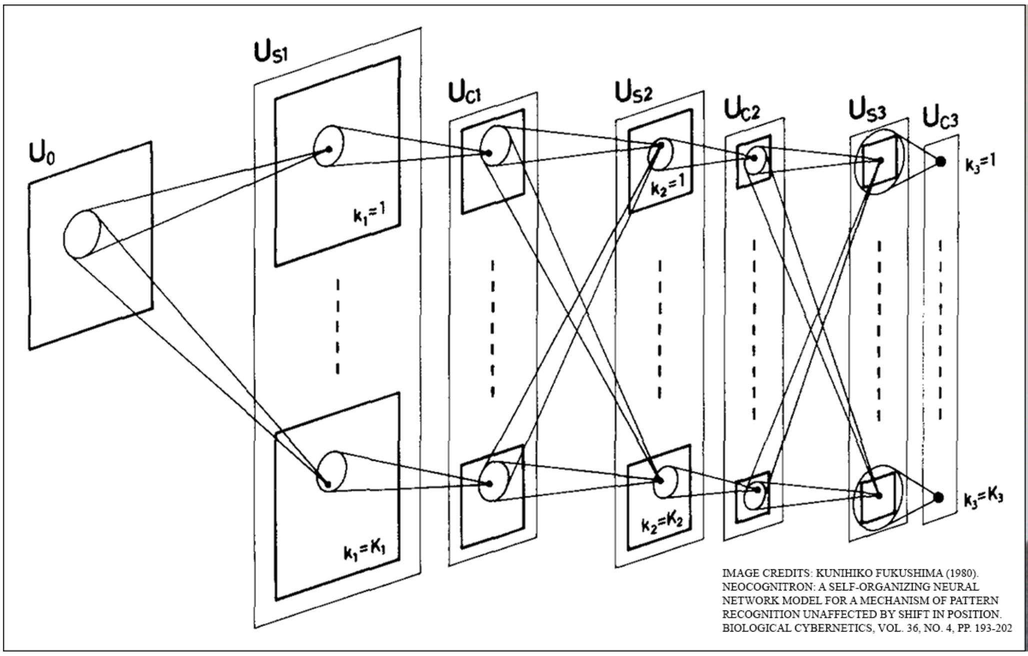


IMAGE CREDITS: KUNIHIKO FUKUSHIMA (1980). NEOCOGNITRON: A SELF-ORGANIZING NEURAL NETWORK MODEL FOR A MECHANISM OF PATTERN RECOGNITION UNAFFECTED BY SHIFT IN POSITION. BIOLOGICAL CYBERNETICS, VOL. 36, NO. 4, PP. 193-202

Jürgen Schmidhuber (August 2025)
Pronounce: You_again Shmidhoobuh

Who invented convolutional neural networks?

Modern AI is based on artificial neural networks (NNs).[DLH] Convolutional NNs (CNNs) are widely used whenever images or videos are involved. But who invented them? Here is the timeline of the origins of CNNs (extending a popular tweet):

★ **1969:** Kunihiko Fukushima published **rectified linear units** or ReLUs[CN69] which are now extensively used in CNNs.

★ **1979:** Fukushima published **the basic CNN architecture with convolution layers and downsampling layers**,[CN79][CN80][CN21] inspired by neurophysiological findings of Hubel and Wiesel.[HUW59-68] His network was trained by *unsupervised* learning rules. Compute was 100 times more expensive than in 1989, and a billion times more expensive than today.

★ **1987:** Alex Waibel (a German researcher working in Japan) trained *supervised* weight sharing NNs with 1-dimensional convolutions (TDNNs) by Linnainmaa's 1970 backpropagation algorithm[BP1-5] to recognise speech.[CN87][CN89c] A similar proposal by Homma et al.[CN87b] introduced the *"convolution"* terminology to NNs.

★ **1988:** Wei Zhang (a Chinese researcher working in Japan) and colleagues had **the first "modern" 2-dimensional CNN trained by backpropagation**, and applied it to character recognition.[CN88] Compute was about 10 million times more expensive than today.

Most of the above was published in Japan 1979-1988.[CN69][CN79][CN87][CN88] Why Japan? Let's look back at the 1980s. Back then, Japan was the envy of the world. Before the 1990 crash,[95-25] the Tokyo stock market was the world's largest, and the world's 6 most valuable public companies were all Japanese. So were the world's richest business men. According to the real estate market, tiny Japan was 4 times more valuable than the much bigger US. The central square mile of Tokyo had the value of California. Japan had far more robots than any other country,[95-25] and by far the most expensive AI project: the 5th Generation Project. Interestingly, this project had little to do with neural networks; it was mostly about logic programming and expert systems. So Fukushima and colleagues were outsiders back then. But at least there was sufficient funding in Japan, even for such unpopular types of blue skies research. Today, the rest of the world can be thankful for that.[CN25]

