PENGENALAN EKSPRESI WAJAH PESERTA DIDIK DI RUANG KELAS MENGGUNAKAN *HYBRID* MOBILENETV3-VIT DENGAN *TOKEN DOWNSAMPLING*

SKRIPSI

Diajukan Untuk Memenuhi Sebagian dari Syarat Memperoleh Gelar Sarjana Komputer Program Studi Ilmu Komputer

Oleh
Mochamad Khaairi
2106416

PROGRAM STUDI ILMU KOMPUTER
FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PENDIDIKAN INDONESIA
2025

PENGENALAN EKSPRESI WAJAH PESERTA DIDIK DI RUANG KELAS MENGGUNAKAN *HYBRID* MOBILENETV3-VIT DENGAN *TOKEN DOWNSAMPLING*

Oleh
Mochamad Khaairi
2106416

Sebuah skripsi yang diajukan untuk memenuhi salah satu syarat memeroleh gelar Sarjana pada Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

MOCHAMAD KHAAIRI

PENGENALAN EKSPRESI WAJAH PESERTA DIDIK DI RUANG KELAS MENGGUNAKAN *HYBRID* MOBILENETV3-VIT DENGAN *TOKEN DOWNSAMPLING*

Disetujui dan disahkan oleh pembimbing:

Pembimbing I

Dr. Rasim, S.T., M.T.

NIP. 197407252006041002

Pembimbing II

Yaya Wihardi, S. Kom., M. Kom.

NIP. 198903252015041001

Mengetahui,

Ketua Program Studi Ilmu Komputer

Dr. Muhamad Nursalman, M.T.

NIP. 197909292006041002

PENGENALAN EKSPRESI WAJAH PESERTA DIDIK DI RUANG KELAS MENGGUNAKAN *HYBRID* MOBILENETV3-VIT DENGAN *TOKEN DOWNSAMPLING*

Oleh

Mochamad Khaairi

2106416

ABSTRAK

Dalam lingkungan kelas besar, pengajar sering mengalami kesulitan dalam memantau secara menyeluruh ekspresi wajah setiap peserta didik selama proses pembelajaran. Padahal, ekspresi wajah mencerminkan kondisi emosional dan tingkat partisipasi peserta didik. Penelitian ini bertujuan membangun dan mengevaluasi sistem pengenalan ekspresi wajah yang tangguh pada kondisi nyata. Diusulkan model berbasis arsitektur *hybrid* yang menggabungkan MobileNetV3 untuk ekstraksi fitur lokal dan *Vision Transformer* untuk pemodelan konteks global, serta dilengkapi *Token Downsampling* guna mengurangi jumlah *token* yang diproses. Model dilatih pada set data FER-2013 dan mencapai akurasi 71.24%, lebih tinggi dari *baseline* 70.40%. Penggunaan *Token Downsampling* dapat mengurangi kompleksitas komputasi model hingga dua kali lipat. Evaluasi sistem *end-to-end* pada set data ruang kelas menunjukkan bahwa pendekatan ini berhasil mendeteksi hampir semua ekspresi wajah yang ada (*recall* mencapai 99.88% dari sudut pandang tengah). Meskipun presisi klasifikasi masih menjadi tantangan, strategi pelatihan gabungan terbukti mampu meningkatkan performa secara signifikan, menegaskan bahwa model ini adaptif untuk diterapkan di lingkungan pembelajaran nyata.

Kata kunci: *Hybrid Vision Transformer*, MobileNetV3, Pengenalan Ekspresi Wajah, Ruang Kelas, *Token Downsampling*

FACIAL EXPRESSION RECOGNITION OF STUDENT IN CLASSROOM USING HYBRID MOBILENETV3-VIT WITH TOKEN DOWNSAMPLING

Arranged by

Mochamad Khaairi

2106416

ABSTRACT

In large classroom environments, teachers often face difficulties in thoroughly monitoring the facial expressions of each student during the learning process. However, facial expressions can reflect students' emotional states and levels of participation. This study aims to develop and evaluate a robust facial expression recognition system under real-world conditions. A hybrid model is proposed, combining MobileNetV3 for local feature extraction and a Vision Transformer for global context modeling, enhanced with Token Downsampling to reduce the number of processed tokens. The model was trained on the FER-2013 dataset and achieved an accuracy of 71.24%, outperforming the baseline of 70.40%. Token Downsampling significantly reduced the model's computational complexity by up to half. End-to-end system evaluation on a classroom dataset shows that this approach successfully detects nearly all existing facial expressions (recall reached 99.88% from a front viewpoint). Although classification precision remains a challenge, the combined training strategy proved to significantly improve performance, confirming that this model is adaptive for implementation in a real learning environment.

