KLASIFIKASI AKSI PESERTA DIDIK DI DALAM RUANG KELAS MENGGUNAKAN *VIDEO MASKED AUTOENCODER*

SKRIPSI

Diajukan untuk Memenuhi Sebagian dari Syarat Memperoleh Gelar Sarjana Ilmu Komputer

oleh:

Meutia Jasmine Annisa Herawan

2000188

PROGRAM STUDI ILMU KOMPUTER

FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS PENDIDIKAN INDONESIA

2025

KLASIFIKASI AKSI PESERTA DIDIK DI DALAM RUANG KELAS MENGGUNAKAN *VIDEO MASKED AUTOENCODER*

Disusun Oleh:

Meutia Jasmine Annisa Herawan

NIM 2000188

Diajukan untuk memenuhi salah satu syarat memperoleh gelar Sarjana Komputer pada Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

Universitas Pendidikan Indonesia

Agustus 2025

MEUTIA JASMINE ANNISA HERAWAN

2000188

KLASIFIKASI AKSI PESERTA DIDIK DI DALAM RUANG KELAS MENGGUNAKAN *VIDEO MASKED AUTOENCODER*

DISETUJUI DAN DISAHKAN OLEH:

Pembimbing I,

Dr. Rani Megasari, S.Kom., M.T.

NIP. 198705242014042002

Pembimbing II,

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

NIP. 198903252015041001

Mengetahui,

Ketua Program Studi Ilmu Komputer

Dr. Muhammad Nursalman, M.T.

NIP. 197909292006041002

PERNYATAAN

Dengan ini penulis menyatakan bahwa skripsi dengan judul "Klasifikasi Aksi Peserta Didik di dalam Ruang Kelas Menggunakan *Video Masked Autoencoder*" ini beserta seluruh isinya adalah benar-benar karya penulis sendiri. Penulis tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika ilmu yang berlaku dalam masyarakat keilmuan. Atas pernyataan ini, penulis siap menanggung risiko/sanksi apabila di kemudian hari ditemukan adanya pelanggaran etika keilmuan atau ada klaim dari pihak lain terhadap keaslian karya penulis ini.

Bandung, Agustus 2025

Yang Membuat Pernyataan,

Meutia Jasmine Annisa Herawan

2000188

KATA PENGANTAR

Segala puji bagi Allah SWT. Penguasa semesta alam serta pemilik segala kesempurnaan. Alhamdulillah penulis ucapkan karena atas segala pertolongan, rahmat, dan kasih sayang-Nya penulis dapat menyelesaikan skripsi dengan judul "Klasifikasi Aksi Peserta Didik Di Dalam Ruang Kelas Menggunakan *Video Masked Autoencoder*". Shalawat serta salam penulis persembahkan kepada Rasulullah Shallallahu Alaihi Wassalam yang telah menjadi sumber inspirasi dan teladan yang baik bagi umat manusia.

Skripsi ini ditulis dengan tujuan sebagai salah satu syarat untuk memperoleh gelar sarjana komputer pada jenjang studi S1 Program Studi Ilmu Komputer, Universitas Pendidikan Indonesia.

Akhir kata penulis menyadari jika penelitian pada skripsi ini bukanlah skripsi yang sempurna. Oleh karena itu, penulis meminta maaf atas kesalahan yang dilakukan penulis di dalam penelitian ini. Penulis berharap jika penelitian ini dapat bermanfaat bagi orang lain dan dapat dikembangkan menjadi lebih baik lagi di masa mendatang.

Bandung, Agustus 2025

Meutia Jasmine Annisa Herawan

UCAPAN TERIMA KASIH

Alhamdulillahi Rabbil'alamin, puji dan syukur kehadirat Allah SWT atas rahmat, hidayah dan nikmat yang sangat luar biasa. Atas karunia serta berkat yang engkau berikan, akhirnya karya tulis ilmiah ini dapat diselesaikan. Shalawat serta salam selalu tercurah limpahkan kepada Rasulullah SAW.

Dalam proses penyelesaian skripsi, penulis tidak luput dihadapkan dengan berbagai kendala dan rintangan, namun dengan adanya bimbingan, dukungan, dan bantuan dari berbagai pihak baik secara langsung maupun tidak langsung, penulis dapat menyelesaikan tugas akhir ini dengan baik. Pada kesempatan ini penulis mengucapkan terima kasih sebesar besarnya pada semua pihak yang telah terlibat dalam memberikan bimbingan, dukungan dan bantuan selama proses penyelesaian tugas akhir skripsi ini sebagai sebuah penghargaan kepada:

1. Kedua orang tua penulis, Bapak Heldy Wahyu Dwi Herawan dan Ibu Cut Mytha Fitriana, yang selalu memberikan kasih sayang, doa, dan pengorbanan senantiasa menjadi sumber kekuatan dan inspirasi bagi penulis. Dukungan, semangat, serta bimbingan telah menjadi fondasi utama yang mengantarkan penulis hingga dapat menyelesaikan studi dan penelitian ini.
2. Abang E'an dan Adik Kumay, yang selalu memberikan dukungan moral, hiburan, serta semangat di saat penulis menghadapi kesulitan.
3. Keluarga Besar Thaib Ubit, terutama nenek penulis, yang selalu mendoakan dan memberi dorongan sehingga penulis dapat menyelesaikan studi.
4. Ibu Dr. Rani Megasari, S.Kom. M.T. selaku dosen pembimbing I yang telah meluangkan waktu, tenaga, dan pikiran serta memberikan ilmu pengetahuan dan bimbingan kepada penulis, sehingga skripsi ini dapat terselesaikan dengan baik.
5. Bapak Yaya Wihardi, M.Kom. selaku dosen pembimbing II dan dosen wali yang telah meluangkan waktu, tenaga, dan pikiran serta memberikan ilmu pengetahuan dan bimbingan kepada penulis, sehingga skripsi ini dapat terselesaikan dengan baik.

6. Bapak Dr. Muhammad Nursalman, M.T. selaku Ketua Program Studi Ilmu Komputer Universitas Pendidikan Indonesia yang telah memberikan arahan kepada penulis.
7. Bapak dan Ibu Dosen Program Studi Ilmu Komputer yang telah membimbing dan memberikan ilmu yang bermanfaat semasa kuliah pada penulis.
8. Sahabat-sahabat penulis, Azka, Amirah, Sarah, dan Devia yang telah banyak membantu, menemani, dan mendukung penulis selama masa perkuliahan.
9. Semua pihak yang tidak dapat penulis sebutkan satu persatu yang telah membantu dalam menyelesaikan skripsi ini.

KLASIFIKASI AKSI PESERTA DIDIK DI DALAM RUANG KELAS MENGGUNAKAN VIDEO MASK AUTOENCODER

oleh

Meutia Jasmine Annisa Herawan —meutia.ja@upi.edu

2000188

ABSTRAK

Dalam dunia pendidikan, evaluasi terhadap aksi peserta didik di ruang kelas merupakan indikator penting keberhasilan proses pembelajaran. Pendekatan berbasis *deep learning* menggunakan CNN telah banyak diterapkan untuk pengenalan aksi, tetapi CNN memiliki keterbatasan dalam menangkap hubungan spasial-temporal data. Saat ini, ViT menjadi metode baru yang menunjukkan performa unggul pada tugas pengenalan visual, termasuk pengenalan aksi manusia berbasis video. Penelitian ini bertujuan mengembangkan model klasifikasi aksi peserta didik di dalam kelas menggunakan arsitektur *video transformer*, yaitu VideoMAE, dan membandingkannya dengan TimeSformer sebagai model *baseline* melalui teknik *fine-tuning*. Karena belum tersedia dataset standar yang sesuai dengan konteks penelitian, penulis menyusun dataset khusus yang mencakup lima kelas aksi, yaitu mengangguk, mengangkat tangan, menggunakan ponsel, menopang kepala, dan menunduk, dengan total 403 klip video. Hasil penelitian menunjukkan bahwa VideoMAE mencapai akurasi terbaik sebesar 87.23% pada data uji, dengan parameter optimal berupa 15 epoch, batch size 4, learning rate 5e-5, *weight decay* 0.01, dan linear learning rate scheduler, melampaui TimeSformer yang memperoleh akurasi 79.80%. Temuan ini menunjukkan bahwa VideoMAE efektif untuk tugas klasifikasi aksi peserta didik di ruang kelas. Pendekatan ini diharapkan dapat menjadi solusi yang lebih objektif dan efisien untuk pengamatan aksi peserta didik, sekaligus menjadi rujukan pengembangan model *video transformer* pada domain pendidikan dengan dataset berukuran kecil.

Kata Kunci: Klasifikasi Aksi, Fine-tune, Aksi Peserta Didik, Pengenalan Aksi Manusia, *Time-Series Transformers(TimeSformer),* Video Masked Auto Encoder (VideoMAE)

