BIBLIOGRAPHY

Abdallah, S., Alencar-Brayner, A., Benetos, E., Cottrell, S., Dykes, J., Gold, N., ... Wolff, D. (2015). Automatic Transcription and Pitch Analysis of the British Library World and Traditional Music Collection. In *Proceedings of the International Workshop on Folk Music Analysis (FMA)* (pp. 10–12). 10.

Abdallah, S., & Plumbley, M. (2004). Polyphonic Music Transcription by Non-Negative Sparse Coding of Power Spectra. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*. 22.

Agostini, A., & Ghisi, D. (2013). Real-Time Computer-Aided Composition with bach. *Contemporary Music Review*, *32*(1), 41–48. 10 and 73.

Agostini, G., Longari, M., & Pollastri, E. (2003). Musical Instrument Timbres Classification With Spectral Features. *EURASIP Journal on Advances in Signal Processing*, 97–102. 109.

Argenti, F., Nesi, P., & Pantaleo, G. (2011). Automatic Transcription of Polyphonic Music Based on the Constant-Q Bispectral Analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(6), 1610–1630. 18.

Arjovsky, M., & Bottou, L. (2017). Towards Principled Methods for Training Generative Adversarial Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)* (pp. 1–17). 44 and 48.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. *arXiv preprint arXiv:1701.07875*. 45.

Babacan, O., Drugman, T., D'Alessandro, N., Henrich, N., & Dutoit, T. (2013). A Comparative Study of Pitch Extraction Algorithms on a Large Variety of Singing Sounds. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7815–7819). 21.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 35.

Barry, S., & Kim, Y. (2018). Style Transfer for Musical Audio Using Multiple Time-Frequency Representations. *OpenReview:BybQ7zWCb*. 26.

Bell, A. J., & Sejnowski, T. J. (1995). An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, *7*(6), 1129–1159. 26.

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. B. (2005). A Tutorial on Onset Detection in Music Signals. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 1035–1046. 18.

Benetos, E., & Dixon, S. (2011). Polyphonic Music Transcription Using Note Onset and Offset Detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 19.

Benetos, E., & Dixon, S. (2012). A Shift-Invariant Latent Variable Model for Automatic Music Transcription. *Computer Music Journal*, *36*(4), 81–94. 23.

Benetos, E., & Dixon, S. (2013). Multiple-Instrument Polyphonic Music Ttranscription Using a Temporally Constrained Shift-Invariant Model. *The Journal of the Acoustical Society of America*, *133*(3), 1727–1741. 22.

Benetos, E., Dixon, S., Duan, Z., & Ewert, S. (2019). Automatic Music Transcription: An Overview. *IEEE Signal Processing Magazine*, *36*(1), 20–30. 22, 24, and 90.

Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., & Klapuri, A. P. (2013). Automatic Music Transcription: Challenges and Future Directions. *Journal of Intelligent Information Systems*, *41*(3), 407–434. 2.

Benetos, E., & Weyde, T. (2015). An Efficient Temporally-Constrained Probabilistic Model for Multiple-Instrument Music Transcription. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 701–707). 109 and 111.

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, *2*(1), 1–127. 41.

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., & Others. (2007). Greedy Layer-Wise Training of Deep Networks. In *Advances in Neural Information Processing Systems (NIPS)* (Vol. 19, p. 153). 38.

Berg-Kirkpatrick, T., Andreas, J., & Klein, D. (2014). Unsupervised Transcription of Piano Music. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 1538–1546). 110 and 113.

Berthelot, D., Schumm, T., & Metz, L. (2017). BEGAN: Boundary Equilibrium Generative Adversarial Networks. *arXiv preprint arXiv:1703.10717*. 45.

Bertin, N., Badeau, R., & Vincent, E. (2009). Fast Bayesian NMF Algorithms Enforcing Harmonicity and Temporal Continuity in Polyphonic Music Transcription. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 22.

Bertin, N., Badeau, R., & Vincent, E. (2010). Enforcing Harmonicity and Smoothness in Bayesian Non-Negative Matrix Factorization Applied to Polyphonic Music Transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(3), 538–549. 22.

Bittner, R. M., Mcfee, B., Salamon, J., Li, P., & Bello, J. P. (2017). Deep Salience Representations for F0 Estimation in Polyphonic Music. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 23–27). 18, 24, 51, 54, 88, and 109.

Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., & Bello, J. P. (2014). MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (Vol. 14, pp. 155–160). 50, 55, and 64.

Blaauw, M., & Bonada, J. (2017). A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs. *Applied Sciences*, *7*(12), 1313. 28.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. 39.

Boccardi, F., & Drioli, C. (2001). Sound Morphing With Gaussian Mixture Models. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (pp. 6–9). 72.

Böck, S., & Schedl, M. (2011). Enhanced Beat Tracking with Context-Aware Neural Networks. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*. 51.

Böck, S., & Schedl, M. (2012). Polyphonic Piano Note Transcription with Recurrent Neural Networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1–4). 25.

Boersma, P. (1993). Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratio of a Sampled Sound. In *Proceedings of the Institute of Phonetic Sciences* (Vol. 17, pp. 97–110). 20.

Bosch, J. J., & Gómez, E. (2014). Melody Extraction in Symphonic Classical Music: a Comparative Study of Mutual Agreement Between Humans and Algorithms. In *Proceedings of the Conference on Interdisciplinary Musicology*. 50.

Boulanger-Lewandowski, N., Bengio, Y., & Vincent, P. (2012). Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription. In *Proceedings of the International Conference on Machine Learning (ICML)*. 25, 90, and 95.

Bregman, A. S. (1990). *Auditory Scene Analysis*. MIT Press. 25.

Brown, G. J., & Cooke, M. (1994). *Computational Auditory Scene Analysis* (Vol. 8). 25.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... Amodei, D. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv preprint arXiv:1802.07228*. xi and 12.

Burraston, D., Edmonds, E., Livingstone, D., & Miranda, E. R. (2004). Cellular Automata in MIDI based Computer Music. In *Proceedings of the International Computer Music Conference (ICMC)* (Vol. 4, pp. 71–78). 29.

Caetano, M., & Rodet, X. (2010). Automatic Timbral Morphing of Musical Instrument Sounds by High-Level Descriptors. In *Proceedings of the International Computer Music Conference (ICMC)* (pp. 254–261). 72.

Caetano, M., & Rodet, X. (2013). Musical Instrument Sound Morphing Guided by Perceptually Motivated Features. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(8), 1666–1675. 72.