★ **April 1989:** Wei Zhang et al. also had the first journal submission on "modern" backpropagation-trained CNNs (with applications to character recognition).[CN89] In the early 1990s, Zhang et al. published several additional important CNN papers.[CN91-CN94]

★ **July 1989:** Yann LeCun et al. at Bell Labs had the second journal submission on backpropagation-trained CNNs for character recognition (zip codes),[CN89b][DLP] following the work of Zhang et al.[CN88][CN89] See also Hampshire & Waibel (1989).[CN89d]

★ **1990-93:** Fukushima’s downsampling based on spatial averaging[CN79] was replaced by **max-pooling** for 1-D convolutional NNs (Yamaguchi et al.)[CN90] and for 2-D CNNs (Weng et al.).[CN93]

Many additional CNN papers were published in the 1990s and early 2000s.e.g., [CN98-CN10]

★ **2011:** Many years after the original work on max-pooling CNNs, Schmidhuber's postdoc Dan Ciresan and colleagues dramatically accelerated such CNNs on NVIDIA GPUs,[GPUCNN1][DAN][CN25b] extending their 2010 work on record-breaking GPU-based NNs.[MLP1-3] In 2011, the resulting DanNet achieved the first superhuman visual pattern recognition result.[DAN1] For a while, it enjoyed a monopoly: from May 2011 to Sept 2012, DanNet won every image recognition challenge it entered, 4 of them in a row.[GPUCNN5][MIR][MOST] Admittedly, however, this was mostly about engineering & scaling up the basic insights from the previous millennium, profiting from much faster hardware.

Some "AI experts" claim that "making CNNs work" (e.g., [CN88][CN89][CN89b][DAN]) was as important as inventing them. But "making them work" largely depended on whether your lab was rich enough to buy the latest computers required to scale up the original work. It's the same as today. Basic research is the essential foundation for later engineering/development—the R vs the D in R&D.

More in the Annotated History of Modern AI and Deep Learning.[DLH]

Acknowledgments

Thanks to several expert reviewers for useful comments. (Let me know under *juergen@idsia.ch* if you can spot any remaining error.) The title image was adapted from Fukushima's 1980 paper.[CN80] The remaining contents of this article may be used for educational and non-commercial purposes, including articles for Wikipedia and similar sites.

References

[BP1] S. Linnainmaa. The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. Master's Thesis (in Finnish), Univ. Helsinki, 1970. *See chapters 6-7 and FORTRAN code on pages 58-60.* PDF. See also BIT 16, 146-160, 1976. Link. *The first publication on "modern" backpropagation, also known as the reverse mode of automatic differentiation.*

[BP2] P. J. Werbos. Applications of advances in nonlinear sensitivity analysis. In R. Drenick, F. Kozin, (eds): System Modeling and Optimization: Proc. IFIP, Springer, 1982. PDF. *First application of backpropagation[BP1] to NNs (concretizing thoughts in Werbos' 1974 thesis).*

[BP4] J. Schmidhuber (AI Blog, 2014; updated 2025). Who invented backpropagation? See also LinkedIn post (2025).

[BP5] A. Griewank (2012). Who invented the reverse mode of differentiation? Documenta Mathematica, Extra Volume ISMP (2012): 389-400.

[BPA] H. J. Kelley. Gradient Theory of Optimal Flight Paths. ARS Journal, Vol. 30, No. 10, pp. 947-954, 1960. *Precursor of modern backpropagation.[BP1-4]*

[BPB] A. E. Bryson. A gradient method for optimizing multi-stage allocation processes. Proc. Harvard Univ. Symposium on digital computers and their applications, 1961.

[BPC] S. E. Dreyfus. The numerical solution of variational problems. Journal of Mathematical Analysis and Applications, 5(1): 30-45, 1962.

[CN69] K. Fukushima (1969). Visual feature extraction by a multilayered network of analog threshold elements. IEEE Transactions on Systems Science and Cybernetics. 5 (4): 322-333. doi:10.1109/TSSC.1969.300225. *This work introduced rectified linear units or ReLUs, now widely used in CNNs and other neural nets.*

[CN79] K. Fukushima (1979). Neural network model for a mechanism of pattern recognition unaffected by shift in position—Neocognitron. Trans. IECE, vol. J62-A, no. 10, pp. 658-665, 1979. *The first deep convolutional neural network architecture, with alternating convolutional layers and downsampling layers. In Japanese. English version: [CN80]. More in Scholarpedia.*

[CN80] K. Fukushima: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, vol. 36, no. 4, pp. 193-202 (April 1980). Link.

[CN87] A. Waibel. Phoneme Recognition Using Time-Delay Neural Networks. Meeting of IEICE, Tokyo, Japan, 1987. *Application of backpropagation[BP1][BP2] and weight sharing to a 1-dimensional convolutional architecture.*

[CN87b] T. Homma, L. Atlas; R. Marks II (1987). An Artificial Neural Network for Spatio-Temporal Bipolar Patterns: Application to Phoneme Classification. Advances in Neural Information Processing Systems (N(eur)IPS), 1:31-40.