STUDENT ACTIONS CLASSIFICATION INSIDE THE CLASSROOM USING VIDEO MASK AUTOENCODER

arranged by

Meutia Jasmine Annisa Herawan — meutia.ja@upi.edu

2000188

ABSTRACT

In the education world, evaluating students' actions in the classroom is an important success indicator of the learning process. Deep learning approaches based on CNNs have been widely applied for action recognition. However, CNNs have limitations in capturing spatio-temporal relations within data. Recently, Vision Transformers (ViT) have emerged as a promising method, demonstrating superior performance on visual recognition tasks, including video-based human action recognition. This study aims to develop a model for classifying students' actions in the classroom using a video transformer architecture, namely VideoMAE, and compare its performance with TimeSformer as a baseline model through fine-tuning techniques. Since no standard dataset suitable for the research context was available, a custom dataset was constructed comprising five action classes—nodding, raising a hand, using a phone, resting the head on the desk, and looking down, for a total of 403 video clips. The results show that VideoMAE achieved the highest accuracy of 87.23% on the test set, with optimal parameters of 15 epochs, batch size of 4, learning rate of 5e-5, weight decay of 0.01, and a linear learning rate scheduler, outperforming TimeSformer, which achieved 79.80% accuracy. These findings demonstrate that VideoMAE is effective for classroom action classification. This approach is expected to provide a more objective and efficient solution for observing students' actions and serve as a reference for further research of video transformer models in the educational domain, particularly with small-scale datasets.`

Keywords: *Action Classification, Fine-Tune, Student Action, Human Action Recognition, Time-Series Transformers(TimeSformer), Video Masked AutoEncoder (VideoMAE).*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR PUSTAKA

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6836-6846).

Beddu, M. (2017). Evaluasi Belajar Peserta Didik (Siswa). Jurnal Idaarah (Vol. 1, No. 2), 257-267.

Bertasius, G., Wang, H., & Torresani, L. (2021, July). Is space-time attention all you need for video understanding?. In Icml (Vol. 2, No. 3, p. 4).

Bhuiyan Shawon, J. A., Mahmud, H., & Hasan, K. (2025). Fine-tuning video transformers for word-level Bangla sign language: A comparative analysis for classification tasks. arXiv preprint arXiv:2506.04367. https://arxiv.org/abs/2506.04367

Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R. (2005). Action as Space - Time Shapes. Tenth IEEE International Conference on Computer Vision (Vol. 2, 1395–1402)

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages[J]. Research in higher education, 47(1), 1-32.

Chollet, F. (2018,). Deep Learning with Python. Manning Publications Co.

Chathanadath, D. (2023). "What is positive and negative behavior?" [Forum Diskusi Daring Quora]. Diakses dari https://knowyourpersonality1.quora.com/What-is-positive-and-negative-behavior.

DeepWiki. (2025, 21 April). "Model Architecture | facebookresearch/TimeSformer | DeepWiki". [*Online*]. Diakses di https://deepwiki.com/facebookresearch/TimeSformer/2-model-architecture

Deshpande, Anagha & Deshpande, Vedant. (2023). Student Activity Recognition in Classroom Environments Using Transfer Learning. 1-6. 10.1109/ICCINS58907.2023.10450091.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Gelly, S., Uszkoreit J., & Houlsby, N. (2020). An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale (Version 2).

Fanucchi, R.Z., Junior, A.R., Marques, G.D., Rodrigues, L.B., Soares, A.D., & Soares, T.W. (2024). Fine-Tuning a Video Masked Autoencoder to Develop an Augmented Reality Application for Brazilian Sign Language Interpretation. Symposium on Virtual and Augmented Reality.

Fatima, N. (2024, 4 Juli). *Understanding Tensors in Machine Learning: From 0D to 5D*. [*Online*]. Diakses dari https://medium.com/@noorfatimaafzalbutt/understanding-tensors-in-machine-learning-from-0d-to-5d-9313eb77c999

Febriana, R. (2021). Evaluasi Pembelajaran. Bumi Aksara.

Feng, J. &Shi, Y. (2025). A bibliometric studies of pre-trained model and fine-tune method. Procedia Computer Science, Volume 266, Pages 1295-1304, ISSN 1877-0509. https://doi.org/10.1016/j.procs.2025.08.159.

Gantini, R. (2022). Pengenalan Aksi pada Video Menggunakan Metode Transformer. (Skripsi). Program Studi Ilmu Komputer, Universitas Pendidikan Indonesia, Bandung.

Glasscock, J & Tenenbaum, S. (2023, 11 Januari). Action. [*Online*]. Diakses dari https://plato.stanford.edu/archives/spr2023/entries/action/

Gowda, S. N., Arnab, A., & Huang, J. (2023). Optimizing vivit training: Time and memory reduction for action recognition. arXiv preprint arXiv:2306.04822.

Halim, S., Wahid, R., & Halim, T. (2018). Classroom Observation- a Powerful Tool for Continous Professional Development (CPD). International Journal on Language Research and Education Studies.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16000-16009).

Jahne, B., Haussecker, H., & Geissler, P. (1999). Handbook of Computer Vision and Applications: Volume 2: From Images to Features. Academic Press, Inc., USA.

Jin, H. (2022). Hyperparameter importance for machine learning algorithms. arXiv preprint arXiv:2201.05132.

Khatimah, H. (2021). Major Impact of Classroom Environment in Students' Learning.

