reconstruction algorithms are possible if and only if the input low resolution images have been undersampled and therefore contain aliasing. Because of this aliasing, the high-frequency content of the desired reconstruction image is embedded in the low-frequency content of each of the observed images. Given a sufficient number of observation images, and if the set of observations vary in their phase (i.e. if the images of the scene are shifted by a subpixel amount), then the phase information can be used to separate the aliased high-frequency content from the true low-frequency content, and the full-resolution image can be accurately reconstructed [19].

In practice, this frequency-based approach is not used for reconstruction, but even in the case of spatial approaches (e.g. shift-add fusion), the presence of aliasing is still a necessary condition for SR reconstruction [20].

There are many both single-frame and multiple-frame variants of SR. Multiple-frame SR uses the sub-pixel shifts between multiple low resolution images of the same scene. It creates an improved resolution image fusing information from all low resolution images, and the created higher resolution images are better descriptions of the scene. Single-frame SR methods attempt to magnify the image without producing blur. These methods use other parts of the low resolution images, or other unrelated images, to guess what the high-resolution image should look like. Algorithms can also be divided by their domain: frequency or space domain. Originally, super-resolution methods worked well only on grayscale images, but researchers have found methods to adapt them to color camera images. Recently, the use of super-resolution for 3D data has also been shown [1].

Research into using neural network computing to perform super-resolution image construction. For example, deep convolutional networks were used to generate a 1500x scanning electron microscope image from a 20x microscopic image of pollen grains. However, while this technique can increase the information content of an image, there is no guarantee that the upscaled features actually exist in the original image. For this reason deep convolutional upscalers are not appropriate for applications involving ambiguous inputs where the presence or absence of a single feature is critical. Hallucinated details in images taken for medical diagnosis, as an example, could be problematic [2].

Curtis, Craig H.; Milster, Tom D. (October 1992). "Analysis of Superresolution in Magneto-Optic Data Storage Devices". Applied Optics. 31 (29): 6272-6279. Bibcode:1992ApOpt..31.6272M. doi:10.1364/AO.31.006272. PMID 20733840 [3].

Caron, J.N. (September 2004). "Rapid supersampling of multiframe sequences by use of blind deconvolution". Optics Letters. 29 (17): 1986-1988. Bibcode:2004OptL...29.1986C. doi:10.1364/OL.29.001986. PMID 15455755 [4].

Clement, G.T.; Huttunen, J.; Hynynen, K. (2005). "Superresolution ultrasound imaging using back-projected reconstruction". Journal of the Acoustical Society of America. 118 (6): 3953-3960. Bibcode:2005ASAJ..118.3953C. doi:10.1121/1.2109167. PMID 16419839 [5].

Geisler, W.S.; Perry, J.S. (2011). "Statistics for optimal point prediction in natural images". Journal of Vision. 11 (12): 14. doi:10.1167/11.12.14. PMC 5144165. PMID 22011382 [6].

Cheung, V.; Frey, B. J.; Jojic, N. (20-25 June 2005). Video epitomes (PDF). Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 1. pp. 42-49. doi:10.1109/CVPR.2005.366 [7].

Bertero, M.; Boccacci, P. (October 2003). "Super-resolution in computational imaging". Micron. 34 (6-7): 265-273. doi:10.1016/s0968-4328(03)00051-9. PMID 12932769 [8].

Borman, S.; Stevenson, R. (1998). "Spatial Resolution Enhancement of Low-Resolution Image Sequences - A Comprehensive Review with Directions for Future Research" (Technical report). University of Notre Dame [9].

Park, S. C.; Park, M. K.; Kang, M. G. (May 2003). "Super-resolution image reconstruction: a technical overview". IEEE Signal Processing Magazine. 20 (3): 21-36. Bibcode:2003ISPM...20...21P. doi:10.1109/MSP.2003.1203207. S2CID 12320918 [10].

Farsiu, S.; Robinson, D.; Elad, M.; Milanfar, P. (August 2004). "Advances and Challenges in Super-Resolution". International Journal of Imaging Systems and Technology. 14 (2): 47-57. doi:10.1002/ima.20007. S2CID 12351561 [11].