Camacho, A., & Harris, J. G. (2008). A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music. *The Journal of the Acoustical Society of America*, *124*(3), 1638–1652. 20 and 57.

Carreira-Perpiñán, M. A., & Hinton, G. E. (2005). On Contrastive Divergence Learning. *Artificial Intelligence and Statistics*, 17. 40.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-Based Music Information Retrieval: Current Directions and Future Challenges. *Proceedings of the IEEE*, *96*(4), 668–696. 15.

Celma, Ó. (2010). *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer. 10.

Cemgil, A. T., Kappen, B., & Barber, D. (2003). Generative Model Based Polyphonic Music Transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 23.

Cemgil, A. T., Kappen, B., Desain, P., & Honing, H. (2000). On Tempo Tracking: Tempogram Representation and Kalman filtering. *Journal of New Music Research*, *29*(1967), 259–273. 18.

Cemgil, A. T., Kappen, H. J., Member, S., Barber, D., Member, S., & Barber, D. (2006). A Generative Model for Music Transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(2), 679–694. 2 and 113.

Chemilier, M. (2001). Improvising Jazz Chord Sequences by Means of Formal Grammars.

Journee d'Informatique Musicale, 121–126. 29.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 2172–2180). 46.

Cheng, T., Mauch, M., Benetos, E., & Dixon, S. (2016). An Attack/Decay Model for Piano Transcription. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*. 19.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 34 and 115.

Cho, T., Weiss, R. J., & Bello, J. P. (2010). Exploring Common Variations in State-of-the-Art Chord Recognition Systems. *Sound and Music Computing*, 11–22. 19.

Choi, K., & Cho, K. (2019). Deep Unsupervised Drum Transcription. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*. 113.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2017). StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. *arXiv preprint arXiv:1711.09020*. 46.

Chollet, F. (2015). *Keras: The Python Deep Learning Library*. *https://keras.io/*. 55.

Chomsky, N. (1966). *Topics in the Theory of Generative Grammar*. Mouton. 29.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-Based Models for Speech Recognition. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 577–585). 35.

Cogliati, A., & Duan, Z. (2015). Piano Music Transcription Modeling Note Temporal Evolution. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 429–433). 19.

Cont, A. (2006). Realtime Multiple Pitch Observation using Sparse Non-Negative Constraints. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 206–211). 22.

Cook, P. R. (2002). *Real Sound Synthesis for Interactive Applications*. CRC Press. 11.

Costantini, G., Todisco, M., & Perfetti, R. (2013). NMF Based Dictionary Learning for Automatic Transcription of Polyphonic Piano Music. *WSEAS Transactions on Signal Processing*, *9*(3), 148–157. 22.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2017). Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*, 53–65. 92.

Dannenberg, R. B. (1985). An On-Line Algorithm for Real-Time Accompaniment. In *Proceedings of the International Computer Music Conference (ICMC)* (pp. 193–198). 10.

Dannenberg, R. B., & Hu, N. (2003). Polyphonic Audio Matching for Score Following and Intelligent Audio Editors. In *Proceedings of the International Computer Music Conference (ICMC)* (pp. 27–33). 10.

Daskalakis, C., Ilyas, A., Syrgkanis, V., & Zeng, H. (2018). Training GANs with Optimism. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 48.

Davy, M., Godsill, S., & Idier, J. (2006). Bayesian Analysis of Polyphonic Western Tonal Music. *The Journal of the Acoustical Society of America*, *119*(4), 2498–517. 23.

Davy, M., & Godsill, S. J. (2003). Bayesian Harmonic Models for Musical Signal Analysis. *Bayesian Statistics*, 1–16. 23.

de Cheveigné, A., & Kawahara, H. (2002). YIN, a Fundamental Frequency Estimator for Speech and Music. *The Journal of the Acoustical Society of America*, *111*(4), 1917–1930. 20 and 51.

Denton, E. L., Chintala, S., Szlam, A., Fergus, R., & Others. (2015). Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 1486–1494). 44.

Dessein, A., Cont, A., & Lemaitre, G. (2010). Real-Time Polyphonic Music Transcription With Non-Negative Matrix Factorization and Beta-Divergence. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 489–494). 22.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*. 132.

Ding, C., Li, T., & Peng, W. (2008). On the Equivalence Between Non-Negative Matrix Factorization and Probabilistic Latent Semantic Indexing. *Computational Statistics and Data Analysis*, *52*(8), 3913–3927. 23.

Dixon, S. (2000). On the Computer Recognition of Solo Piano Music. In *Proceedings of the Australasian Computer Music Conference* (pp. 31–37). 21.

Doersch, C. (2016). Tutorial on Variational Autoencoders. *arXiv preprint arXiv:1606.05908*. 42.

Donahue, C., Li, B., & Prabhavalkar, R. (2017). Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition. *arXiv preprint arXiv:1711.05747*. 27.

Donahue, C., McAuley, J., & Puckette, M. (2018). Synthesizing Audio with Generative Adversarial Networks. *arXiv preprint arXiv:1802.04208*. 28 and 44.

Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., & Yang, Y.-H. (2017). MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 34–41). 92.

Dosovitskiy, A., & Brox, T. (2016). Generating Images with Perceptual Similarity Metrics based on Deep Networks. *Advances in Neural Information Processing Systems (NIPS)*, *1*(c), 1–9. 89.

Downie, J. S., Hu, X., Lee, J. H., Choi, K., Cunningham, S. J., & Hao, Y. (2014). Ten Years of MIREX: Reflections, Challenges, and Opportunities. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 657–662). 21.

Dozat, T. (2016). Incorporating Nesterov Momentum into Adam. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 37.

Duan, Z., Pardo, B., & Zhang, C. (2010). Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(8), 2121–2133. 64.

Dubnowski, J. J., Schafer, R. W., & Rabiner, L. R. (1976). Real-Time Digital Hardware Pitch Detector. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *24*(1), 2–8. 20.

Dubois, C., & Davy, M. (2005). Harmonic Tracking Using Sequential Monte Carlo. In *IEEE/SP Workshop on Statistical Signal Processing* (pp. 1292–1297). 113.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*(Jul), 2121–2159. 36.

Durrieu, J. L., David, B., & Richard, G. (2011). A Musically Motivated Mid-Level Representation for Pitch Estimation and Musical Audio Source Separation. *IEEE Journal on Selected Topics in Signal Processing*, *5*(6), 1180–1191. 26.

