**REFERENCES**

[1] Q. Hao and L. Qin, ‘‘The design of intelligent transportation video

processing system in big data environment,’’ *IEEE Access*, vol. 8,

pp. 13769–13780, 2020.

[2] K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, and S. W. Baik,

‘‘Efficient deep CNN-based fire detection and localization in video surveillance

applications,’’ *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 7,

pp. 1419–1434, Jul. 2019.

[3] L. Simoni, A. Scarton, C. Macchi, F. Gori, G. Pasquini, and S. Pogliaghi,

‘‘Quantitative and qualitative running gait analysis through an innovative

video-based approach,’’ *Sensors*, vol. 21, no. 9, p. 2977, Apr. 2021.

[4] M. Shorfuzzaman, M. S. Hossain, and M. F. Alhamid, ‘‘Towards the

sustainable development of smart cities through mass video surveillance:

A response to the COVID-19 pandemic,’’ *Sustain. Cities Soc.*, vol. 64,

Jan. 2021, Art. no. 102582.

[5] A. B. Mabrouk and E. Zagrouba, ‘‘Abnormal behavior recognition for

intelligent video surveillance systems: A review,’’ *Expert Syst. Appl.*,

vol. 91, pp. 480–491, Jan. 2018.

[6] A. B. Tanfous, H. Drira, and B. B. Amor, ‘‘Sparse coding of shape trajectories

for facial expression and action recognition,’’ *IEEE Trans. Pattern*

*Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2594–2607, Oct. 2020.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘ImageNet classification

with deep convolutional neural networks,’’ in *Proc. NIPS*, Dec. 2012,

pp. 1097–1105.

[8] C. Szegedy,W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan,

V. Vanhoucke, and A. Rabinovich, ‘‘Going deeper with convolutions,’’

in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015,

pp. 1–9.

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, ‘‘Rethinking

the inception architecture for computer vision,’’ in *Proc. IEEE Conf.*

*Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[10] K. He, X. Zhang, S. Ren, and J. Sun, ‘‘Deep residual learning for image

recognition,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*,

Jun. 2016, pp. 770–778.

[11] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He,

J. Mueller, R. Manmatha, M. Li, and A. Smola, ‘‘ResNeSt: Split-attention

networks,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.Workshops*

*(CVPRW)*, Jun. 2022, pp. 2735–2745.

[12] K. Simonyan and A. Zisserman, ‘‘Two-stream convolutional networks for

action recognition in videos,’’ in *Proc. Adv. Neural Inf. Process. Syst.*,

vol. 27, Dec. 2014, pp. 1–9.

[13] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. van Gool,

‘‘Temporal segment networks: Towards good practices for deep action

recognition,’’ in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland:

Springer, Oct. 2016, pp. 20–36.

[14] Z. Lan, Y. Zhu, A. G. Hauptmann, and S. Newsam, ‘‘Deep local video

feature for action recognition,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern*

*Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1219–1225.

[15] A. Diba, V. Sharma, and L. Van Gool, ‘‘Deep temporal linear encoding

networks,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*,

Jul. 2017, pp. 1541–1550.

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, ‘‘Learning

spatiotemporal features with 3D convolutional networks,’’ in *Proc. IEEE*

*Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[17] J. Carreira and A. Zisserman, ‘‘Quo vadis, action recognition? A new

model and the kinetics dataset,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern*

*Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[18] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri,

‘‘A closer look at spatiotemporal convolutions for action recognition,’’

in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018,

pp. 6450–6459.

[19] C. Feichtenhofer, ‘‘X3D: Expanding architectures for efficient video

recognition,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*

*(CVPR)*, Jun. 2020, pp. 200–210.

[20] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, ‘‘Rethinking spatiotemporal

feature learning: Speed-accuracy trade-offs in video classification,’’

in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 305–321.

[21] M. Zolfaghari, K. Singh, and T. Brox, ‘‘ECO: Efficient convolutional

network for online video understanding,’’ in *Proc. Eur. Conf. Comput. Vis.*

*(ECCV)*, Sep. 2018, pp. 713–730.

[22] C. Feichtenhofer, H. Fan, J. Malik, and K. He, ‘‘SlowFast networks for

video recognition,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*,

Oct. 2019, pp. 6202–6211.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai,

T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly,

J. Uszkoreit, and N. Houlsby, ‘‘An image is worth 16 × 16 words: Transformers

for image recognition at scale,’’ 2020, *arXiv:2010.11929*.

[24] M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. S. Khan, and

M.-H. Yang, ‘‘Intriguing properties of vision transformers,’’ in *Proc.*

*NeurIPS*, Dec. 2021, pp. 23296–23308.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo,

‘‘Swin transformer: Hierarchical vision transformer using shifted windows,’’

in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021,

pp. 10012–10022.