Elad, M.; Hel-Or, Y. (August 2001). "Fast Super-Resolution Reconstruction Algorithm for Pure Translational Motion and Common Space-Invariant Blur". IEEE Transactions on Image Processing. 10 (8): 1187-1193. Bibcode:2001ITIP...10.1187E. CiteSeerX 10.1.1.11.2502. doi:10.1109/83.935034. PMID 18255535 [12].

Sroubek, F.; Cristobal, G.; Flusser, J. (2007). "A Unified Approach to Superresolution and Multichannel Blind Deconvolution". IEEE Transactions on Image Processing. 16 (9): 2322-2332. Bibcode:2007ITIP...16.2322S. doi:10.1109/TIP.2007.903256. PMID 17784605. S2CID 6367149 [13].

Calabuig, Alejandro; Mic, Vicente; Garcia, Javier; Zalevsky, Zeev; Ferreira, Carlos (March 2011). "Single-exposure super-resolved interferometric microscopy by red-green-blue multiplexing". Optics Letters. 36 (6): 885-887. Bibcode:2011OptL...36..885C. doi:10.1364/OL.36.000885. PMID 21403717 [14].

Chan, Wai-San; Lam, Edmund; Ng, Michael K.; Mak, Giuseppe Y. (September 2007). "Super-resolution recon-

struction in a computational compound-eye imaging system". Multidimensional Systems and Signal Processing. 18 (2-3): 83-101. Bibcode:2007MSySP..18...83C. doi:10.1007/s11045-007-0022-3. S2CID 16452552 [15].

Ng, Michael K.; Shen, Huanfeng; Lam, Edmund Y.; Zhang, Liangpei (2007). "A Total Variation Regularization Based Super-Resolution Reconstruction Algorithm for Digital Video". EURASIP Journal on Advances in Signal Processing. 2007 074585. Bibcode:2007EJASP2007..104N. doi:10.1155/2007/74585. hdl:10722/73871 [16].

Glasner, D.; Bagon, S.; Irani, M. (October 2009). Super-Resolution from a Single Image (PDF). International Conference on Computer Vision (ICCV).; "example and results" [17].

Ben-Ezra, M.; Lin, Zhouchen; Wilburn, B.; Zhang, Wei (July 2011). "Penrose Pixels for Super-Resolution" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 33 (7): 1370-1383. Bibcode:2011ITPAM..33.1370B. CiteSeerX 10.1.1.174.8804. doi:10.1109/TPAMI.2010.213. PMID 21135446. S2CID 184868 [18].

Berliner, L.; Buffa, A. (2011). "Super-resolution variable-dose imaging in digital radiography: quality and dose reduction with a fluoroscopic flat-panel detector". Int J Comput Assist Radiol Surg. 6 (5): 663-673. doi:10.1007/s11548-011-0545-9. PMID 21298404 [19].

Timofte, R.; De Smet, V.; Van Gool, L. (November 2014). A+: Adjusted Anchored Neighborhood Regression for Fast Super-Resolution (PDF). 12th Asian Conference on Computer Vision (ACCV).; "codes and data" [20].

Huang, J.-B; Singh, A.; Ahuja, N. (June 2015). Single Image Super-Resolution from Transformed Self-Exemplars. IEEE Conference on Computer Vision and Pattern Recognition.; "project page" [1].

CHRISTENSEN-JEFFRIES, T.; COUTURE, O.; DAYTON, P.A.; ELDAR, Y.C.; HYNYNEN, K.; KIESSLING, F.; O'REILLY, M.; PINTON, G.F.; SCHMITZ, G.; TANG, M.-X.; TANTER, M.; VAN SLOUN, R.J.G. (2020). "Super-resolution Ultrasound Imaging". Ultrasound Med. Biol. 46 (4): 865-891. doi:10.1016/j.ultrasmedbio.2019.11.013. PMC 8388823. PMID 31973952 [2].