Emiya, V., Badeau, R., & David, B. (2010). Multipitch Estimation of Piano Sounds Using a New Probabilistic Spectral Smoothness Principle. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(6), 1643–1654. 19 and 79.

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., & Roberts, A. (2019).

GANSynth: Adversarial Neural Audio Synthesis. *arxiv preprint arXiv:1902.08710*. 28 and 92.

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Eck, D., Simonyan, K., & Norouzi, M. (2017). Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. *arXiv preprint arXiv:1704.01279*. 64, 69, and 73.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., & Bengio, S. (2010). Why Does Unsupervised Pre-training Help Deep Learning? *Journal of Machine Learning Research*, *11*(Feb), 625–660. 38.

Esling, P., Chemla-Romeu-Santos, A., & Bitton, A. (2018). Bridging Audio Analysis, Perception and Synthesis with Perceptually-Regularized Variational Timbre Spaces. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*. 73.

Everitt, B. S., & Hand, D. J. (1981). *Finite Mixture Distributions*. Chapman and Hall. 39.

Ewert, S., Pardo, B., Müller, M., & Plumbley, M. D. (2014). Score-Informed Source Separation for Musical Audio Recordings: An Overview. *IEEE Signal Processing Magazine*, 116–124. 26.

Ewert, S., & Sandler, M. (2016). Piano Transcription in the Studio Using an Extensible Alternating Directions Framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(11), 1983–1997. 22.

Ewert, S., & Sandler, M. B. (2017). An Augmented Lagrangian Method for Piano Transcription Using Equal Loudness Thresholding and LSTM-Based Decoding. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 146–150). 22.

Ezzat, T., Meyers, E., Glass, J. R., & Poggio, T. (2005). Morphing Spectral Envelopes Using Audio Flow. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)* (pp. 2545–2548). 72.

Fan, Z.-C., Lai, Y.-L., & Jang, J.-S. R. (2017). SVSGAN: Singing Voice Separation via Generative Adversarial Network. *arXiv preprint arXiv:1710.11428*. 27.

Farbood, M., & Schoner, B. (2001). Analysis and Synthesis of Palestrina-Style Counterpoint Using Markov Chains. In *Proceedings of the International Computer Music Conference (ICMC)*. 29.

Fedus, W., Rosca, M., Lakshminarayanan, B., Dai, A. M., Mohamed, S., & Goodfellow, I. (2018). Many Paths to Equilibrium: GANs Do Not Need to Decrease a Divergence At Every Step. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 48.

Fernández, J. D., & Vico, F. (2013). AI Methods in Algorithmic Composition: A Comprehensive Survey. *Journal of Artificial Intelligence Research, 48,* 513–582. 11 and 29.

Fuentes, B., Badeau, R., & Richard, G. (2013). Harmonic Adaptive Latent Component Analysis of Audio and Application to Music Transcription. *IEEE Transactions on Audio, Speech, and Language Processing*. 22.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A Neural Algorithm of Artistic Style. *arXiv preprint arXiv:1508.06576*. 26 and 34.

Gaussier, E., & Goutte, C. (2005). Relation between PLSA and NMF and Implications. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 601–602). 23.

Ghias, A., Logan, J., Chamberlin, D., & Smith, B. C. (1995). Query By Humming: Musical Information Retrieval in An Audio Database. *Proceedings of the ACM International Conference on Multimedia, 6,* 110–113. 10.

Glorot, X., & Bengio, Y. (2010). Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*. 38.

Goertzel, B., & Pennachin, C. (2007). *Artificial General Intelligence*. Springer. 31.

Goodfellow, I. (2016). NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv preprint arXiv:1701.00160*. 92 and 96.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 2672–2680). xi, 12, 43, 92, 96, and 98.

Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2002). RWC Music Database: Popular, Classical and Jazz Music Databases. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 287–288). 56.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research, 62,* 729–754. 31.

Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., & Wierstra, D. (2015). DRAW: A Recurrent Neural Network For Image Generation. In *Proceedings of the International Conference on Machine Learning (ICML)*. 40.

Gregor, K., Danihelka, I., Mnih, A., Blundell, C., & Wierstra, D. (2014). Deep AutoRegressive Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*. 40.

Grey, J. M. (1977). Multidimensional Perceptual Scaling of Musical Timbres. *The Journal of the Acoustical Society of America, 61*(5), 1270–7. 73.

Griffin, D., & Lim, J. S. (1984). Signal Estimation from Modified Short-Time Fourier transform. *IEEE Transactions on Audio, Speech, and Language Processing, 32*(2), 236–243. 17.

Grindlay, G., & Ellis, D. P. W. (2009). Multi-Voice Polyphonic Music Transcription Using Eigeninstruments. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 53–56). 23, 109, and 111.

Grindlay, G., & Ellis, D. P. W. (2010). A Probabilistic Subspace Model for Multi-Instrument Polyphonic Transcription. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 21–26). 23.

Grosche, P., Muller, M., & Kurth, F. (2010). Cyclic Tempogram - a Mid-Level Tempo Representation for Music Signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 18.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NIPS)* (Vol. 30, pp. 5769–5779). 45.

Hamanaka, M., Hirata, K., & Tojo, S. (2013). Computational Music Theory and Its Applications to Expressive Performance and Composition. In *Guide to Computing for Expressive Music Performance* (pp. 205–234). 29.

Harte, C. A., & Sandler, M. B. (2005). Automatic Chord Identifcation using a Quantised Chromagram. In *Proceedings of the AES Convention* (pp. 1–21). 18.

Harte, C. A., Sandler, M. B., & Gasser, M. (2006). Detecting Harmonic Change in Musical Audio. In *Proceedings of the ACM workshop on Audio and Music Computing Multimedia* (p. 21). 18.

Hartmann, W. M. (1997). *Signals, Sound, and Sensation*. Springer. 20.

Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., ... Eck, D. (2018). Onsets and Frames: Dual-Objective Piano Transcription. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*. 25, 75, 88, 89, 90, 91, 98, 99, 100, 110, 114, and 117.

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-z. A., Dieleman, S., ... Brain, G. (2019). Enabling Factorized Piano Music Modeling and Generation with the MAESTRO dataset. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 74, 85, 98, 100, and 106.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing

Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1026–1034). 37 and 38.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). 34.

Heittola, T., Klapuri, A., & Virtanen, T. (2009). Musical Instrument Recognition in Polyphonic Audio Using Source-Filter Model for Sound Separation. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 327–332). 26 and 111.

Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016). Deep Clustering: Discriminative Embeddings for Segmentation and Separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 31–35). IEEE. 112.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems (NIPS)*. 47 and 96.

Hinton, G. (2012). *rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude*. 36.

Hinton, G., Osindero, S., & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation, 18*(7), 1527–1554. 24 and 40.

Hinton, G. E., Sejnowski, T. J., & Ackley, D. H. (1984). *Boltzmann Machines: Constraint Satisfaction Networks that Learn* (Vol. 9; Tech. Rep. No. CMS-CS-84-119). 39.

Hjelm, R. D., Jacob, A. P., Che, T., Cho, K., & Bengio, Y. (2018). Boundary-Seeking Generative Adversarial Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 47.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*(8), 1735–1780. 34.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)* (pp. 289–296). 22.

Hsu, C.-L., & Jang, J. S. R. (2010). On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(2), 310–319. 64.

Hua, K. (2018). Do WaveNets Dream of Acoustic Waves? *arXiv preprint arXiv:1802.08370*. 27.

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Hawthorne, C., Dai, A. M., … Eck, D. (2019). Music Transformer. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 29 and 132.

Huang, P. S., Chen, S. D., Smaragdis, P., & Hasegawa-Johnson, M. (2012). Singing-Voice Separation from Monaural Recording using Robust Principal Component Analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 57–60). 26.

Humphrey, E. J., & Bello, J. P. (2012). Rethinking Automatic Chord Recognition with Convolutional Neural Networks. In *Proceedings of the International Conference on Machine Learning and Applications and Workshops (ICMLA)* (Vol. 2, pp. 357–362). 51.

Humphrey, E. J., Glennon, A. P., & Bello, J. P. (2011). Non-Linear Semantic Embedding for Organizing Large Instrument Sample Libraries. In *Proceedings of the International Conference on Machine Learning and Applications and Workshops (ICMLA)* (Vol. 2, pp. 142–147). 73.

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 448–456). 37 and 55.

Isik, Y., Le Roux, J., Chen, Z., Watanabe, S., & Hershey, J. R. (2016). Single-Channel Multi-Speaker Separation Using Deep Clustering. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)* (pp. 545–549). 112.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1125–1134). 46, 93, 94, 96, and 99.

Itoyama, K., Goto, M., Komatani, K., Ogata, T., & Okuno, H. G. (2011). Simultaneous Processing of Sound Source Separation and Musical Instrument Identification using Bayesian Spectral Modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3816–3819). IEEE. 111.

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., & Weyde, T. (2017). Singing Voice Separation With Deep U-Net Convolutional Networks. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 745–751). 27.

Jin, X., Xu, C., Feng, J., Wei, Y., Xiong, J., & Yan, S. (2015). Deep Learning with S-shaped Rectified Linear Activation Units. *arXiv preprint arXiv:1512.07030*. 37.

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., … Kavukcuoglu, K. (2018). Efficient Neural Audio Synthesis. *arXiv preprint*

arXiv:1802.08435. 28 and 40.

Kalingeri, V., & Grandhe, S. (2016). Music Generation with Deep Learning. *arXiv preprint arXiv:1612.04928*. 27.

Karpathy, A., & Fei-Fei, L. (2017). Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 35.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proceedings of the International Conference on Learning Representations (ICLR)*. xi, 12, and 44.

Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4401–4410). xi, 12, 92, and 133.

Kassler, M. (1966). Toward Musical Information Retrieval. *Perspectives of New Music*, *4*(2), 59. 15.

Kelz, R., Dorfer, M., Korzeniowski, F., Böck, S., Arzt, A., & Widmer, G. (2016). On the Potential of Simple Framewise Approaches to Piano Transcription. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*. xii, 24, 88, and 91.

Kelz, R., & Widmer, G. (2017). An Experimental Analysis of the Entanglement Problem in Neural-Network-based Music Transcription Systems. In *Proceedings of the AES Conference on Semantic Audio*. 25.

Khadkevich, M., & Omologo, M. (2009). Use of Hidden Markov Models and Factored Language Models for Automatic Chord Recognition. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 561–566). 19.

Kim, J. W., & Bello, J. P. (2019). Adversarial Learning for Improved Onsets and Frames Music Transcription. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*. 88.

Kim, J. W., Bittner, R., Kumar, A., & Bello, J. P. (2019). Neural Music Synthesis for Flexible Timbre Control. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 176–180). 71, 82, and 106.

Kim, J. W., Salamon, J., Li, P., & Bello, J. P. (2018). CREPE: A Convolutional Representation for Pitch Estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 50 and 88.

Kim, T., Cha, M., Kim, H., Lee, J. K., & Kim, J. (2017). Learning to Discover Cross-Domain

Relations with Generative Adversarial Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*. 46.

Kindermann, R., & Snell, J. L. (1980). *Markov Random Fields and Their Applications* (Vol. 16) (No. 4). American Mathematical Society. 39.

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)* (pp. 1–15). 36, 55, 99, and 117.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., & Welling, M. (2016). Improving Variational Inference with Inverse Autoregressive Flow. In *Advances in Neural Information Processing Systems (NIPS)*. 42.

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 41 and 73.

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-Normalizing Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*. 38.

Klapuri, A. P. (2003). Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness. *IEEE Transactions on Speech and Audio Processing*, *11*(6), 804–816. 19.

Klapuri, A. P., & Davy, M. (2006). *Signal Processing Methods for Music Transcription*. Springer. 2.

Kleene, S. C. (1951). *Representation of Events in Nerve Nets and Finite Automata*. 31.

Kodali, N., Abernethy, J., Hays, J., & Kira, Z. (2017). On Convergence and Stability of GANs. *arXiv preprint arXiv:1705.07215*. 48.

Koushik, J., & Hayashi, H. (2016). Improving Stochastic Gradient Descent with Feedback. *arXiv preprint arXiv:1611.01505*. 36.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 1097–1105). 34.

Larochelle, H., & Murray, I. (2011). The Neural Autoregressive Distribution Estimator. In *Proceedings of the International Conference on Machine Learning (ICML)* (Vol. 15, pp. 29–37). 40 and 90.

Le Roux, J., Kameoka, H., Ono, N., & Sagayama, S. (2010). Fast Signal Reconstruction Frommagnitude Stft Spectrogram Based on Spectrogram Consistency. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*. 17.

LeCun, Y. (2016). *Unsupervised Learning*. NYU. 11.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, *521*(7553), 436–444. 24 and 91.