[CN88] W. Zhang, J. Tanida, K. Itoh, Y. Ichioka. Shift-invariant pattern recognition neural network and its optical architecture. Proc. Annual Conference of the Japan Society of Applied Physics, 1988. PDF. *First "modern" backpropagation-trained 2-dimensional CNN, applied to character recognition.*

[CN89] W. Zhang, J. Tanida, K. Itoh, Y. Ichioka (received 13 April 1989). Parallel distributed processing model with local space-invariant interconnections and its optical architecture. Applied Optics / Vol. 29, No. 32, 1990. PDF. *First journal submission on a "modern" backpropagation-trained 2-dimensional CNN (applied to character recognition).*

[CN89b] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel (received July 1989). Backpropagation Applied to Handwritten Zip Code Recognition, Neural Computation, 1(4):541-551, 1989. *Second journal submission on a "modern" backpropagation-trained 2-dimensional CNN (applied to character recognition). Compare [CN88][CN89].*

[CN89c] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. J. Lang. Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 3, pp. 328-339, March 1989. *Based on [CN87] (1-dimensional convolutions).*

[CN89d] J. Hampshire, A. Waibel (1989). Connectionist architectures for multi-speaker phoneme recognition. In Advances in Neural Information Processing Systems, N(eur)IPS'2. *Conference publication on 2D-TDNNs or 2D-CNNs for speech recognition.*

[CN90] K. Yamaguchi, K. Sakamoto, A. Kenji, T. Akabane, Y. Fujimoto. A Neural Network for Speaker-Independent Isolated Word Recognition. First International Conference on Spoken Language Processing (ICSLP 90), Kobe, Japan, Nov 1990. *A 1-dimensional NN with convolutions using Max-Pooling instead of Fukushima's Spatial Averaging.*[CN79]

[CN93] Weng, J., Ahuja, N., and Huang, T. S. (1993). Learning recognition and segmentation of 3-D objects from 2-D images. Proc. 4th Intl. Conf. Computer Vision, Berlin, Germany, pp. 121-128. *A 2-dimensional CNN whose downsampling layers use Max-Pooling (which has become very popular) instead of Fukushima's Spatial Averaging.*[CN79]

[CN91] W. Zhang, A. Hasegawa, K. Itoh, Y. Ichioka. Image processing of human corneal endothelium based on a learning network. Applied Optics, Vol. 30, No. 29, 1991. *First published CNN-based image segmentation.*

[CN91b] W. Zhang, A. Hasegawa, K. Itoh, Y. Ichioka. Error Back Propagation With Minimum-Entropy Weights: A Technique for Better Generalization of 2-D Shift-Invariant NNs. IJCNN, 1991. *Got an IJCNN student paper award.*

[CN92] W. Zhang, A. Hasegawa, O. Matoba, K. Itoh, Y. Ichioka, K. Doi. Shift-invariant Neural Network for Image Processing: Learning and Generalization. SPIE Vol. 1709, Application of Artificial Neural Networks III, Orlando, 1992.

[CN94] W. Zhang, K. Doi, M. Giger, Y. Wu, R. Nishikawa, R. Schmidt. Computerized detection of cluster microcalcifications in digital mammogram using a shift-invariant neural network. Medical Physics, 21(4), 1994. *First CNN for object detection, commercialised by R2 Technology, which processed over 30 million mammography exams annually to aid radiologists in breast cancer detection.*

[CN98] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE. 86 (11): 2278-2324. *This work about backpropagation-trained 2-dimensional CNNs for character recognition failed to cite the original work on this by Zhang et al. (1988).*[CN88][CN89]

[CN99] S. Behnke. Learning iterative image reconstruction in the neural abstraction pyramid. International Journal of Computational Intelligence and Applications, 1(4):427-438, 1999.

[CN03] S. Behnke. Hierarchical Neural Networks for Image Interpretation, volume LNCS 2766 of Lecture Notes in Computer Science. Springer, 2003.

[CN07] M. A. Ranzato, Y. LeCun: A Sparse and Locally Shift Invariant Feature Extractor Applied to Document Images. Proc. ICDAR, 2007

[CN10] D. Scherer, A. Mueller, S. Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In Proc. International Conference on Artificial Neural Networks (ICANN), pages 92-101, 2010.

[CN21] Bower Award Ceremony 2021: Jürgen Schmidhuber lauds Kunihiko Fukushima. YouTube video, 2021.

[CN25] J. Schmidhuber (AI Blog, 2025). Who invented convolutional neural networks? See popular tweet.

