Journal of Computer Technology and Applied Mathematics
Vol. 1, No. 1, 2024

Cross-Task Multi-Branch Vision Transformer for Facial Expression and Mask Wearing Classification

ZHU, Armando [1*] LI, Keqin [2] WU, Tong [3] ZHAO, Peng [4] HONG, Bo [5]

[1] Carnegie Mellon University, USA
[2] AMA University, Philippines
[3] University of Washington, USA
[4] Microsoft, China
[5] Northern Arizona University, USA

**** ZHU, Armando is the corresponding author, E-mail: armandozhu@cmu.edu***

Abstract: With wearing masks becoming a new cultural norm, facial expression recognition (FER) while taking masks into account has become a significant challenge. In this paper, we propose a unified multi-branch vision transformer for facial expression recognition and mask wearing classification tasks. Our approach extracts shared features for both tasks using a dual-branch architecture that obtains multi-scale feature representations. Furthermore, we propose a cross-task fusion phase that processes tokens for each task with separate branches, while exchanging information using a cross attention module. Our proposed framework reduces the overall complexity compared with using separate networks for both tasks by the simple yet effective cross-task fusion phase. Extensive experiments demonstrate that our proposed model performs better than or on par with different state-of-the-art methods on both facial expression recognition and facial mask wearing classification task.

Keywords: Vision Transformer, Facial Expression Recognition, Facial Mask Wearing Classification, Deep Learning.

DOI: https://doi.org/10.5281/zenodo.11083875

1 Introduction

Understanding human through analyzing face has been an increasingly important task for technologies interacting with humans, which allows for various application scenarios, such as robot human interaction [12, 23], security surveillance, and so on. It also serves as a basis for various vision tasks [7, 19, 20, 33]. Facial expression recognition (FER), in particular, has been extensively studied and reasonable accuracy has been achieved. However, with wearing facial masks becoming a cultural norm in the post COVID-19 era, new challenges arise where wearing facial masks impairs the ability of existing methods for facial expression recognition [14, 21]. Since facial masks occlude a big part of the human face, previous methods for facial expression recognition that assume no or minimal occlusion can be seriously affected. Therefore, the challenge to tackle face-mask aware facial expression recognition has begun to draw attention.

In this paper, we propose a model based on ViT, that allows for multi-task classification for facial expression and mask wearing classification within a unified framework. This is based on the observation of the inherent correlation between both tasks. In this case, the number of parameters required for the model is much less than the approach to classify expression and mask wearing condition with two individual networks. Moreover, to explicitly utilize information targeting both tasks, we draw inspiration from CrossViT [5] and propose a cross-task fusion module. The experiment result shows that our proposed approach outperforms existing methods while maintaining a low computational cost.

Figure 1 Sample Images from Facial Expression Dataset, CK+, and Mask Classification Dataset, MMD

2 Related Work

2.1 Facial Expression Recognition

Facial Expression Recognition (FER) started from handcrafted [6] solutions for handcrafted features and attributes to solutions based on deep learning [8, 9]. Turan et al. [6] proposed a system using region-based handcrafted

features for facial expression recognition. More recently, since more sophisticated datasets have become publicly available, many CNN deep learning models [4, 16, 17, 32] have been proposed for this task.

Hybrid solutions have also been proposed which combines deep learning techniques with handcrafted techniques [11, 31, 35, 44]. Levi et al. [35] proposed to apply CNN on the image, with LBP features using Multi Dimensional Scaling (MDS). More recently, many Transformer models have been introduced for different computer vision tasks. Ma et al. [11] proposed a convolutional vision transformer model that extracts features from the input image as well as form its LBP using a ResNet18.

Other researchers have thought about handling occlusion in face recognition. These approaches have been presented by [13] and [48]. An existing study on FER that takes face masks into account and recognizes emotions only from the the eyes region. This approach was evaluated on a masked FER-2013 dataset which included seven emotions [48] However, the other facial areas such as the forehead are discarded while only the eyes region is isolated using landmark detection methods, which decreases the accuracy of the FER system.