LeCun, Y., Jackel, L. D., Bottou, L., Brunot, A., Cortes, C., Denker, J. S., ... Vapnik, V. (1995). Comparison of Learning Algorithms for Handwritten Digit Recognition. In *Proceedings of the International Conference on Artificial Neural Networks* (Vol. 60, pp. 53–60). 33.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., ... Shi, W. (2017). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 85.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. 22.

Lee, D. D., & Seung, H. S. (2001). Algorithms for Non-Negative Matrix Factorization. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 556–562). 22 and 90.

Leglaive, S., Badeau, R., & Richard, G. (2016). Multichannel Audio Source Separation with Probabilistic Reverberation Priors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(12), 2453–2465. 26.

Lerdahl, F., & Jackendoff, R. (1983). *Generative Theory of Tonal Music*. MIT Press. 29.

Li, Q., Myaeng, S. H., & Kim, B. M. (2007). A Probabilistic Music Recommender Considering User Opinions and Audio Features. *Information Processing and Management*, *43*(2), 473–487. 10.

Li, S. (2017). Context-Independent Polyphonic Piano Onset Transcription with an Infinite Training Dataset. *arXiv preprint arXiv:1707.08438*. 113.

Li, Y., & Wang, D. (2007). Separation of Singing Voice From Music Accompaniment for Monaural Recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(4), 1475–1487. 26.

Linnainmaa, S. (1970). The Representation of the Cumulative Rounding Error of an Algorithm as a Taylor Expansion of the Local Rounding Errors. *Master's Thesis (in Finnish), University of Helsinki*, 6–7. 36.

Liu, M.-Y., & Tuzel, O. (2016). Coupled Generative Adversarial Networks. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 469–477). xi and 12.

Liutkus, A., Rafii, Z., Badeau, R., Pardo, B., & Richard, G. (2012). Adaptive Filtering for Music/Voice Separation Exploiting the Repeating Musical Structure. In *Proceedings of*

the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 53–56). 26.

Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*. 18, 64, and 73.

Long, J., Shelhamer, E., & Darrell, T. (2018). Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 6810–6818. 95.

Lostanlen, V., & Cella, C.-E. (2016). Deep Convolutional Networks on the Pitch Spiral for Music Information Recognition. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 612–618). 109.

Lu, C. C., & Tseng, V. S. (2009). A Novel Method for Personalized Music Recommendation. *Expert Systems with Applications*, *36*(6), 10035–10044. 10.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., & Bousquet, O. (2017). Are GANs Created Equal? A Large-Scale Study. *arXiv preprint arXiv:1711.10337*. 45.

Luo, Y., Chen, Z., Hershey, J. R., Le Roux, J., & Mesgarani, N. (2017). Deep Clustering and Conventional Networks for Music Separation: Stronger Together. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 61–65). 112.

Magno, T., & Sable, C. (2008). A Comparison of Signal-Based Music Recommendation To Genre Labels, Collaborative Filtering, Musicological Analysis, Human Recommendation, and Random Baseline. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 161–166). 10.

Mahendran, A., & Vedaldi, A. (2016). Visualizing Deep Convolutional Neural Networks using Natural Pre-Images. *International Journal of Computer Vision*, *120*(3), 233–255. 34.

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., & Smolley, S. P. (2017). Least Squares Generative Adversarial Networks. *Proceedings of the International Conference on Computer Vision (ICCV)*, 2794–2802. 45 and 92.

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., & Smolley, S. P. (2018). On the Effectiveness of Least Squares Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 45.

Marolt, M. (1999). A Comparison of Feed Forward Neural Network Architectures for Piano Music Transcription. In *Proceedings of the International Computer Music Conference (ICMC)*. 24.

Marolt, M. (2004). A Connectionist Approach to Automatic Transcription of Polyphonic Piano Music. *IEEE Transactions on Multimedia*, *6*(3), 439–449. 24.

Maron, M. E. (1961). Automatic Indexing: An Experimental Inquiry. *Journal of the ACM*, *8*(3), 404–417. 39.

Martin, K. D. (1996). Automatic Transcription of Simple Polyphonic Music. *The Journal of the Acoustical Society of America*, *100*(399), 2813. 21.

Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., ... Dixon, S. (2015). Computer-Aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency. In *Proceedings of the First International Conference on Technologies for Music Notation and Representation* (p. 8). 50.

Mauch, M., & Dixon, S. (2014). pYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 659–663). 20, 51, 56, 57, 64, 66, and 89.

Mauch, M., & Ewert, S. (2013). The Audio Degradation Toolbox and Its Application To Robustness Evaluation. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*. 57.

McFee, B., Humphrey, E. J., & Bello, J. P. (2015). A Software Framework for Musical Data Augmentation. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 248–254). 69, 126, and 131.

McFee, B., Kim, J. W., Cartwright, M., Salamon, J., Bittner, R. M., & Bello, J. P. (2019). Open-Source Practices for Music Signal Processing Research: Recommendations for Transparent, Sustainable, and Reproducible Audio Research. *IEEE Signal Processing Magazine*, *36*(1), 128–137. 63.

Mcfee, B., Raffel, C., Liang, D., Ellis, D. P. W., Mcvicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the Python in Science Conference* (pp. 18–25). 117.

Mcleod, A., & Steedman, M. (2018). Evaluating Automatic Polyphonic Music Transcription. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 42–49). 25.

Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., ... Bengio, Y. (2017). SampleRNN: An Unconditional End-to-End Neural Audio Generation Model. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 27 and 40.

Mescheder, L., Nowozin, S., & Geiger, A. (2017). The Numerics of GANs. In *Advances in*

Neural Information Processing Systems (NIPS). 48.

Miranda, E. R., Al, J., & Eds, B. (2007). *Evolutionary Computer Music*. 29.

Miron, M. (2018). Source Separation Methods for Orchestral Music: Timbre-Informed and Score-Informed Strategies (Doctoral dissertation, Pompeu Fabra University, Barcelona).
26.

Mirza, M., & Osindero, S. (2014). Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*. 46 and 93.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral Normalization for Generative Adversarial Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 45.

Miyato, T., & Koyama, M. (2018). cGANs with Projection Discriminator. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 46.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. *arXiv preprint arXiv:1312.5602*. 35.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Ostrovski, G. (2015). Human-Level Control Through Deep Reinforcement Learning. *Nature*, *518*(7540), 529–533. 35.

Mohamed, S., & Lakshminarayanan, B. (2017). Learning in Implicit Generative Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 48.