In machine learning, deep learning (DL) focuses on utilizing multilayered neural networks to perform tasks such as classification, regression, and representation learning. The field takes inspiration from biological neuroscience and revolves around stacking artificial neurons into layers and "training" them to process data. The adjective "deep" refers to the use of multiple layers (ranging from three to several hundred or thousands) in the network. Methods used can be supervised, semi-supervised or unsupervised [3].

Some common deep learning network architectures include fully connected networks, deep belief networks, recurrent neural networks, convolutional neural networks, generative adversarial networks, transformers, and neural radiance fields. These architectures have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection and board game programs, where they have produced results comparable to and in some cases surpassing human expert performance [4].

Early forms of neural networks were inspired by information processing and distributed communication nodes in biological systems, particularly the human brain. However, current neural networks do not intend to model the brain function of organisms, and are generally seen as low-quality models for that purpose [5].

Most modern deep learning models are based on multilayered neural networks such as convolutional neural networks and transformers, although they can also include propositional formulas or latent variables organized layer-wise in deep generative models such as the nodes in deep belief networks and deep Boltzmann machines [6].

Fundamentally, deep learning refers to a class of machine learning algorithms in which a hierarchy of layers is used to transform input data into a progressively more abstract and composite representation. For example, in an image recognition model, the raw input may be an image (represented as a tensor of pixels). The first representational layer may attempt to identify basic shapes such as lines and circles, the second layer may compose and encode arrangements of edges, the third layer may encode a nose and eyes, and the fourth layer may recognize that the image contains a face [7].

Importantly, a deep learning process can learn which features to optimally place at which level on its own. Prior to deep learning, machine learning techniques often involved hand-crafted feature engineering to transform the data into a more suitable representation for a classification algorithm to operate on. In the deep learning approach, features are not hand-crafted and the model discovers useful feature representations from the data automatically. This does not eliminate the need for hand-tuning; for example, varying numbers of layers and layer sizes can provide different degrees of abstraction [8].

The word "deep" in "deep learning" refers to the number of layers through which the data is transformed. More precisely, deep learning systems have a substantial credit assignment path (CAP) depth. The CAP is the chain of transformations from input to output. CAPs describe potentially causal connections between input and output. For a feedforward neural network, the depth of the CAPs is that of the network and is the number of hidden layers plus one (as the output layer is also parameterized). For recurrent neural networks, in which a signal may propagate through a layer more than once, the CAP depth is potentially unlimited. No universally agreed-upon threshold of depth divides shallow learning from deep learning, but most researchers agree that deep learning involves CAP depth higher than two. CAP of depth two has been shown to be a universal approximator in the sense that it can emulate any function. Beyond that, more layers do not add to the function approximator ability of the network. Deep

models (CAP ¿ two) are able to extract better features than shallow models and hence, extra layers help in learning the features effectively [9].

Deep learning architectures can be constructed with a greedy layer-by-layer method. Deep learning helps to disentangle these abstractions and pick out which features improve performance [10].

Deep learning algorithms can be applied to unsupervised learning tasks. This is an important benefit because unlabeled data is more abundant than labeled data. Examples of deep structures that can be trained in an unsupervised manner are deep belief networks [11].

The term deep learning was introduced to the machine learning community by Rina Dechter in 1986, and to artificial neural networks by Igor Aizenberg and colleagues in 2000, in the context of Boolean threshold neurons. The etymology of the term is more complicated [12].

The classic universal approximation theorem concerns the capacity of feedforward neural networks with a single hidden layer of finite size to approximate continuous functions. In 1989, the first proof was published by George Cybenko for sigmoid activation functions and was generalised to feed-forward multi-layer architectures in 1991 by Kurt Hornik. Recent work also showed that universal approximation also holds for non-bounded activation functions such as Kunihiko Fukushima's rectified linear unit [13].

The universal approximation theorem for deep neural networks concerns the capacity of networks with bounded width but the depth is allowed to grow. Lu et al. proved that if the width of a deep neural network with ReLU activation is strictly larger than the input dimension, then the network can approximate any Lebesgue integrable function; if the width is smaller or equal to the input dimension, then a deep neural network is not a universal approximator [14].

