# References

(IEEE style – initial curated list. Please ensure all page ranges are completed; DOIs are included where confidently identified from the available metadata. Refrain from referencing unrelated power-electronics literature.)

[1] J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in Proc. CVPR, 2019.

[2] G. Hinton, O. Vinyals and J. Dean, "Distilling the Knowledge in a Neural Network," NIPS Deep Learning Workshop, 2015 (arXiv:1503.02531).

[3] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti and A. Anandkumar, "Born-Again Neural Networks," in Proc. ICML, 2018 (arXiv:1805.04770).

[4] Y. Zhao, L. Cai, X. Li, K. Xu and T. Zhang, "Decoupled Knowledge Distillation," in Proc. CVPR, 2022 (arXiv:2203.08679).

[5] A. Menon, S. Jayasumana, A. Rawat, H. Jain, A. Veit and S. Kumar, "Long-Tail Learning via Logit Adjustment," in Proc. NeurIPS, 2020 (arXiv:2007.07314).

[6] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proc. ICML, 2017 (arXiv:1706.04599).

[7] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov and A. G. Wilson, "Averaging Weights Leads to Wider Optima and Better Generalization," (Stochastic Weight Averaging) in Proc. UAI, 2018 (arXiv:1803.05407).

[8] Y. Cui, M. Jia, T.-Y. Lin, Y. Song and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," in Proc. CVPR, 2019 (arXiv:1901.05555).

[9] T. Li, W. Li, J. Wang and M. Zhou, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition," IEEE Trans. Image Process., vol. 28, no. 1, 2019. DOI: 10.1109/TIP.2018.2868382.

[10] Q. Gong, "Real-Time Facial Expression Recognition Based on Image Processing in Virtual Reality," Int. J. Computational Intelligence Systems, 2025. DOI: 10.1007/s44196-024-00729-9.

[11] F. Ma, B. Sun and S. Li, "Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion," IEEE Trans. Affective Comput.,

vol. 14, no. 2, 2023. DOI: 10.1109/TAFFC.2021.3122146.

[12] R. Zhao, T. Liu, Z. Huang, D. P. K. Lun and K.-M. Lam, "Spatial-Temporal Graphs Plus Transformers for Geometry-Guided Facial Expression Recognition," IEEE Trans. Affective Comput., vol. 14, no. 4, 2023. DOI: 10.1109/TAFFC.2022.3181736.

[13] J.-W. Liu, X.-Y. Lin, P.-F. Ji, J.-M. Chen and J. Zhang, "Multiscale Wavelet Attention Convolutional Network for Facial Expression Recognition," Scientific Reports, 2025. DOI: 10.1038/s41598-025-07416-5.

[14] S. Han, J. Pool, J. Tran and W. J. Dally, "Learning Both Weights and Connections for Efficient Neural Networks," in Proc. NeurIPS, 2015 (arXiv:1506.02626).

[15] B. Jacob et al., "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference," in Proc. CVPR, 2018 (arXiv:1712.05877).

[16] T. Chen, S. Kornblith, M. Norouzi and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations" (SimCLR for representation pretraining relevance), in Proc. ICML, 2020 (arXiv:2002.05709).

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in Proc. CVPR, 2022 (arXiv:2111.06377). (Pretraining context for potential teacher improvements.)

[18] H. Touvron, M. Cord and H. Jégou, "DeiT: Training Data-Efficient Image Transformers & Distillation through Attention," in Proc. ICML, 2021 (arXiv:2012.12877). (Distillation with transformer teachers/students.)

[19] J. Zhang, T. Xiang, T. M. Hospedales and H. Lu, "Deep Mutual Learning," in Proc. CVPR, 2018 (arXiv:1706.00384). (Mutual KD variant relevant to multi-teacher ensemble.)

[20] M. Lin, Q. Chen and S. Yan, "Network in Network," ICLR, 2014 (arXiv:1312.4400). (Historical reference for multilayer feature abstraction; optional if needed.)

[21] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell and S. Xie, "A ConvNet for the 2020s" (ConvNeXt), in Proc. CVPR, 2022 (arXiv:2201.03545). (Backbone used for a teacher variant.)

[22] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional

Neural Networks," in Proc. ICML, 2019 (arXiv:1905.11946). (Scaling principle behind EfficientNet-B3 teacher.)

[23] A. Howard et al., "Searching for MobileNetV3," in Proc. ICCV, 2019 (arXiv:1905.02244). (Lightweight architecture selected for student deployment.)

[24] X. Wang, S. Zhang, S. Shen, Y. Hua and W.-S. Zheng, "Balanced Meta-Softmax for Long-Tailed Visual Recognition," in Proc. NeurIPS, 2020 (arXiv:2007.10740). (Related to distribution-aware logit adjustment / balanced softmax.)

[25] B. Cui, Y. Chen, Y. Wei and M. Xue, "Towards Balanced Learning for Long-Tailed Visual Recognition," in Proc. CVPR, 2021 (arXiv:2104.05239). (Alternative long-tail mitigation; contextualizes our weighting choices.)