[CN25b] F. Chollet, ex-Google, creator of Keras (tweets, 3 Aug 2025): *"The big breakthrough for convnets was the first GPU-accelerated CUDA implementation, which immediately started winning first place in image classification competitions. Remember when that happened? I do. That was Dan Ciresan in 2011 [...] it was an actual system that won 2 competitions and even got significant media coverage at the time. Most computer vision researchers knew about it in late 2011 / early 2012."*

[DAN] J. Schmidhuber (AI Blog, 2021). In 2011, DanNet triggered the deep convolutional neural network (CNN) revolution. *Named after Schmidhuber's outstanding postdoc Dan Ciresan, it was the first deep and fast CNN to win international computer vision contests, and had a temporary monopoly on winning them, driven by a very fast implementation based on graphics processing units (GPUs). 1st superhuman result in 2011.*[DAN1] *Now everybody is using this approach.*

[DAN1] J. Schmidhuber (AI Blog, 2011; updated 2021 for 10th birthday of DanNet): First superhuman visual pattern recognition. *At the IJCNN 2011 computer vision competition in Silicon Valley, the artificial neural network called* DanNet *performed twice better than humans, three times better than the closest artificial competitor (from LeCun's team), and six times better than the best non-neural method.*

[DEC] J. Schmidhuber (AI Blog, 02/20/2020, updated 2025). The 2010s: Our Decade of Deep Learning / Outlook on the 2020s. *The recent decade's most important developments and industrial applications based on our AI, with an outlook on the 2020s, also addressing privacy and data markets.*

[DEEP1] Ivakhnenko, A. G. and Lapa, V. G. (1965). Cybernetic Predicting Devices. CCM Information Corporation. *First working Deep Learners with many layers, learning internal representations.*

[DEEP1a] Ivakhnenko, Alexey Grigorevich. The group method of data of handling; a rival of the method of stochastic approximation. Soviet Automatic Control 13 (1968): 43-55.

[DEEP2] Ivakhnenko, A. G. (1971). Polynomial theory of complex systems. IEEE Transactions on Systems, Man and Cybernetics, (4):364-378.

[DL1] J. Schmidhuber, 2015. Deep learning in neural networks: An overview. Neural Networks, 61, 85-117. More. *Got the first Best Paper Award ever issued by the journal Neural Networks, founded in 1988.*

[DLC1] Y. LeCun. IEEE Spectrum Interview by L. Gomes, Feb 2015. *Quote: "A lot of us involved in the resurgence of Deep Learning in the mid-2000s, including Geoff Hinton, Yoshua Bengio, and myself—the so-called 'Deep Learning conspiracy' ..."*

[DLC2] M. Bergen, K. Wagner (2015). Welcome to the AI Conspiracy: The 'Canadian Mafia' Behind Tech's Latest Craze. Vox recode, 15 July 2015. *Quote: "... referred to themselves as the 'deep learning conspiracy.' Others called them the 'Canadian Mafia.'"*

[DLH] J. Schmidhuber (AI Blog, 2022). Annotated History of Modern AI and Deep Learning. Technical Report IDSIA-22-22, IDSIA, Lugano, Switzerland, 2022. Preprint arXiv:2212.11279. Tweet of 2022.

[DLP] J. Schmidhuber (AI Blog, 2023). How 3 Turing awardees republished key methods and ideas whose creators they failed to credit. Technical Report IDSIA-23-23, Swiss AI Lab IDSIA, 14 Dec 2023. Tweet of 2023.

[Drop1] S. J. Hanson (1990). A Stochastic Version of the Delta Rule, PHYSICA D,42, 265-272. *What's now called "dropout" is a variation of the stochastic delta rule—compare preprint arXiv:1808.03578, 2018.*

[Drop2] N. Frazier-Logue, S. J. Hanson (2020). The Stochastic Delta Rule: Faster and More Accurate Deep Learning Through Adaptive Weight Noise. Neural Computation 32(5):1018-1032.

[Drop3] J. Hertz, A. Krogh, R. Palmer (1991). Introduction to the Theory of Neural Computation. Redwood City, California: Addison-Wesley Pub. Co., pp. 45-46.

[Drop4] N. Frazier-Logue, S. J. Hanson (2018). Dropout is a special case of the stochastic delta rule: faster and more accurate deep learning. Preprint arXiv:1808.03578, 2018.