Keywords: *Classroom, Facial Expression Recognition, Hybrid Vision Transformer, MobileNetV3, Token Downsampling*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR PUSTAKA

Ahdiat, A. (2024). *Ini Perbandingan Jumlah Mahasiswa dan Dosen di Indonesia*. Databoks.Katadata.Co.Id. https://databoks.katadata.co.id/demografi/statistik/04fd4a4ef248a8f/ini-perbandingan-jumlah-mahasiswa-dan-dosen-di-indonesia

Bengio, Y., Goodfellow, I., & Courville, A. (2016). Deep learning. *MIT Press*, *29*(7553), 1–73. http://deeplearning.net/

Bieg, M., Goetz, T., Sticca, F., Brunner, E., Becker, E., Morger, V., & Hubbard, K. (2017). Teaching methods and their impact on students' emotions in mathematics: an experience-sampling approach. *ZDM - Mathematics Education*, *49*(3), 411–422. https://doi.org/10.1007/s11858-017-0840-1

Bouhlal, M., Aarika, K., AitAbdelouahid, R., Elfilali, S., & Benlahmar, E. (2020). Emotions recognition as innovative tool for improving students' performance and learning approaches. *Procedia Computer Science*, *175*, 597–602. https://doi.org/10.1016/j.procs.2020.07.086

Cerqueira, V., Santos, M., Roque, L., Baghoussi, Y., & ... (2025). Online data augmentation for forecasting with deep learning. *ArXiv Preprint ArXiv …*. https://arxiv.org/abs/2404.16918

Damasio, A. R. (1998). Emotion in the perspective of an integrated nervous system. *Brain Research Reviews*, *26*(2–3), 83–86. https://doi.org/10.1016/S0165-0173(97)00064-7

Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 5202–5211. https://doi.org/10.1109/CVPR42600.2020.00525

Deng, L., & Yu, D. (2013). Deep learning: Methods and applications. *Foundations and Trends in Signal Processing*, *7*(3–4), 197–387. https://doi.org/10.1561/2000000039

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE. *ICLR 2021 - 9th International Conference on Learning Representations*.

Dukić, D., & Krzic, A. S. (2022). Real-Time Facial Expression Recognition Using Deep Learning with Application in the Active Classroom Environment. *Electronics (Switzerland)*, *11*(8). https://doi.org/10.3390/electronics11081240

Ekman, P. (1970). Universal facial expressions of emotion. California Mental Health Research Digest, 8(4), 151–158.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system (Issue v. 1). Consulting Psychologists Press.

El Boudouri, Y., & Bohi, A. (2023). EmoNeXt: an Adapted ConvNeXt for Facial Emotion Recognition. *2023 IEEE 25th International Workshop on Multimedia Signal Processing, MMSP 2023*. https://doi.org/10.1109/MMSP59012.2023.10337732

Fang, B., Li, X., Han, G., & He, J. (2023). Facial Expression Recognition in Educational Research from the Perspective of Machine Learning: A Systematic Review. *IEEE Access*, *11*(August), 112060–112074. https://doi.org/10.1109/ACCESS.2023.3322454

Foret, P., Kleiner, A., Mobahi, H., & Neyshabur, B. (2021). Sharpness-Aware Minimization for Efficiently Improving Generalization. *ICLR 2021 - 9th International Conference on Learning Representations*.

Gkontzis, A. F., Karachristos, C. V., Panagiotakopoulos, C. T., Stavropoulos, E. C., & Verykios, V. S. (2017). Sentiment analysis to track emotion and polarity in student fora. *ACM International Conference Proceeding Series*, *Part F1325*(February 2018). https://doi.org/10.1145/3139367.3139389

Goodfellow, I. J., Erhan, D., Luc Carrier, P., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D. H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., … Bengio, Y. (2015). Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, *64*, 59–63. https://doi.org/10.1016/j.neunet.2014.09.005

Goyal, S., Choudhury, A. R., Raje, S. M., Chakaravarthy, V. T., Sabharwal, Y., & Verma, A. (2020). PoWER-BERT: Accelerating BERT inference via progressive word-vector elimination. *37th International Conference on Machine Learning, ICML 2020*, *PartF16814*, 3648–3657.