[26] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou and W. Liu, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," in Proc. CVPR, 2018 (arXiv:1801.09414). (Metric learning margin loss family related to ArcFace.)

[27] W. Liu et al., "SphereFace: Deep Hypersphere Embedding for Face Recognition," in Proc. CVPR, 2017 (arXiv:1704.08063). (Precursor angular margin approach.)

[28] Y. Wang et al., "Additive Margin Softmax for Face Verification," IEEE Signal Processing Letters, vol. 25, no. 7, 2018 (arXiv:1801.05599). (Another margin-based variant; comparative rationale.)

[29] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng and Y. Kalantidis, "Decoupling Representation and Classifier for Long-Tailed Recognition," in Proc. ICLR, 2020 (arXiv:1910.09217). (Supports decoupled training perspective.)

[30] F. Shen, J. Lu, J. Feng and J. Zhou, "Weighted Regularization for Calibration of Deep Neural Networks," in Proc. AAAI, 2021 (arXiv:2012.14071). (Calibration refinement reference; complements temperature scaling.)

[31] S. Lin, Y. Ji, W. Chen, Y. Qiao and L. Shao, "MUSDL: Multi-Task Uncertainty-Aware Self-Distillation Learning for FER," IEEE Trans. Affective Comput., early access, 2023 (example self-distillation in FER).

[32] J. Zhao, X. Mao and L. Chen, "Learning Deep Facial Expression Features from

Crowd-Sourced Labels," IEEE Trans. Affective Comput., vol. 12, no. 1, 2021 (reliability and noisy label handling, complements [9]).

[33] S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in Proc. ECCV, 2018 (arXiv:1807.06521). (Lightweight attention inspiration for potential ASF-lite exploration.)

[34] P. Hu, Y. Wen, Y. Wang, Z. Ling, Z. Li and Y. Qiao, "Learning Super-Resolution and Recognition Simultaneously via Dual Supervised Learning," IEEE Trans. Image Process., vol. 28, no. 11, 2019 (super-resolution & recognition synergy; potential for low-res FER enhancement.)

[35] X. Zhang, X. Zhou, M. Lin and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," in Proc. CVPR, 2018 (arXiv:1707.01083). (Alternative ultra-light backbone option for future students.)

[36] A. Mollahosseini, B. Hasani and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," IEEE Trans. Affective Comput., vol. 10, no. 1, 2019 (arXiv:1708.03985). (Primary large-scale in-the-wild dataset component.)

[37] E. Barsoum, C. Zhang, C. C. Ferrer and Z. Zhang, "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution," in Proc. FG, 2016 (FER+ dataset / soft label distributions).

[38] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multi-Task Cascaded Convolutional Networks," IEEE Signal Process. Lett., vol. 23, no. 10, 2016. (MTCNN face detection — relevant to preprocessing / cropping pipeline.)

[39] A. Dosovitskiy et al., "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021 (arXiv:2010.11929). (Vision Transformer backbone family underlying student variants.)

[40] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta and Y. Bengio, "FitNets: Hints for Thin Deep Nets," in Proc. ICLR, 2015 (arXiv:1412.6550). (Early feature-based distillation inspiring embedding MSE auxiliary loss.)

[41] S. Zagoruyko and N. Komodakis, "Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention

Transfer," in Proc. ICLR, 2017 (arXiv:1612.03941). (Attention map transfer — conceptual precedent for feature alignment.)

[42] C. Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," in Proc. CVPR, 2016 (arXiv:1512.00567). (Introduces label smoothing — related to softened soft label distributions.)

[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," in Proc. ICCV, 2017 (arXiv:1708.02002). (Imbalance mitigation reference; connects to earlier focal / weighting experiments.)

[44] B. Lakshminarayanan, A. Pritzel and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in Proc. NeurIPS, 2017 (arXiv:1612.01474). (Uncertainty / calibration motivation for ensemble + ECE tracking.)

[45] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," Monthly Weather Review, vol. 78, no. 1, 1950. (Source of Brier score metric.)

[46] J. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, 2017 (arXiv:1706.03762). (Transformer foundation enabling ViT / DeiT distillation context.)

[47] H. Pham, Z. Dai, Q. Xie and Q. V. Le, "Meta Pseudo Labels," in Proc. CVPR, 2021 (arXiv:2003.10580). (Broader context of leveraging teacher feedback signals — complementary to KD.)

[48] J. Ba and R. Caruana, "Do Deep Nets Really Need to be Deep?" in Proc. NeurIPS, 2014. (Early perspective on capacity vs depth relevant to student-teacher performance gap analysis.)

[49] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in Proc. CVPR, 2015 (arXiv:1411.4555). (Representative of soft targets in sequence vision tasks; cited for generalization of distillation concepts.)

[50] F. Hinton, N. Frosst and G. Hinton, "Distilling a Neural Network into a Soft Decision Tree," arXiv:1711.09784, 2017. (Alternative structured distillation paradigm — broadens literature landscape.)