[GDa] Y. Z. Tsypkin (1966). Adaptation, training and self-organization automatic control systems, Avtomatika I Telemekhanika, 27, 23-61. *On gradient descent-based on-line learning for non-linear systems.*

[GDb] Y. Z. Tsypkin (1971). Adaptation and Learning in Automatic Systems, Academic Press, 1971. *On gradient descent-based on-line learning for non-linear systems.*

[GD1] S. I. Amari (1967). A theory of adaptive pattern classifier, IEEE Trans, EC-16, 279-307 (Japanese version published in 1965). PDF. *Probably the first paper on using stochastic gradient descent[STO51-52] for learning in multilayer neural networks (without specifying the specific gradient descent method now known as reverse mode of automatic differentiation or backpropagation[BP1]).*

[GD2] S. I. Amari (1968). Information Theory—Geometric Theory of Information, Kyoritsu Publ., 1968 (in Japanese). OCR-based PDF scan of pages 94-135 (see pages 119-120). *Contains computer simulation results for a five layer network (with 2 modifiable layers) which learns internal representations to classify non-linearily separable pattern classes.*

[GD2a] H. Saito (1967). Master's thesis, Graduate School of Engineering, Kyushu University, Japan. *Implementation of Amari's 1967 stochastic gradient descent method for multilayer perceptrons.[GD1] (S. Amari, personal communication, 2021.)*

[95-25] J. Schmidhuber (AI Blog, 2025). 1995-2025: The Decline of Germany & Japan vs US & China. Can All-Purpose Robots Fuel a Comeback? In 1995, in terms of nominal gross domestic product (GDP), a combined Germany and Japan were almost 1:1 economically with a combined USA and China, according to IMF. Only 3 decades later, this ratio is now down to 1:5! Self-replicating AI-driven all-purpose robots may be the answer. Based on a 2024 F.A.Z. guest article.

[GPUNN] Oh, K.-S. and Jung, K. (2004). GPU implementation of neural networks. Pattern Recognition, 37(6):1311-1314. *Speeding up traditional NNs on GPU by a factor of 20.*

[GPUCNN] K. Chellapilla, S. Puri, P. Simard. High performance convolutional neural networks for document processing. International Workshop on Frontiers in Handwriting Recognition, 2006. *Speeding up shallow CNNs on GPU by a factor of 4.*

[GPUCNN1] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, J. Schmidhuber. Flexible, High Performance Convolutional Neural Networks for Image Classification. *International Joint Conference on Artificial Intelligence (IJCAI-2011, Barcelona)*, 2011. PDF. ArXiv preprint. *Speeding up deep CNNs on GPU by a factor of 60. Used to win four important computer vision competitions 2011-2012 before others won any with similar approaches.*

[GPUCNN2] D. C. Ciresan, U. Meier, J. Masci, J. Schmidhuber. A Committee of Neural Networks for Traffic Sign Classification. *International Joint Conference on Neural Networks (IJCNN-2011, San Francisco)*, 2011. PDF. HTML overview. *First superhuman performance in a computer vision contest, with half the error rate of humans, and one third the error rate of the closest competitor.*[DAN1] *This led to massive interest from industry.*

[GPUCNN3] D. C. Ciresan, U. Meier, J. Schmidhuber. Multi-column Deep Neural Networks for Image Classification. Proc. *IEEE Conf. on Computer Vision and Pattern Recognition CVPR 2012*, p 3642-3649, July 2012. PDF. Longer TR of Feb 2012: arXiv:1202.2745v1 [cs.CV]. More.

[GPUCNN4] A. Krizhevsky, I. Sutskever, G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. NIPS 25, MIT Press, Dec 2012. PDF. *This paper describes AlexNet, which is similar to the earlier DanNet,*[DAN,DAN1][R6] *the first pure deep CNN to win computer vision contests in 2011*[GPUCNN2-3,5] *(AlexNet and VGG Net*[GPUCNN9] *followed in 2012-2014). [GPUCNN4] emphasizes benefits of Fukushima's ReLUs (1969)*[CN69] *and dropout (a variant of Hanson 1990 stochastic delta rule)*[Drop1-4] *but neither cites the original work*[CN69][Drop1] *nor the basic CNN architecture (Fukushima, 1979).*[CN79]

[GPUCNN5] J. Schmidhuber (AI Blog, 2017; updated 2021 for 10th birthday of DanNet): History of computer vision contests won by deep CNNs since 2011. DanNet was the first CNN to win one, and won 4 of them in a row before the similar AlexNet/VGG Net and the Resnet (a Highway Net with open gates) joined the party. Today, deep CNNs are standard in computer vision.

[GPUCNN6] J. Schmidhuber, D. Ciresan, U. Meier, J. Masci, A. Graves. On Fast Deep Nets for AGI Vision. In Proc. Fourth Conference on Artificial General Intelligence (AGI-11), Google, Mountain View, California, 2011. PDF.

[GPUCNN7] D. C. Ciresan, A. Giusti, L. M. Gambardella, J. Schmidhuber. Mitosis Detection in Breast Cancer Histology Images using Deep Neural Networks. MICCAI 2013. PDF.

[GPUCNN8] J. Schmidhuber (AI Blog, 2017; updated 2021 for 10th birthday of DanNet). First deep learner to win a contest on object detection in large images— first deep learner to win a medical imaging contest (2012). Link. *How the Swiss AI Lab IDSIA used GPU-based CNNs to win the ICPR 2012 Contest on Mitosis Detection and the MICCAI 2013 Grand Challenge.*

[GPUCNN9] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition. Preprint arXiv:1409.1556 (2014).