Kong, Yu & Fu, Y. (2022). Human Action Recognition: A Survey: V3. ArXiv.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural net-works. NeurIPS (Vol. 25).

Learned-Miller, E. G. (2012). Introduction to Computer Vision. University of Massachusetts. Amherst, Amherst.

Li, X. Wang, M., Zeng, W., & Lu, W. (2019). A Students' Action Recognition Database In Smart Classroom. The 14th International Conference on Computer Science & Education (ICCSE 2019) August 19-21. Toronto, Canada.

Lin, F. Ngo, H., Dow, C. Lam, K., & Le, H. (2021). Student Behavior Recognition System for the Classroom. Sensors 2021, 21, 5314. https://doi.org/10.3390/s21165314.

Meta. (2021, 12 Maret). "*TimeSformer: A new architecture for video understanding*". [*Online*]. Diakses di https://ai.meta.com/blog/timesformer-a-new-architecture-for-video-understanding/

Mardapi, D. (2015). Pengukuran, Penilaian, dan Evaluasi Pendidikan. Yogyakarta: Nuha litera.

Miller, P. W. (2008). Measurement and teaching. Munster, Indiana: Patrick W. Miller & Associates.

Mulyanto, E., Pritadi, E., Purnomo, O., & Hafidz, I. (2024). Improvement of Tradition Dance Classification Process Using Video Vision Transformer based on Tubelet Embedding.International Journal, of Intelligent Engineering System (INAAS).

Panagakis, Y., Kossaifi, J., Chrysos, G. G., Oldfield, J., Nicolaou, M. A., Anandkumar, A., & Zafeiriou, S. (2021). Tensor methods in computer vision and deep learning. Proceedings of the IEEE, 109(5), 863-890.

Pandelu, A. P. (2024, 15 Desember). Day 48: Training Neural Networks — Hyperparameters, Batch Size, Epochs. Medium. [*Online*]. https://medium.com/%40bhatadithya54764118/day-48-training-neural-networks-hyperparameters-batch-size-epochs-712c57d9e30c.

Pareek, P. & Thakkar, (2020). A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications. Springer Nature.

Rasmi, M., Ashwin, T. S., & Guddeti, R. M. R. (2020). Surveillance Video Analysis for Student Action Recognition and Localization Inside Computer Laboratories of A Smart Campus. *Multimed Tools Appl* **80**, 2907–2929 (2021). https://doi.org/10.1007/s11042-020-09741-5.

Ruan, L. & Jin, Q. (2022). Survey: Transformer based video-language pre-training. AI Open Volume 3, 2022, Pages 1-13, ISSN 2666-6510. https://doi.org/10.1016/j.aiopen.2022.01.001.

Sammut, C. (2011). Encyclopedia of machine learning, G. I. Webb, Ed. Springer, 3 2011. [Online]. Available: https://books.google.com/books/about/ Encyclopedia of Machine Learning.html?hl=id&id=i8hQhp1a62UC.

Sharma, V., Gupta, M., Kumar, A., & Mishra, D. EduNet: A New Video Dataset for Understanding Human Activity in the Classroom Environment. Sensors 2021, 21, 5699. https://doi.org/10.3390/s21175699.

Shou, Z., Yuan, X., Li, D., Mo, J., Zhang, H., Zhang, J., & Wu, Z. (2024). A Dynamic Position Embedding-Based Model for Student Classroom Complete Meta-Action Recognition. Sensors, 24(16), 5371. https://doi.org/10.3390/s24165371.

Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820.

Sugiyono. (2017). Metode Penelitian & Pengembangan (Research Development/R&D). Bandung: Alfabeta.

Sugiyono. (2019). Metode Penelitian & Pengembangan (Research Development/R&D). Bandung: Alfabeta.

Sun, B., Wu, Y., Zhao, K., He, J., Yu, L.,Yan, H., & Luo, A. (2021). Student Class Behavior Dataset: a video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. Springer Nature 2021.

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 (pp. 5998-6008).

Voulodimos, A., Doulamis, N.D., Doulamis, A.D., & Protopapadakis, E.E. (2018). Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience, 2018*.

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

Wortsman, M., Ilharco, G., Kim, J., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., Schmidt, L. (2022). Robust Fine-Tuning of Zero-Shot Models. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7959-7971.

Yao, G., Lei, T., & Zhong, J. 2018. A Review of Convolutional-Neural-Network-Based Action Recognition, Pattern Recognition Letters (2018), doi:10.1016/j.patrec.2018.05.018.

Yu, B., Liu, Y., & Chan, K. 2021. Multimodal Fusion via Teacher-Student Network for Indoor Action Recognition. The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21).

Zheng, R., JIang, F., & Shen, R. 2020. GestureDet: Real-time Student Gesture Analysis with Multi-dimensional Attention-based Detector. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20).