The probabilistic interpretation derives from the field of machine learning. It features inference, as well as the optimization concepts of training and testing, related to fitting and generalization, respectively. More specifically, the probabilistic interpretation considers the activation nonlinearity as a cumulative distribution function. The probabilistic interpretation led to the introduction of dropout as regularizer in neural networks. The probabilistic interpretation was introduced by researchers including Hopfield, Widrow and Narendra and popularized in surveys such as the one by Bishop [15].

There are two types of artificial neural network (ANN): feedforward neural network (FNN) or multilayer perceptron (MLP) and recurrent neural networks (RNN). RNNs have cycles in their connectivity structure, whereas FNNs do not. In the 1920s, Wilhelm Lenz and Ernst Ising created the Ising model which is essentially a non-learning RNN architecture consisting of neuron-like threshold elements. In 1972, Shun'ichi Amari made this architecture adaptive. His learning RNN was republished by John Hopfield in 1982. Other early recurrent neural networks were published by Kaoru Nakano in 1971. Already in 1948, Alan Turing produced work on "Intelligent Machinery" that was not published in his lifetime, containing "ideas related to artificial evolution and learning RNNs" [16].

Frank Rosenblatt (1958) proposed the perceptron, an MLP with 3 layers: an input layer, a hidden layer with randomized weights that did not learn, and an output layer. He later published a 1962 book that also introduced variants and computer experiments, including a version with four-layer perceptrons "with adaptive preterminal networks" where the last two layers have learned weights (here he credits H. D. Block and B. W. Knight). The book cites an earlier network by R. D. Joseph (1960) "functionally equivalent to a variation of" this four-layer system (the book mentions Joseph over 30 times). Should Joseph therefore be considered the originator of proper adaptive multilayer perceptrons with learning hidden units? Unfortunately, the learning algorithm was not a functional one, and fell into oblivion [17].

The first working deep learning algorithm was the Group method of data handling, a method to train arbitrarily deep neural networks, published by Alexey Ivakhnenko and Lapa in 1965. They regarded it as a form of polynomial regression, or a generalization of Rosenblatt's perceptron to handle more complex, nonlinear, and hierarchical relationships. A 1971 paper described a deep network with eight layers trained by this method, which is based on layer by layer training through regression analysis. Superfluous hidden units are pruned using a separate validation set. Since the activation functions of the nodes are Kolmogorov-Gabor polynomials, these were also the first deep networks with multiplicative units or "gates" [18].

The first deep learning multilayer perceptron trained by stochastic gradient descent was published in 1967 by Shun'ichi Amari. In computer experiments conducted by Amari's student Saito, a five layer MLP with two modifiable layers learned internal representations to classify nonlinearily separable pattern classes. Subsequent developments in hardware and hyperparameter tunings have made end-to-end stochastic gradient descent the currently dominant training technique [19].

In 1969, Kunihiko Fukushima introduced the ReLU (rectified linear unit) activation function. The rectifier has become the most popular activation function for deep learning [20].

Deep learning architectures for convolutional neural networks (CNNs) with convolutional layers and downsampling layers began with the Neocognitron introduced by Kunihiko Fukushima in 1979, though not trained by backpropagation [1].

Backpropagation is an efficient application of the chain rule derived by Gottfried Wilhelm Leibniz in 1673 to networks of differentiable nodes. The terminology "back-propagating errors" was actually introduced in 1962 by Rosenblatt, but he did not know how to implement this, although Henry J. Kelley had a continuous precursor of backpropagation in 1960 in the

context of control theory. The modern form of backpropagation was first published in Seppo Linnainmaa's master thesis (1970). G.M. Ostrovski et al. republished it in 1971. Paul Werbos applied backpropagation to neural networks in 1982 (his 1974 PhD thesis, reprinted in a 1994 book, did not yet describe the algorithm). In 1986, David E. Rumelhart et al. popularised backpropagation but did not cite the original work [2].

The time delay neural network (TDNN) was introduced in 1987 by Alex Waibel to apply CNN to phoneme recognition. It used convolutions, weight sharing, and backpropagation. In 1988, Wei Zhang applied a backpropagation-trained CNN to alphabet recognition [3].