Moorer, J. A. (1977). On the Transcription of Musical Sound by Computer. *Computer Music Journal*, *1*(4), 32–38. 2 and 21.

Mor, N., Wolf, L., & Polyak, A. (2019). A Universal Music Translation Network. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 74.

Nagarajan, V., & Kolter, J. Z. (2017). Gradient Descent GAN Optimization is Locally Stable. In *Advances in Neural Information Processing Systems (NIPS)*. 48.

Nair, V., & Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 807–814). 37.

Nam, J., Ngiam, J., Lee, H., & Slaney, M. (2011). A Classification-Based Polyphonic Piano Transcription Approach Using Learned Feature Representations. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 175–180). 24.

Nayebi, A., & Vitelli, M. (2015). GRUV: Algorithmic Music Generation using Recurrent Neural Networks. *Stanford CS224d Class Project*. 27.

Nesterov, Y. (1983). *A Method for Solving the Convex Programming Problem with Convergence Rate O(1/k⊠)* (Vol. 269). 37.

Ng, A. (2011). Sparse Autoencoder. *Stanford CS294a Lecture notes*. 41.

Ni, Y., McVicar, M., Santos-Rodriguez, R., & De Bie, T. (2012). An End-to-End Machine Learning System for Harmonic Analysis of Music. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(6), 1771–1783. 89.

Niedermayer, B. (2008). Non-Negative Matrix Division For The Automatic Transcription Of Polyphonic Music. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 544–549). 22.

Noll, A. M. (1967). Cepstrum Pitch Determination. *The Journal of the Acoustical Society of America*, *41*(2), 293–309. 20.

Nowozin, S., & Cseke, B. (2016). f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In *Advances in Neural Information Processing Systems (NIPS)*. 45.

Odena, A., Olah, C., & Shlens, J. (2016). Conditional Image Synthesis With Auxiliary Classifier GANs. *arXiv preprint arXiv:1610.09585*. 46.

Ono, N., Miyamoto, K., Kameoka, H., & Roux, J. L. (2010). Harmonic and Percussive Sound Separation and Its Application to MIR-Related Tasks. In *Advances in Music Information Retrieval* (pp. 213–236). Springer. 26.

Oramas, S., Nieto, O., Barbieri, F., & Serra, X. (2017). Multi-Label Music Genre Classification from Audio, Text, and Images Using Deep Features. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 23–30). 18.

Orio, N., Lemouton, S., Schwarz, D., & Schnell, N. (2003). Score Following: State of the Art and New Developments. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (pp. 36–41). 10.

Osaka, N. (1995). Timbre Interpolation of Sounds using a Sinusoidal Model. In *Proceedings of the International Computer Music Conference (ICMC)* (pp. 408–411). 72.

Oudre, L., Grenier, Y., & Févotte, C. (2009). Template-Based Chord Recognition: Influence of the Chord Types. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 153–158). 19.

Ozerov, A., Philippe, P., Bimbot, F., & Gribonval, R. (2007). Adaptation of Bayesian Models

for Single-Channel Source Separation and its Application to Voice/Music Separation in Popular Songs. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(5), 1564–1578. 26.

Pascual, S., Bonafonte, A., & Serrà, J. (2017). SEGAN: Speech Enhancement Generative Adversarial Network. *arXiv preprint arXiv:1703.09452*. 27 and 47.

Patel, A. D. (2008). *Music, Language, and the Brain*. Oxford university press. 29.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Vol. 8) (No. 1). 39.

Peeling, P. H., Cemgil, A. T., & Godsill, S. J. (2010). Generative Spectrogram Factorization Models for Polyphonic Piano Transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(3), 519–527. 22.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The Timbre Toolbox: Extracting Audio Descriptors from Musical Signals. *The Journal of the Acoustical Society of America*, *130*(5), 2902–2916. 73.

Pejrolo, A., & Metcalfe, S. B. (2017). *Creating sounds from scratch: A Practical guide to music synthesis for producers and composers*. Oxford University Press. 72.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2017). FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*. 75 and 115.

Pesek, M., Leonardis, A., & Marolt, M. (2017). Robust Real-Time Music Transcription with a Compositional Hierarchical Model. *PLoS ONE*, *12*(1). 23.

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., ... Miller, J. (2018). Deep Voice 3: 2000-Speaker Neural Text-to-Speech. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 73 and 75.

Piszczalski, M., & Galler, B. A. (1977). Automatic Music Transcription. *Computer Music Journal*, *1*(4), 24–31. 2 and 21.

Plumbley, M. D., Abdallah, S. A., Bello, J. P., Davies, M. E., Monti, G., & Sandler, M. B. (2002). Automatic Music Transcription and Audio Source Separation. *Cybernetics and Systems: An International Journal*, *6*, 603–627. 26.

Poliner, G. E., & Ellis, D. P. (2006). A Discriminative Model for Polyphonic Piano Transcription. *EURASIP Journal on Advances in Signal Processing*, 1–16. 24, 89, and 101.

Polyak, B. T. (1964). Some Methods of Speeding Up the Convergence of Iteration Methods. *USSR Computational Mathematics and Mathematical Physics*, *4*(5), 1–17. 36.

Pons, J., Nieto, O., Prockup, M., Schmidt, E. M., Ehmann, A. F., & Serra, X. (2018). End-to-End Learning for Music Audio Tagging at Scale. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*. 19.

Prenger, R., Valle, R., & Catanzaro, B. (2019). WaveGlow : A Flow-Based Generative Network for Speech Synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3617–3621). 28 and 85.

Rabiner, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, *77*(2), 257–286. 39.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv preprint arXiv:1511.06434v2*, 1–16. xi, 12, and 44.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *OpenAI Blog*. 132 and 133.

Raffel, C., Mcfee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., & Ellis, D. P. W. (2014). mir_eval: A Transparent Implementation of Common MIR Metrics. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 367–372). 56 and 101.

Rafii, Z., & Pardo, B. (2013). REpeating pattern extraction technique (REPET): A simple method for music/voice separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(1), 71–82. 26.

Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marcal, A. R., Guedes, C., & Cardoso, J. S. (2012). Optical Music Recognition State-of-the-Art and Open Issues. *International Journal of Multimedia Information Retrieval*, *1*(3), 173–190. 29.

Reddi, S. J., Kale, S., & Kumar, S. (2018). On the Convergence of Adam and Beyond. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 37.