[HUW59] Wiesel, D. H. and Hubel, T. N. (1959). Receptive fields of single neurones in the cat’s striate cortex. J. Physiol., 148:574-591.

[HUW62] Hubel, D. H. andWiesel, T. (1962). Receptive fields, binocular interaction, and functional architecture in the cat’s visual cortex. Journal of Physiology (London), 160:106-154.

[HUW68] Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. The Journal of Physiology, 195(1):215-243.

[HW] J. Schmidhuber (AI Blog, 2015, updated 2025 for 10-year anniversary). Overview of Highway Networks: First working really deep feedforward neural networks with over 100 layers.

[HW] J. Schmidhuber (AI Blog, 2015, updated 2025 for 10-year anniversary). Overview of Highway Networks: First working really deep feedforward neural networks with over 100 layers.

[HW1] R. K. Srivastava, K. Greff, J. Schmidhuber. Highway networks. Preprints arXiv:1505.00387 (May 2015) and arXiv:1507.06228 (Training Very Deep Networks; July 2015). Also at NeurIPS 2015. *The first working very deep gradient-based feedforward neural nets (FNNs) with hundreds of layers (previous FNNs had at most a few tens of layers). Let g, t, h, denote non-linear differentiable functions. Each non-input layer of a Highway Net computes g(x)x + t(x)h(x), where x is the data from the previous layer. The gates g(x) are typically initialised to*

1.0, to obtain residual connections for very deep error propagation (this is what makes Highway NNs so deep). The later Resnets (Dec 2015)[HW2] are like a variant of this where all gates are always open: g(x)=t(x)=const=1. That is, Highway Nets are gated ResNets: set the gates to 1.0→ResNet. Highway Nets perform roughly as well as ResNets[HW2] on ImageNet.[HW3] Variants of Highway gates are also used for certain algorithmic tasks, where the simpler residual layers do not work as well.[NDR] See also [HW25b]. More.

[HW1a] R. K. Srivastava, K. Greff, J. Schmidhuber. Highway networks. Presentation at the Deep Learning Workshop, ICML'15, July 10-11, 2015. Link.

[HW2] He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. Preprint arXiv:1512.03385 (Dec 2015). *Residual nets are essentially open-gated variants of the earlier very deep Highway Nets (May 2015) [HW1]. In turn, Highway Nets are gated ResNets: set the gates to 1.0→ResNet. In fact, the gates of the residual connections in Highway Nets are typically initialised to be open (1.0) anyway, to permit very deep error propagation. More.*

[HW3] K. Greff, R. K. Srivastava, J. Schmidhuber. Highway and Residual Networks learn Unrolled Iterative Estimation. Preprint arxiv:1612.07771 (2016). Also at ICLR 2017.

[LEC] J. Schmidhuber (AI Blog, 2022). LeCun's 2022 paper on autonomous machine intelligence rehashes but does not cite essential work of 1990-2015. *Years ago, Schmidhuber's team published most of what Y. LeCun calls his "main original contributions:" neural nets that learn multiple time scales and levels of abstraction, generate subgoals, use intrinsic motivation to improve world models, and plan (1990); controllers that learn informative predictable representations (1997), etc. This was also discussed on Hacker News, reddit, and in the media. See tweet1. LeCun also listed the "5 best ideas 2012-2022" without mentioning that most of them are from Schmidhuber's lab, and older. See tweet2.*

[MGC] MICCAI 2013 Grand Challenge on Mitosis Detection, organised by M. Veta, M.A. Viergever, J.P.W. Pluim, N. Stathonikos, P. J. van Diest of University Medical Center Utrecht.

[MIR] J. Schmidhuber (Oct 2019, updated 2021, 2022, 2025). Deep Learning: Our Miraculous Year 1990-1991. Preprint arXiv:2005.05744. *The Deep Learning Artificial Neural Networks (NNs) of our team have revolutionised Machine Learning & AI. Many of the basic ideas behind this revolution were published within the 12 months of our "Annus Mirabilis" 1990-1991 at our lab in TU Munich. Back then, few people were interested. But a quarter century later, NNs based on our "Miraculous Year" were on over 3 billion devices, and used many billions of times per day, consuming a significant fraction of the world's compute. In particular, in 1990-91, we laid foundations of Generative AI, publishing principles of (1) Generative Adversarial Networks for Artificial Curiosity and Creativity (now used for deepfakes), (2) Transformers (the T in ChatGPT—see the 1991 Unnormalized Linear Transformer), (3) Pre-training for deep NNs (see the P in ChatGPT), (4) NN distillation (key for DeepSeek), and (5) recurrent World Models for Reinforcement Learning and Planning in partially observable environments. The year 1991 also marks the emergence of the defining features of (6) LSTM, the most cited AI paper of the 20th century (based on constant error flow through residual NN connections), and (7) ResNet, the most cited AI paper of the 21st century, based on our LSTM-inspired Highway Net that was 10 times deeper than previous feedforward NNs.*