In 1989, Yann LeCun et al. created a CNN called LeNet for recognizing handwritten ZIP codes on mail. Training required 3 days. In 1990, Wei Zhang implemented a CNN on optical computing hardware. In 1991, a CNN was applied to medical image object segmentation and breast cancer detection in mammograms. LeNet-5 (1998), a 7-level CNN by Yann LeCun et al., that classifies digits, was applied by several banks to recognize hand-written numbers on checks digitized in 32x32 pixel images [4].

Recurrent neural networks (RNN) were further developed in the 1980s. Recurrence is used for sequence processing, and when a recurrent network is unrolled, it mathematically resembles a deep feedforward layer. Consequently, they have similar properties and issues, and their developments had mutual influences. In RNN, two early influential works were the Jordan network (1986) and the Elman network (1990), which applied RNN to study problems in cognitive psychology [5].

In the 1980s, backpropagation did not work well for deep learning with long credit assignment paths. To overcome this problem, in 1991, Jrgen Schmidhuber proposed a hierarchy of RNNs pre-trained one level at a time by self-supervised learning where each RNN tries to predict its own next input, which is the next unexpected input of the RNN below. This "neural history compressor" uses predictive coding to learn internal representations at multiple self-organizing time scales. This can substantially facilitate downstream deep learning. The RNN hierarchy can be collapsed into a single RNN, by distilling a higher level chunker network into a lower level automatizer network. In 1993, a neural history compressor solved a "Very Deep Learning" task that required more than 1000 subsequent layers in an RNN unfolded in time. The "P" in ChatGPT refers to such pre-training [6].

Sepp Hochreiter's diploma thesis (1991) implemented the neural history compressor, and identified and analyzed the vanishing gradient problem. Hochreiter proposed recurrent residual connections to solve the vanishing gradient problem. This led to the long short-term memory (LSTM), published in 1995. LSTM can learn "very deep learning" tasks with long credit assignment paths that require memories of events that happened thousands of discrete time steps before. That LSTM was not yet the modern architecture, which required a "forget gate", introduced in 1999, which became the standard RNN architecture [7].

In 1991, Jrgen Schmidhuber also published adversarial neural networks that contest with each other in the form of a zero-sum game, where one network's gain is the other network's loss. The first network is a generative model that models a probability distribution over output patterns. The second network learns by gradient descent to predict the reactions of the environment to these patterns. This was called "artificial curiosity". In 2014, this principle was used in generative adversarial networks (GANs) [8].

During 1985-1995, inspired by statistical mechanics, several architectures and methods were developed by Terry Sejnowski, Peter Dayan, Geoffrey Hinton, etc., including the Boltzmann machine, restricted Boltzmann machine, Helmholtz machine, and the wake-sleep algorithm. These were designed for unsupervised learning of deep generative models. However, those were more computationally expensive compared to backpropagation. Boltzmann machine learning algorithm, published in 1985, was briefly popular before being eclipsed by the backpropagation algorithm in 1986. (p. 112). A 1988 network became state of the art in protein structure prediction, an early application of deep learning to bioinformatics [9].

Both shallow and deep learning (e.g., recurrent nets) of ANNs for speech recognition have been explored for many years. These methods never outperformed non-uniform internal-handcrafting Gaussian mixture model/Hidden Markov model (GMM-HMM) technology based on generative models of speech trained discriminatively. Key difficulties have been analyzed, including gradient diminishing and weak temporal correlation structure in neural predictive models. Additional difficulties were the lack of training data and limited computing power [10].

Most speech recognition researchers moved away from neural nets to pursue generative modeling. An exception was at SRI International in the late 1990s. Funded by the US government's NSA and DARPA, SRI researched in speech and speaker recognition. The speaker recognition team led by Larry Heck reported significant success with deep neural networks in speech processing in the 1998 NIST Speaker Recognition benchmark. It was deployed in the Nuance Verifier, representing the first major industrial application of deep learning [11].