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. 1–17. http://arxiv.org/abs/2008.05756

Haurum, J. B., Escalera, S., Taylor, G. W., & Moeslund, T. B. (2023). Which Tokens to Use? Investigating Token Reduction in Vision Transformers. *Proceedings - 2023 IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2023*, 773–783. https://doi.org/10.1109/ICCVW60793.2023.00085

Herwin, H., & Mardapi, D. (2017). An emotion assessment model for Elementary School students. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *21*(1), 80–92. https://doi.org/10.21831/pep.v21i1.14504

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, *18*(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Howard, A., Wang, W., Chu, G., Chen, L., Chen, B., & Tan, M. (2019). Searching for MobileNetV3. *International Conference on Computer Vision*, 1314–1324.

Hu, Y., Zeng, Z., Yin, L., Wei, X., Tu, J., & Huang, T. S. (2008). A study of non-frontal-view facial expressions recognition. *2008 19th International Conference on Pattern Recognition*, 1–4. https://doi.org/10.1109/ICPR.2008.4761052

Huang, Q., Huang, C., Wang, X., & Jiang, F. (2021). Facial expression recognition with grid-wise attention and visual transformer. *Information Sciences*, *580*, 35–54. https://doi.org/10.1016/j.ins.2021.08.043

Huang, Y., Chen, F., Lv, S., & Wang, X. (2019). Facial expression recognition: A survey. *Symmetry*, *11*(10). https://doi.org/10.3390/sym11101189

Indolia, S., Nigam, S., Singh, R., Singh, V. K., & Singh, M. K. (2023). Micro Expression Recognition Using Convolution Patch in Vision Transformer. *IEEE Access*, *11*(August), 100495–100507. https://doi.org/10.1109/ACCESS.2023.3314797

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(19), 7241–7244. https://doi.org/10.1073/pnas.1200155109

Jiang, B., Li, N., Cui, X., Liu, W., Yu, Z., & Xie, Y. (2024). Research on Facial Expression Recognition Algorithm Based on Lightweight Transformer. *Information*, *15*(6). https://doi.org/10.3390/info15060321

Joy, D. T., Kaur, G., Chugh, A., & Bajaj, S. B. (2021). COMPUTER VISION FOR COLOR DETECTION. *International Journal of Innovative Research in Computer Science & Technology*, *9*(3), 53–59. https://doi.org/10.21276/ijircst.2021.9.3.9

Karhunen, J., Raiko, T., & Cho, K. H. (2015). Unsupervised deep learning: A short review. *Advances in Independent Component Analysis and Learning Machines*, 125–142. https://doi.org/10.1016/B978-0-12-802806-3.00007-5

Khan, A., Rauf, Z., Sohail, A., Khan, A. R., Asif, H., Asif, A., & Farooq, U. (2023). A survey of the vision transformers and their CNN-transformer based variants. *Artif. Intell. Rev.*, *56*(Suppl 3), 2917–2970. https://doi.org/10.1007/s10462-023-10595-0

Kim, J. W., Khan, A. U., & Banerjee, I. (2025). Systematic Review of Hybrid Vision Transformer Architectures for Radiological Image Analysis. *Journal of Imaging Informatics in Medicine*. https://doi.org/10.1007/s10278-024-01322-4

Kim, S., Nam, J., & Ko, B. C. (2022). Facial Expression Recognition Based on Squeeze Vision Transformer. *Sensors*, *22*(10). https://doi.org/10.3390/s22103729

Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *Sensors (Switzerland)*, *18*(2). https://doi.org/10.3390/s18020401

Krithika, L. B., & Lakshmi Priya, G. G. (2016). Student Emotion Recognition System (SERS) for e-learning Improvement Based on Learner Concentration Metric. *Procedia Computer Science*, *85*, 767–776. https://doi.org/10.1016/J.PROCS.2016.05.264

Lawpanom, R., Songpan, W., & Kaewyotha, J. (2024). Advancing Facial Expression Recognition in Online Learning Education Using a Homogeneous Ensemble Convolutional Neural Network Approach. *Applied Sciences (Switzerland)*, *14*(3). https://doi.org/10.3390/app14031156