2.2 Mask Wearing Classification

Various types of deep learning methods are capable of the task of human mask detection [10, 15]. Jagadeeswari et al. [15] conducted an extensive review of the typical convolution neural network with different optimizers, and found out MobileNetV2 combined with Adam optimizer achieved the highest accuracy. Different approaches targeting multi-task learning have also been presented [27, 36, 37]. Ge et al. [18] propose a CNN based method, which combines two pre-trained CNNs for extracting facial regions, then described by LLE.

2.3 Vision Transformer

Dosovitskiy et al. [1] proposed ViT for the task of image classification. The ViT model achieves state-of-the-art performance on ImageNet for image classification, using ImageNet and JFT-300M [3] for training. Although the performance of ViT network is promising, its basic version requires a larger amount of data and has non trivial computation cost. To improve ViT models from different aspects, many variants of ViT have also been proposed, using novel tokenization process [43] or pyramid structure like CNNs [22]. T2T-ViT [43], specifically, introduces a T2T tokenization module to encode the local structural information and reduce the token size. Transformer based models [26] have also been adopted for various tasks [25, 39-41]. Cross ViT introduces a dual-path structure to extract multi-scale features to balance fine grained feature extraction with reasonable computational cost.

3 Methodology

In this section, we introduce the proposed solution in three separate paragraphs: some details of the ViT architecture, additive attention mechanism, and the proposed Multi-Branch Vision Transformer model.

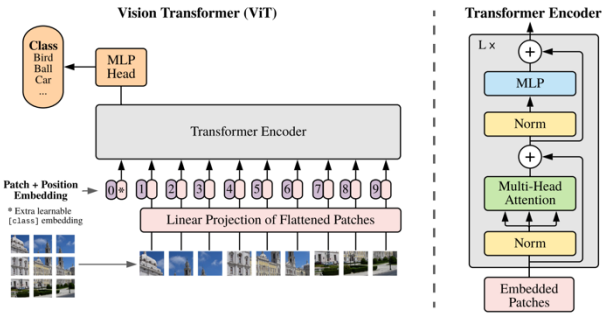

Figure 2 Architecture of Vision Transformer

3.1 Vision Transformer

The overall structure of the Vision Transformer [1] architecture, as illustrated in Figure 2, consists of two phases: tokenization phase, and Transformer encoding phase. During the tokenization phase, image is split into L fixed-size patches of size $(h * h)$ and flattened into a one-dimensional vector. Each patch s is linearly embedded into a lower-dimensional space using a learnable linear projection:

$$Embed(x_i) = x_i W_e + b_e$$

, where W_e and b_e are learnable weight and bias parameters. Positional encodings are added to the patch embeddings to incorporate spatial information.

Figure 2 (right) illustrates the architecture of the Transformer encoder. If consists of L blocks of the attention module. Multi-head self-attention is the main part of the attention block. It is built with parallel heads of self-attention. In the architecture of a self-attention layer, keys, values, and queries all originate from the same source. This structure enables each position within the encoder to consider all positions from the encoder's prior layer. The self-attention function will be a mapping of a query (Q or Q-layer) and a set of key-value (K or K-layer; V or V-layer) pairs to an output. The self-attention function is summarized by:

$$\text{Attention(Q, K, V)} = \text{softmax}\left(QK^T / \sqrt{d_k}\right)V$$

, where d_k is the dimensionality of the keys. Multi-head attention extends the self-attention by performing multiple attention operations in parallel. This is achieved by splitting the queries, keys, and values into multiple smaller vectors and applying self-attention independently to each of these smaller vectors. Afterward, the outputs of the attention heads are concatenated and linearly transformed to obtain the final output. Mathematically, multi-head attention is computed as follows:

$$\text{MultiHead}(Q,\ K,\ V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

, where each head is computed as:

$$head_i\ =\ \text{Attention}\left(QW_i^Q,\ KW_i^K,\ VW_i^V\right)$$

, where W_i^Q, W_i^K, and W_i^V are learnable weight matrices for each attention head i, and W^O is another learnable weight matrix that combines the outputs of all attention heads.