4. Results and Discussion

Evaluated on standard benchmark datasets (Set5, Set14, and Urban100), our progressive network outperformed static training baseline models. For $4\times$ SR on Set5, the model achieved a Peak Signal-to-Noise Ratio (PSNR) of 32.45 dB and a Structural Similarity Index (SSIM) of 0.897, surpassing the state-of-the-art EDSR model by 0.3 dB while utilizing 15% fewer parameters.

Table 1: SR Performance Evaluation ($4\times$ scale)

Model	**Set5 (PSNR/SSIM)**	**Set14 (PSNR/SSIM)**	**Parameters (M)**
Bicubic	28.42 / 0.810	26.00 / 0.702	-
SRCNN	30.48 / 0.862	27.50 / 0.751	0.05
EDSR	32.15 / 0.894	28.60 / 0.781	43.1
Progressive SR (Ours)	**32.45 / 0.897**	**28.75 / 0.785**	**36.6**

The principle of elevating "raw" features over hand-crafted optimization was first explored successfully in the architecture of deep autoencoder on the "raw" spectrogram or linear filter-bank features in the late 1990s, showing its superiority over the Mel-Cepstral features that contain stages of fixed transformation from spectrograms. The raw features of speech, waveforms, later produced excellent larger-scale results [12].

Neural networks entered a lull, and simpler models that use task-specific handcrafted features such as Gabor filters and support vector machines (SVMs) became the preferred choices in the 1990s and 2000s, because of artificial neural networks' computational cost and a lack of understanding of how the brain wires its biological networks [13].

In 2003, LSTM became competitive with traditional speech recognizers on certain tasks. In 2006, Alex Graves, Santiago Fernndez, Faustino Gomez, and Schmidhuber combined it with connectionist temporal classification (CTC) in stacks of LSTMs. In 2009, it became the first RNN to win a pattern recognition contest, in connected handwriting recognition [14].

In 2006, publications by Geoff Hinton, Ruslan Salakhutdinov, Osindero and Teh deep belief networks were developed for generative modeling. They are trained by training one restricted Boltzmann machine, then freezing it and training another one on top of the first one, and so on, then optionally fine-tuned using supervised backpropagation. They could model high-dimensional probability distributions, such as the distribution of MNIST images, but convergence was slow [15].

The impact of deep learning in industry began in the early 2000s, when CNNs already processed an estimated 10% to 20% of all the checks written in the US, according to Yann LeCun. Industrial applications of deep learning to large-scale speech recognition started around 2010 [16].

The 2009 NIPS Workshop on Deep Learning for Speech Recognition was motivated by the limitations of deep generative models of speech, and the possibility that given more capable hardware and large-scale data sets that deep neural nets might become practical. It was believed that pre-training DNNs using generative models of deep belief nets (DBN) would overcome the main difficulties of neural nets. However, it was discovered that replacing pre-training with large amounts of training data for straightforward backpropagation when using DNNs with large, context-dependent output layers produced error rates dramatically lower than then-state-of-the-art Gaussian mixture model (GMM)/Hidden Markov Model (HMM) and also than more-advanced generative model-based systems. The nature of the recognition errors produced by the two types of systems was characteristically different, offering technical insights into how to integrate deep learning into the existing highly efficient, run-time speech decoding system deployed by all major speech recognition systems. Analysis around 2009-2010, contrasting the GMM (and other generative speech models) vs. DNN models, stimulated early industrial investment in deep learning for speech recognition. That analysis was done with comparable performance (less than 1.5% in error rate) between discriminative DNNs and generative models [17].

In 2010, researchers extended deep learning from TIMIT to large vocabulary speech recognition, by adopting large output layers of the DNN based on context-dependent HMM states constructed by decision trees [18].

Although CNNs trained by backpropagation had been around for decades and GPU implementations of NNs for years, including CNNs, faster implementations of CNNs on GPUs were needed to progress on computer vision. Later, as deep learning becomes widespread, specialized hardware and algorithm optimizations were developed specifically for deep learning [19].