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, G., Zhang, J., Zhang, M., Wu, R., Cao, X., & Liu, W. (2022). Efficient depthwise separable convolution accelerator for classification and UAV object detection. *Neurocomputing*, *490*, 1–16. https://doi.org/10.1016/j.neucom.2022.02.071

Li, H., Xu, Z., Taylor, G., Studer, C., & Goldstein, T. (2018). Visualizing the Loss Landscape of Neural Nets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 31). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/a41b3bb3e6b050b6c9067c67f663b915-Paper.pdf

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

Lin, Q., & Lai, X. (2021). Research on the recognition of students' classroom learning state based on facial expressions. *Journal of Physics: Conference Series*, *1914*(1). https://doi.org/10.1088/1742-6596/1914/1/012052

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision -- ECCV 2014* (pp. 740–755). Springer International Publishing.

Marin, D., Chang, J.-H. R., Ranjan, A., Prabhu, A., Rastegari, M., & Tuzel, O. (2021). *Token Pooling in Vision Transformers*. 1–21. http://arxiv.org/abs/2110.03860

Martinez, B., & Valstar, M. F. (2016). Advances, challenges, and opportunities in automatic facial expression recognition. *Advances in Face Detection and Facial Image Analysis*, 63–100. https://doi.org/10.1007/978-3-319-25958-1_4

Mehrabian, A., & Russell, J. A. (1974). *An Approach to Environmental Psychology*.

https://psycnet.apa.org/record/1974-22049-000

Mohamed, O., Ababou, N., Alaoui, S. O. El, & Aouragh, S. L. (2024). Deep Facial Expression Recognition. *Lecture Notes in Networks and Systems*, *838 LNNS*(05), 339–345. https://doi.org/10.1007/978-3-031-48573-2_49

Mukhopadhyay, M., Pal, S., Nayyar, A., Pramanik, P. K. D., Dasgupta, N., & Choudhury, P. (2020). Facial Emotion Detection to Assess Learner's State of Mind in an Online Learning System. *ACM International Conference Proceeding Series*, *June*, 107–115. https://doi.org/10.1145/3385209.3385231

Pabba, C., & Kumar, P. (2022). An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Systems*, *39*(1), 0–2. https://doi.org/10.1111/exsy.12839

Padilla, R., Netto, S. L., & da Silva, E. A. B. (2020). A Survey on Performance Metrics for Object-Detection Algorithms. *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 237–242. https://doi.org/10.1109/IWSSIP48289.2020.9145130

Pan, C., Mu, H., Yuan, Q., & Zhang, Y. (2025). *Research on Methods for Recognizing and Analyzing the Emotional State of College Students*. *7*(1), 27–33. https://doi.org/10.25236/FSST.2025.070105

Ping, H. (2024). Advancing Facial Expression Recognition: A Comparative Study of CNNs and Transformers. *2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering, ICEACE 2024*, 222–226. https://doi.org/10.1109/ICEACE63551.2024.10898937

Qian, S., Ning, C., & Hu, Y. (2021). MobileNetV3 for Image Classification. *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 490–497. https://doi.org/10.1109/ICBAIE52039.2021.9389905

Qin, Y. Y., Cao, J. T., & Ji, X. F. (2021). Fire Detection Method Based on Depthwise Separable Convolution and YOLOv3. *International Journal of Automation and Computing*, *18*(2), 300–310. https://doi.org/10.1007/s11633-020-1269-5

Rajae, A., Amina, R., & El Hassane, I. E. H. (2025). A Hybrid Method for Student Engagement Recognition Using Handcrafted Features and Vision Transformer. In Y. Farhaoui, T. Herawan, A. L. Imoize, & A. El Allaoui (Eds.), *Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment* (pp. 402–409). Springer Nature Switzerland.

Rinn, W. E. (1984). The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin*, *95*(1), 52–77. https://doi.org/10.1037/0033-2909.95.1.52

Roy, A. K., Kathania, H. K., Sharma, A., Dey, A., & Ansari, M. S. A. (2025). ResEmoteNet: Bridging Accuracy and Loss Reduction in Facial Emotion Recognition. *IEEE Signal Processing Letters*, *32*, 491–495. https://doi.org/10.1109/LSP.2024.3521321

Santoso, R. R., Megasari, R., & Hambali, Y. A. (2020). Implementasi Metode Machinelearning. Jurnal Aplikasi Dan Teori Ilmu Komputer, 3(2), 85–97. https://ejournal.upi.edu/index.php/JATIKOM

Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*(6), 1–20. https://doi.org/10.1007/s42979-021-00815-1

Sebe, N. (2005). Machine Learning in Computer Vision. In *Machine Learning in Computer Vision*. https://doi.org/10.1007/1-4020-3275-7

Sharma, P., Sharma, P., Deep, V., & Shukla, V. K. (2021). Facial Emotion Recognition Model. In N. Kumar, S. Tibor, R. Sindhwani, J. Lee, & P. Srivastava (Eds.), *Advances in Interdisciplinary Engineering* (pp. 751–761). Springer Singapore.