3.2 Additive Attention

Additive attention, also known as Bahdanau Attention, is an alternative mechanism for computing attention scores in sequence-to-sequence models. It uses a one-hidden layer feed-forward network to calculate the attention alignment score:

$$f_{att}(h_i, s_j)\ =\ v_a^T\ \tanh(W_a[h_i;\ s_j])$$

Where v_a and W_a are learnable attention parameters. Here h is the hidden states for the encoder, and s represents the hidden states for the decoder. Additive attention provides flexibility in capturing complex relationships between query and key vectors, making it suitable for tasks where non-linear or intricate dependencies exist. The function mentioned serves as an alignment score function. In the context of a neural network, after obtaining these alignment scores, we further process them by applying a softmax function to derive the final scores.

3.3 Multi-Branch Vision Transformer

Figure 4 illustrates the pipeline of our proposed multi-branch vision transformer model. In brief, the model consists of two phases, unified feature extraction as the first phase, and cross task feature fusion as the second phase. This architecture allows for cross task learning while sharing the same low-level feature.

For the first phase, we adopt a dual branch ViT [5] to extract features, including two branches for different purposes: (1) L-Branch: a large branch that utilizes bigger patch size (P_l) with more transformer encoders, and (2) S-Branch: a small, complementary branch that operates at fine-grained patch size (P_s) with smaller embedding dimensions. Output of both branches are fused using cross attention module L_1 times, and the classification token for L-Branch is used for prediction. This dual branch architecture balances performance by incorporating fine-grained information from S-Branch, while having relatively small computational cost when processing L-Branch.

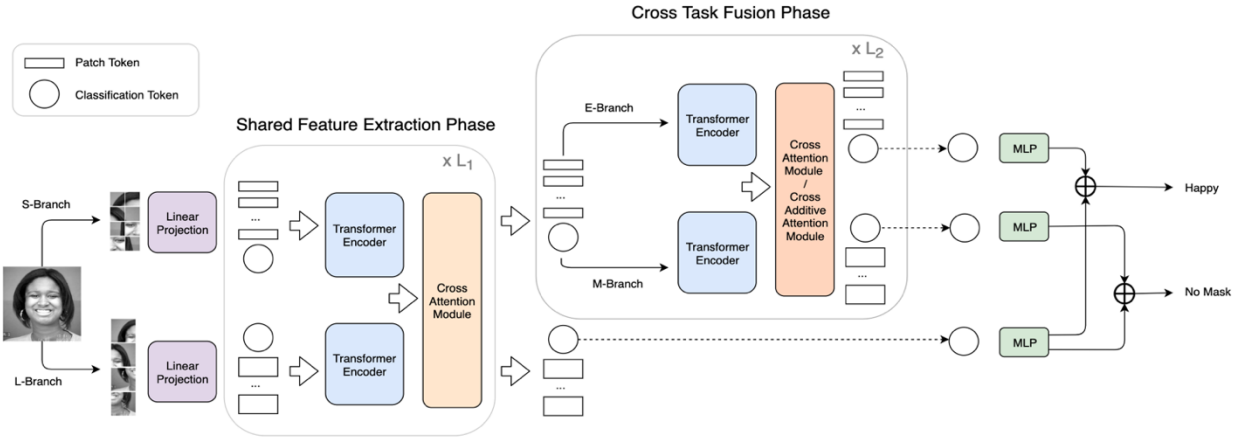

Figure 3 Architecture of Our Proposed Pipeline

The second phase takes in outputs from S-Branch during the first phase. Two branches are introduced in this phase: (1) E-branch, branch dedicated for emotion recognition task, and (2) M-branch, branch dedicated for mask wearing classification task. Both branches share same patch size (P_s) as S-Branch from phase 1. The number of encoders is independently configurable for both branches. Resulting features, i.e. outputs from S-Branch are duplicated and fed to both branches. They are fused using cross attention module L_2 times, and the last fusion is performed using cross-additive-attention to aggregate the resulting features. Classification token from each branch is fused with classification token from L-Branch for prediction of both tasks. This design allows us to explictly exploit information and correlation between two tasks without the additional cost of having two separate networks for different tasks.