[MLP1] D. C. Ciresan, U. Meier, L. M. Gambardella, J. Schmidhuber. Deep Big Simple Neural Nets For Handwritten Digit Recognition. Neural Computation 22(12): 3207-3220, 2010. ArXiv Preprint. *Showed that plain backprop for deep standard NNs is sufficient to break benchmark records, without any unsupervised pre-training.*

[MLP2] J. Schmidhuber (AI Blog, Sep 2020). 10-year anniversary of supervised deep learning breakthrough (2010). No unsupervised pre-training. *By 2010, when compute was 100 times more expensive than today, both the feedforward NNs[MLP1] and the earlier recurrent NNs of Schmidhuber's team were able to beat all competing algorithms on important problems of that time.*

[MLP3] J. Schmidhuber (AI Blog, 2025). 2010: Breakthrough of end-to-end deep learning (no layer-by-layer training, no unsupervised pre-training). The rest is history. *By 2010, when compute was 1000 times more expensive than in 2025, both our feedforward NNs[MLP1] and our earlier recurrent NNs were able to beat all competing algorithms on important problems of that time. This deep learning revolution quickly spread from Europe to North America and Asia.*

[MOST] J. Schmidhuber (AI Blog, 2021, updated 2025). The most cited neural networks all build on work done in my labs: *1. Long Short-Term Memory (LSTM), the most cited AI of the 20th century. 2. ResNet (open-gated Highway Net), the most cited AI of the 21st century. 3. AlexNet & VGG Net (the similar but earlier DanNet of 2011 won 4 image recognition challenges before them). 4. GAN (an instance of Adversarial Artificial Curiosity of 1990). 5. Transformer variants—see the 1991 unnormalised linear Transformer (ULTRA). Foundations of Generative AI were published in 1991: the principles of GANs (now used for deepfakes), Transformers (the T in ChatGPT), Pre-training for deep NNs (the P in ChatGPT), NN distillation, and the famous DeepSeek—see the tweet.*

[NDR] R. Csordas, K. Irie, J. Schmidhuber. The Neural Data Router: Adaptive Control Flow in Transformers Improves Systematic Generalization. Proc. ICLR 2022. Preprint arXiv/2110.07732.

[NOB] J. Schmidhuber. A Nobel Prize for Plagiarism. Technical Report IDSIA-24-24 (7 Dec 2024, updated 31 July 2025). *Sadly, the Nobel Prize in Physics 2024 for Hopfield & Hinton is a Nobel Prize for plagiarism. They republished methodologies for artificial neural networks developed in Ukraine and Japan by Ivakhnenko and Amari in the 1960s & 1970s, as well as other techniques, without citing the original papers. Even in later surveys, they didn't credit the original inventors (thus turning what may have been unintentional plagiarism into a deliberate form). None of the important algorithms for modern Artificial Intelligence were created by Hopfield & Hinton. See also popular tweet1, tweet2, and LinkedIn post.*

[R1] Reddit/ML, 2019. Hinton, LeCun, Bengio receive ACM Turing Award. *This announcement contains more comments about Schmidhuber than about any of the awardees.*

[R4] Reddit/ML, 2019. Five major deep learning papers by G. Hinton did not cite similar earlier work by J. Schmidhuber.

[R6] Reddit/ML, 2019. DanNet, the CUDA CNN of Dan Ciresan in J. Schmidhuber's team, won 4 image recognition challenges prior to AlexNet.

[RELU1] K. Fukushima (1969). Visual feature extraction by a multilayered network of analog threshold elements. IEEE Transactions on Systems Science and Cybernetics. 5 (4): 322-333. doi:10.1109/TSSC.1969.300225. *This work introduced rectified linear units or ReLUs, now widely used.*

[RELU2] C. v. d. Malsburg (1973). Self-Organization of Orientation Sensitive Cells in the Striate Cortex. Kybernetik, 14:85-100, 1973. *See Table 1 for rectified linear units or ReLUs. Possibly this was also the first work on applying an EM algorithm to neural nets.*

[SCAN] J. Masci, A. Giusti, D. Ciresan, G. Fricout, J. Schmidhuber. A Fast Learning Algorithm for Image Segmentation with Max-Pooling Convolutional Networks. ICIP 2013. Preprint arXiv:1302.1690.