Shen, Z. (2024). A Comparative Study of Hybrid CNN and Vision Transformer Models for Facial Emotion Recognition. *2024 11th International Conference on Dependable Systems and Their Applications (DSA)*, 401–408. https://doi.org/10.1109/DSA63982.2024.00061

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0197-0

Swain, A., Laha, S. R., Sahoo, S., Dalei, A., Srivastav, V., & Kumar Nayak, D. S. (2025). Facial Emotion Recognition for University Students using CNN: Transforming Learning Environment. *2025 International Conference on Intelligent and Cloud Computing (ICoICC)*, 1–6. https://doi.org/10.1109/ICoICC64033.2025.11052034

Szeliski, R. (2011). Computer vision: algorithms and applications. *Choice Reviews Online*, *48*(09), 48-5140-48–5140. https://doi.org/10.5860/choice.48-5140

Tishby, N., Pereira, F. C., & Bialek, W. (2000). *The information bottleneck method*. https://arxiv.org/abs/physics/0004057

Tonguç, G., & Ozaydın Ozkara, B. (2020). Automatic recognition of student emotions from facial expressions during a lecture. *Computers and Education*, *148*(August 2019), 103797. https://doi.org/10.1016/j.compedu.2019.103797

Valiente, C., Swanson, J., & Eisenberg, N. (2012). Linking students emotions to academic acheivement. *Child Development Perspect*, *6*(2), 129–135. https://doi.org/10.1111/j.1750-8606.2011.00192.x.Linking

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N.,

Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Villegas-Ch, W. E., García-Ortiz, J., & Sánchez-Viteri, S. (2023). Identification of Emotions From Facial Gestures in a Teaching Environment With the Use of Machine Learning Techniques. *IEEE Access*, *11*, 38010–38022. https://doi.org/10.1109/ACCESS.2023.3267007

Walecki, R., Rudovic, O., Pavlovic, V., Schuller, B., & Pantic, M. (2017). Deep structured learning for facial action unit intensity estimation. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-Janua*, 5709–5718. https://doi.org/10.1109/CVPR.2017.605

Wei, W., Jia, Q., & Chen, G. (2016). Real-time facial expression recognition for affective computing based on Kinect. *Proceedings of the 2016 IEEE 11th Conference on Industrial Electronics and Applications, ICIEA 2016*, 161–165. https://doi.org/10.1109/ICIEA.2016.7603570

Wihardi, Y., Junaeti, E., Setiawan, W., Wahyudin, W., & Erlangga, E. (2022). Smart Classroom System (SCS) Berbasis Kamera Untuk Memantau Keadaan Peserta Didik. *INFORMATION SYSTEM FOR EDUCATORS AND PROFESSIONALS : Journal of Information System*, *6*(1), 67. https://doi.org/10.51211/isbi.v6i1.1771

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., & Vajda, P. (2020). *Visual Transformers: Token-based Image Representation and Processing for Computer Vision*. http://arxiv.org/abs/2006.03677

Wu, Q., Liu, Y., Li, Q., Jin, S., & Li, F. (2017). The application of deep learning in computer vision. *Proceedings - 2017 Chinese Automation Congress, CAC 2017*, *2017-Janua*, 6522–6527. https://doi.org/10.1109/CAC.2017.8243952

Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., & Girshick, R. (2021). Early Convolutions Help Transformers See Better. *Advances in Neural Information Processing Systems*, *36*(NeurIPS), 30392–30400.

xiaoju ye. (2023). *calflops: a FLOPs and Params calculate tool for neural networks in pytorch framework*. https://github.com/MrYxJ/calculate-flops.pytorch

Zhao, X., Shi, X., & Zhang, S. (2015). Facial Expression Recognition via Deep Learning. *IETE Technical Review*, *32*(5), 347–355. https://doi.org/10.1080/02564602.2015.1017542