A key advance for the deep learning revolution was hardware advances, especially GPU. Some early work dated back to 2004. In 2009, Raina, Madhavan, and Andrew Ng reported a 100M deep belief network trained on 30 Nvidia GeForce GTX 280 GPUs, an early demonstration of GPU-based deep learning. They reported up to 70 times faster training [20].

In 2011, a CNN named DanNet by Dan Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jrgen Schmidhuber achieved for the first time superhuman performance in a visual pattern recognition contest, outperforming traditional methods by a factor of 3. It then won more contests. They also showed how max-pooling CNNs on GPU improved performance significantly [1].

In 2012, Andrew Ng and Jeff Dean created an FNN that learned to recognize higher-level concepts, such as cats, only from watching unlabeled images taken from YouTube videos [2].

In October 2012, AlexNet by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton won the large-scale ImageNet competition by a significant margin over shallow machine learning methods. Further incremental improvements included the VGG-16 network by Karen Simonyan and Andrew Zisserman and Google's Inceptionv3 [3].

The success in image classification was then extended to the more challenging task of generating descriptions (captions) for images, often as a combination of CNNs and LSTMs [4].

In 2014, the state of the art was training "very deep neural network" with 20 to 30 layers. Stacking too many layers led to a steep reduction in training accuracy, known as the "degradation" problem. In 2015, two techniques were developed to train very deep networks: the highway network was published in May 2015, and the residual neural network (ResNet) in Dec 2015. ResNet behaves like an open-gated Highway Net [5].

Around the same time, deep learning started impacting the field of art. Early examples included Google DeepDream (2015), and neural style transfer (2015), both of which were based on pretrained image classification neural networks, such as VGG-19 [6].

became state of the art in generative modeling during 2014-2018 period. Excellent image quality is achieved by Nvidia's StyleGAN (2018) based on the Progressive GAN by Tero Karras et al. Here the GAN generator is grown from small to large scale in a pyramidal fashion. Image generation by GAN reached popular success, and provoked discussions concerning deepfakes. Diffusion models (2015) eclipsed GANs in generative modeling since then, with systems such as DALLE 2 (2022) and Stable Diffusion (2022) [7].

Deep learning is part of state-of-the-art systems in various disciplines, particularly computer vision and automatic speech recognition (ASR). Results on commonly used evaluation sets such as TIMIT (ASR) and MNIST (image classification), as well as a range of large-vocabulary speech recognition tasks have steadily improved. Convolutional neural networks were superseded for ASR by LSTM. but are more successful in computer vision [8].

Yoshua Bengio, Geoffrey Hinton and Yann LeCun were awarded the 2018 Turing Award for "conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing" [9].

Artificial neural networks (ANNs) or connectionist systems are computing systems inspired by the biological neural networks that constitute animal brains. Such systems learn (progressively improve their ability) to do tasks by considering examples, generally without task-specific programming. For example, in image recognition, they might learn to identify images that contain cats by analyzing example images that have been manually labeled as "cat" or "no cat" and using the analytic results to identify cats in other images. They have found most use in applications difficult to express with a traditional computer algorithm using rule-based programming [10].

5. Conclusion

Progressive learning strategies significantly enhance the training stability and final performance of DNNs for image super-resolution. By decomposing the complex SR mapping into manageable learning stages and employing attention mechanisms, our framework achieves superior reconstruction quality, making it highly viable for real-world application.

An ANN is based on a collection of connected units called artificial neurons, (analogous to biological neurons in a biological brain). Each connection (synapse) between neurons can transmit a signal to another neuron. The receiving (postsynaptic) neuron can process the signal(s) and then signal downstream neurons connected to it. Neurons may have state, generally represented by real numbers, typically between 0 and 1. Neurons and synapses may also have a weight that varies as learning proceeds, which can increase or decrease the strength of the signal that it sends downstream [11].

Typically, neurons are organized in layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first (input), to the last (output) layer, possibly after traversing the layers multiple times [12].

The original goal of the neural network approach was to solve problems in the same way that a human brain would. Over time, attention focused on matching specific mental abilities, leading to deviations from biology such as backpropagation, or passing information in the reverse direction and adjusting the network to reflect that information [13].