Rezende, D. J., & Mohamed, S. (2015). Variational Inference with Normalizing Flows. In *Proceedings of the International Conference on Machine Learning (ICML)* (Vol. 37). 42.

Riedmiller, M. (1994). Advanced Supervised Learning in Multi-layer Perceptrons - From Backpropagation to Adaptive Learning Algorithms. *Computer Standards and Interfaces*, *16*, 265–278. 33.

Rifai, S., Vincent, P., Muller, X., Glorot, X., & Bengio, Y. (2011). Contractive Auto-Encoders: Explicit Invariance During Feature Extraction. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 833–840). 41.

Robbins, H., & Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, *22*(3), 400–407. 36.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (pp. 234–241). Springer International Publishing. 26.

Rosenblatt, F. (1957). *The Perceptron, a Perceiving and Recognizing Automaton (Project Para)*. Cornell Aeronautical Laboratory. 32.

Ross, M. J., Shaffer, H. L., Cohen, A., Freudberg, R., & Manley, H. J. (1974). Average Magnitude Difference Function Pitch Extractor. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *ASSP-22*(5), 353–362. 20.

Rowe, R. (2003). *Machine Musicianship*. 13.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Chapter 8. Learning Internal Representations. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* (pp. 354–361). 34.

Rumelhart, D. E., Hinton, G. E., & Williams., R. J. (1986). Learning Representations by Back-Propagating Errors. *Nature, 323*(6088), 533–538. 36.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision, 115*(3), 211–252. 46 and 47.

Russell, S. J., & Norvig, P. (2009). *Artificial Intelligence: a Modern Approach (3rd Edition)*. Pearson. 13.

Saito, S., Kameoka, H., Takahashi, K., Nishimoto, T., & Sagayama, S. (2008). Specmurt Analysis of Polyphonic Music Signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(3), 639–650. 18.

Sakaue, D., Itoyama, K., Ogata, T., & Okuno, H. (2013). Robust Multipitch Analyzer against Initialization based on Latent Harmonic Allocation using Overtone Corpus. *Journal of Information Processing*, *21*(2), 246–255. 23.

Salamon, J., Bittner, R. M., Bonada, J., Bosch, J. J., Gomez, E., & juan pablo Bello. (2017). An Analysis/Synthesis Framework for Automatic F0 Annotation of Multitrack Datasets. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 71–78). 51, 56, and 64.

Salamon, J., Gómez, E., Ellis, D. P. W., & Richard, G. (2014). Melody Extraction from Polyphonic Music Signals: Approaches, Applications, and Challenges. *IEEE Signal Processing Magazine, 31*(2), 118–134. 56.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016).

Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 2226–2234). 44 and 47.

Saruwatari, H., Kawamura, T., Nishikawa, T., Lee, A., & Shikano, K. (2006). Blind Source Separation Based on a Fast-Convergence Algorithm Combining ICA and Beamforming. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(2), 666–678. 26.

Schörkhuber, C., & Klapuri, A. P. (2010). Constant-Q Transform Toolbox for Music Processing. In *Proceedings of the Sound and Music Computing (SMC) Conference* (pp. 3–64). 18.

Shelhamer, E., Long, J., & Darrell, T. (2017). Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 640–651. 34.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... Wu, Y. (2018). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2–6). 27, 73, 75, and 106.

Shepard, R. N. (1964). Circularity in Judgments of Relative Pitch. *The Journal of the Acoustical Society of America*, *36*(12), 2346–2353. 9.

Sigtia, S., Benetos, E., Boulanger-Lewandowski, N., Weyde, T., d'Avila Garcez, A. S., & Dixon, S. (2015). A Hybrid Recurrent Neural Network For Music Transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 25.

Sigtia, S., Benetos, E., Cherla, S., Weyde, T., d'Avila Garcez, a., & Dixon, S. (2014). An RNN-based Music Language Model for Improving Automatic Music Transcription. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 53–58). 29.

Sigtia, S., Benetos, E., & DIxon, S. (2016). An End-to-End Neural Network for Polyphonic Piano Music Transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(5), 927–939. 25, 89, and 95.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... Others (2016). Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, *529*(7587), 484–489. 35.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*. 34.

Slaney, M., Covell, M., & Lassiter, B. (1996). Automatic Audio Morphing. In *Proceedings of*

the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1001–1004). 72.

Smaragdis, P. (2009). Relative-Pitch Tracking of Multiple Arbitrary Sounds. *The Journal of the Acoustical Society of America, 125*(5), 3406–13. 23.

Smaragdis, P., & Brown, J. C. (2003). Non-Negative Matrix Factorization for Polyphonic Music Transcription. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 177–180). 22.

Smaragdis, P., Raj, B., & Shashanka, M. (2006). A Probabilistic Latent Variable Model for Acoustic Modeling. In *Advances in Neural Information Processing Systems (NIPS)*. 23.

Smilkov, D., Thorat, N., Assogba, Y., Yuan, A., Kreeger, N., Yu, P., ... Wattenberg, M. (2019). TensorFlow.js: Machine Learning for the Web and Beyond. In *Proceedings of the SysML Conference*. 66.

Smolensky, P. (1986). Chapter 6. Information Processing in Dynamical Systems: Foundations of Harmony Theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* (pp. 194–281). 39.

Sønderby, C. K., Caballero, J., Theis, L., Shi, W., & Huszár, F. (2017). Amortised MAP Inference for Image Super-resolution. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 47 and 48.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research, 15*(1), 1929–1958. 37, 55, and 96.

Stoller, D., Ewert, S., & Dixon, S. (2017). Adversarial Semi-Supervised Audio Source Separation applied to Singing Voice Extraction. *arXiv preprint arXiv:1711.00048*. 27.

Sturm, B. L. (2013). Classification Accuracy is Not Enough: On the Evaluation of Music Genre Recognition Systems. *Journal of Intelligent Information Systems, 41*(3), 371–406. 56 and 126.

Subakan, C., & Smaragdis, P. (2017). Generative Adversarial Source Separation. *arXiv preprint arXiv:1710.10779*. 27.

Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-To-End Memory Networks. In *Advances in Neural Information Processing Systems (NIPS)*. 35.

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the Importance of Initialization and Momentum in Deep Learning. *30th International Conference on Machine Learning, ICML 2013*, 2176–2184. 36.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with

neural networks. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 3104–3112). 34.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction (2nd Edition)*. MIT press Cambridge. 35.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–9). 34.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 47.