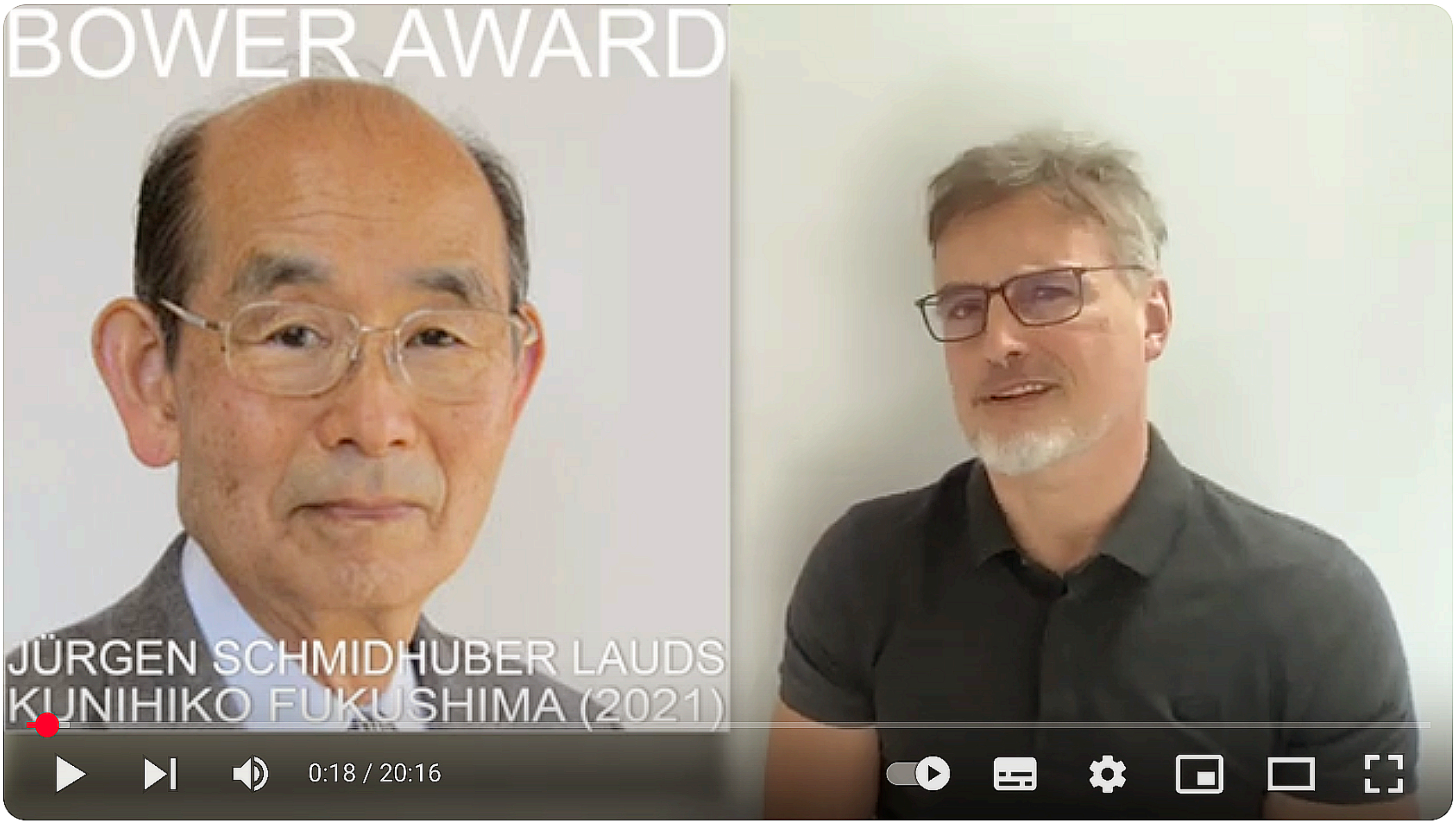
[ST] J. Masci, U. Meier, D. Ciresan, G. Fricout, J. Schmidhuber Steel Defect Classification with Max-Pooling Convolutional Neural Networks. Proc. IJCNN 2012. PDF.

[VAN1] S. Hochreiter. Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, TUM, 1991 (advisor J. Schmidhuber). PDF. *More on the Fundamental Deep Learning Problem.*

[VAN2] Y. Bengio, P. Simard, P. Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE TNN 5(2), p 157-166, 1994. *Results are essentially identical to those of Schmidhuber's diploma student Sepp Hochreiter (1991).[VAN1] Even after a common publication,[VAN3] the first author of [VAN2] published papers[VAN4] that cited only their own [VAN2] but not the original work.*

[VAN3] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In S. C. Kremer and J. F. Kolen, eds., A Field Guide to Dynamical Recurrent Neural Networks. IEEE press, 2001. PDF.

[VAN4] Y. Bengio. Neural net language models. Scholarpedia, 3(1):3881, 2008. Link.


BOWER AWARD
JÜRGEN SCHMIDHUBER LAUDS
KUNIHIKO FUKUSHIMA (2021)
0:18 / 20:16