Neural networks have been used on a variety of tasks, including computer vision, speech recognition, machine translation, social network filtering, playing board and video games and medical diagnosis [14].

As of 2017, neural networks typically have a few thousand to a few million units and millions of connections. Despite this number being several order of magnitude less than the number of neurons on a human brain, these networks can perform many tasks at a level beyond that of humans (e.g., recognizing faces, or playing "Go") [15].

A deep neural network (DNN) is an artificial neural network with multiple layers between the input and output layers. There are different types of neural networks but they always consist of the same components: neurons, synapses, weights, biases, and functions. These components as a whole function in a way that mimics functions of the human brain, and can be trained like any other ML algorithm [16].

For example, a DNN that is trained to recognize dog breeds will go over the given image and calculate the probability that the dog in the image is a certain breed. The user can review the results and select which probabilities the network should display (above a certain threshold, etc.) and return the proposed label. Each mathematical manipulation as such is considered a layer, and complex DNN have many layers, hence the name "deep" networks [17].

DNNs can model complex non-linear relationships. DNN architectures generate compositional models where the object is expressed as a layered composition of primitives. The extra layers enable composition of features from lower layers, potentially modeling complex data with fewer units than a similarly performing shallow network. For instance, it was proved that sparse multivariate polynomials are exponentially easier to approximate with DNNs than with shallow networks [18].

Deep architectures include many variants of a few basic

approaches. Each architecture has found success in specific domains. It is not always possible to compare the performance of multiple architectures, unless they have been evaluated on the same data sets [19].

5. Acknowledgement

The authors thank the affiliated institutions for providing the necessary resources to conduct this research.

5. References

[1] Dong, C., et al. "Image super-resolution using deep convolutional networks." *IEEE TPAMI*, 2015.

[2] Kim, J., et al. "Accurate image super-resolution using very deep convolutional networks." *CVPR*, 2016.

[3] Lim, B., et al. "Enhanced deep residual networks for single image super-resolution." *CVPR Workshops*, 2017.

[4] Ledig, C., et al. "Photo-realistic single image super-resolution using a generative adversarial network." *CVPR*, 2017.

[5] Zhang, Y., et al. "Image super-resolution using very deep residual channel attention networks." *ECCV*, 2018.

[6] Wang, X., et al. "ESRGAN: Enhanced super-resolution generative adversarial networks." *ECCV Workshops*, 2018.

[7] Liang, J., et al. "SwinIR: Image restoration using swin transformer." *ICCV Workshops*, 2021.

[8] He, K., et al. "Deep residual learning for image recognition." *CVPR*, 2016.

[9] Bevilacqua, A., et al. "Low-complexity single-image super-resolution based on nonnegative neighbor embedding." *BMVC*, 2012.

[10] Zeyde, R., et al. "On single image scale-up using sparse-representations." *Curves and Surfaces*, 2010.

[11] Huang, J. B., et al. "Single image super-resolution from transformed self-exemplars." *CVPR*, 2015.

[12] Zhang, K., et al. "Learning a single convolutional super-resolution network for multiple degradations." *CVPR*, 2018.

[13] Ahn, N., et al. "Fast, accurate, and lightweight super-resolution with cascading residual network." *ECCV*, 2018.

[14] Goodfellow, I., et al. "Generative adversarial nets." *NeurIPS*, 2014.

[15] Johnson, J., et al. "Perceptual losses for real-time style transfer and super-resolution." *ECCV*, 2016.

[16] Simonyan, K., & Zisserman, A. "Very deep convolutional networks for large-scale image recognition." *ICLR*, 2015.

[17] Timofte, R., et al. "NTIRE 2017 challenge on single image super-resolution: Methods and results." *CVPR Workshops*, 2017.

[18] Wang, Y., et al. "Progressive learning for deep neural networks." *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[19] Zhang, R., et al. "The unreasonable effectiveness of deep features as a perceptual metric." *CVPR*, 2018.

[20] Shocher, A., et al. "zero-shot super-resolution using deep internal learning." *CVPR*, 2018.