Talkin, D., Kleijn, W. B., & Paliwal, K. K. (1995). A Robust Algorithm for Pitch Tracking (RAPT). In *Speech Coding and Synthesis* (pp. 495–518). Elsevier Science. 20.

Teng, Y., Zhao, A., & Goudeseune, C. (2017). Generating Nontrivial Melodies for Music as a Service. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*. 29.

Theis, L., & Bethge, M. (2015). Generative Image Modeling Using Spatial LSTMs. In *Advances in Neural Information Processing Systems (NIPS)*. 40.

Theis, L., van den Oord, A., & Bethge, M. (2016). A Note on the Evaluation of Generative Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 47.

Thickstun, J., Harchaoui, Z., Foster, D., & Kakade, S. M. (2018). Invariances and Data Augmentation for Supervised Music Transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 25 and 129.

Thickstun, J., Harchaoui, Z., & Kakade, S. (2017). Learning Features of Music from Scratch. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 78, 117, and 125.

Thomé, C., & Ahlbäck, S. (2017). Polyphonic Pitch Detection With Convolutional Recurrent Neural Networks. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*. 25.

Tikhonov, A., & Yamshchikov, I. P. (2017). Music Generation with Variational Recurrent Autoencoder Supported by History. In *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR)*. 29.

Tolonen, T., & Karjalainen, M. (2000). A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio Processing*, *8*(6), 708–716. 18.

Ullrich, K., & Van Der Wel, E. (2017). Music Transcription With Convolutional Sequence-To-Sequence Models. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*. 25.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. *arXiv preprint arXiv:1609.03499*. 27, 40, 73, 76, and 114.

van den Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel Recurrent Neural Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*. 40 and 90.

van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., & Kavukcuoglu, K. (2016). Conditional Image Generation with PixelCNN Decoders. *Advances in Neural Information Processing Systems (NIPS)*. 90.

van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., ... Hassabis, D. (2018). Parallel WaveNet: Fast High-Fidelity Speech Synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*. 85.

van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems (NIPS)*. 42 and 132.

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(2008), 2579–2605. 83.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is All You Need. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 5999–6009). 132.

Vercoe, B. (1984). The Synthetic Performer in the Context of Live Performance. In *Proceedings of the International Computer Music Conference (ICMC)* (pp. 199–200). 10.

Vincent, E., Bertin, N., & Badeau, R. (2010). Adaptive Harmonic Spectral Decomposition for Multiple Pitch Estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(3), 528–537. 22.

Vincent, E., Bertin, N., Gribonval, R., & Bimbot, F. (2014). From Blind to Guided Audio Source Separation: How Models and Side Information Can Improve the Separation of Sound. *IEEE Signal Processing Magazine*, *31*(3), 107–115. 26.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and Composing Robust Features with Denoising Autoencoders. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 1096–1103). 41.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., ... Silver, D. (2019). Grandmaster Level in StarCraft II using Multi-Agent Reinforcement

Learning. *Nature, 575*(7782), 350–354. 1.

Viterbi, A. J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269. 66.

Vogl, R., Dorfer, M., Widmer, G., & Knees, P. (2017). Drum Transcription via Joint Beat and Drum Modeling using Convolutional Recurrent Neural Networks. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference* (pp. 150–157). 18.

von dem Knesebeck, A., & Zölzer, U. (2010). Comparison of Pitch Trackers for Real-Time Guitar Effects. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*. 21.

Wang, Q., Zhou, R., & Yan, Y. (2018). Polyphonic Piano Transcription with a Note-Based Music Language Model. *Applied Sciences*, *8*(3), 470. 25.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., … Others (2017). Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model. *arXiv preprint arXiv:1703.10135*, *2017-Augus*, 4006–4010. 27.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. 47.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multi-Scale Structural Similarity for Image Quality Assessment. In *Proceedings of the IEEE Asilomar Conference on Signals, Systems and Computers* (Vol. 2, pp. 9–13). 47.

Wang, Z. Q., Roux, J. L., & Hershey, J. R. (2018). Alternative Objective Functions for Deep Clustering. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 686–690). IEEE. 112.

Weninger, F., Kirst, C., & Bungartz, H.-j. (2013). A Discriminative Approach to Polyphonic Piano Note Transcription Using Supervised Non-Negative Matrix Factorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6–10). 24.

Werbos, P. J. (1982). Applications of Advances in Nonlinear Sensitivity Analysis. In *System Modeling and Optimization* (pp. 762–770). Springer. 36.

Wessel, D. L. (1979). Timbre Space as a Musical Control Structure. *Computer Music Journal*, 45–52. 73.

Wu, J., Zhang, C., Xue, T., Freeman, W. T., & Tenenbaum, J. B. (2016). Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. In *Advances in Neural Information Processing Systems (NIPS)*. 44.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*. 35.

Wu, Y.-T., Chen, B., & Su, L. (2019). Polyphonic Music Transcription with Semantic Segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 166–170). 111.

Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical Evaluation of Rectified Activations in Convolutional Network. *arXiv preprint arXiv:1505.00853*. 37.

Yang, L.-C., Chou, S.-Y., & Yang, Y.-H. (2017). MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation using 1D and 2D Conditions. *arXiv preprint arXiv:1703.10847*. 29 and 92.

Yeh, C., Roebel, A., & Rodet, X. (2010). Multiple Fundamental Frequency Estimation and Polyphony Inference of Polyphonic Music Signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(6), 1116–1126. 19.

Yoshii, K., & Goto, M. (2012). A Nonparametric Bayesian Multipitch Analyzer Based on Infinite Latent Harmonic Allocation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(3), 717–730. 23.

Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent Neural Network Regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 105.

Zeiler, M. D. (2012). ADADELTA: an Adaptive Learning Rate Method. *arXiv preprint arXiv:1212.5701*. 36.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond Empirical Risk Minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 96.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. (2017a). StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. *arXiv preprint arXiv:1710.10916*. 44 and 47.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. (2017b). StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 47.

Zhang, W., Chen, Z., Member, S., Yin, F., & Zhang, Q. (2018). Melody Extraction From Polyphonic Music Using Particle Filter and Dynamic Programming. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(9), 1620–1632. 109.

Zhao, J., Mathieu, M., & LeCun, Y. (2017). Energy-based Generative Adversarial Network.

In *Proceedings of the International Conference on Learning Representations (ICLR)*. 45.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2223–2232). 46.

Zubizarreta, M. L. (1998). *Prosody, Focus, and Word Order*. MIT Press. 51.