Figure 5 illustrates the cross-attention module. Specifically, for one of the two branches, it first collects the patch tokens from the other branch, and concatenates its own classification tokens. The cross-additive-attention module replaces the dot-product attention with an additive attention module to enhance the aggregation capability of the feature fusion.

We use two-stage training to finetune the whole model. For the first stage, we train the shared classifier, which is the classification token output from L-branch and S-branch, and then in the second stage, we jointly train two output classifiers from two branches and the shared classifier.

4 Results

4.1 Experimental Setup

4.1.1 Datasets

M-FER-2013: The FER-2013 [24] is a dataset that contains 35,887 facial images in 48x48 size in grayscale and seven different types of expressions: angry, disgust, fear, happy, neutral, sad, and surprise. We follow the automatic wearing face mask (AWFM) method used in FMA-3D [38] to form a new dataset M-FER-2013 with 35,609 images. We use the training and testing splits for training and evaluation.

M-RAF-DB: The Real-world Affective Faces Database (RAF-DB) [42] is published on 2017, consisting of around 30K great-diverse facial images downloaded from the Internet. We use the version with seven basic expressions for training, and use AWFM to randomly add masks with a probability of 0.5.

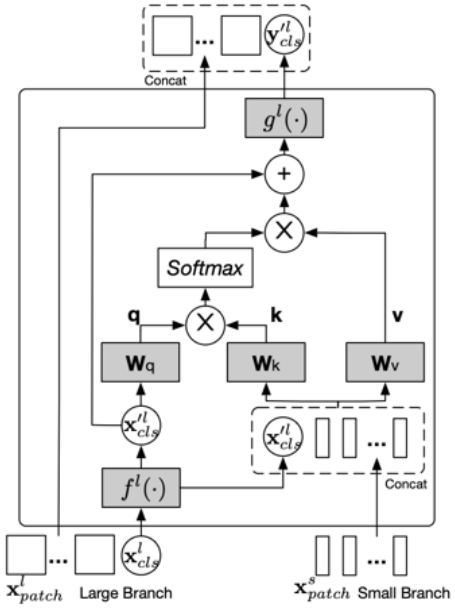

Figure 4 Cross Attention Module

CK+/M-CK+: The CK+ dataset [52] consists of 593 video sequences with 10-60 frames captured from 123 subjects. In our case, we evaluate our results on the seven basic expressions to ensure a consistent comparison with different methods. We uniformly sampled and collected 1508 images for testing, and use AWFM to randomly add masks with a probability of 0.5 for M-CK+.

JAFFE: The JAFFE [46] dataset contains 213 images of different facial expressions captured from 10 different Japanese female subjects. Each subject was asked to do 7 basic facial expressions and the images were annotated with average semantic ratings. We use the dataset for validation.

MMD-FMD: [50] proposed a merged version of combining Medical Masks Dataset (MMD) and Face Mask Dataset (FMD). The MMD Dataset consists of 682 pictures with over 3k medical masked faces wearing masks, while the FMD dataset consists of 853 images. The resulting merged dataset contained 1,415 after data cleansing.

4.1.2 Implementation Details

In all the following experiments, we use the model architecture based on CrossViT-B [5], the base version of CrossViT with 224×224 as input imge size, 12×12 as the small patch and 16×16 for the large patch, and 3 multi-scale fusion iterations. We use the model weights pretrained on ImageNet1K [2]. All images are converted to 3 channels, resized to (224 × 224) and normalized. We use batch size of 16 and fixed learning rate ot $1 * 1.0^{-4}$, and train for 8 epochs for the first phase, and 2 epochs for the second phase.

4.2 Main Results

4.2.1 Facial Expression Recognition

Figure 5 Sample images from M-RAF-DB with mask (up) and without mask (bottom)

In this section, we compare our proposed model trained on FER-2013 and MMD-FMD with different solutions on M-CK+ dataset, and further compare with state-of-the-art methods on JAFFE, RAF-DB and FER, which are in-the-wild datasets. It can be seen from Table 1 that accuracy on scenarios with masks is significantly affected. Under this case, our proposed model outperforms other traditional and handcrafted methods and improves the accuracy based on CrossViT. We achieve this with only a small increase in FLOPs and model parameters compared with a two-fold increase for CrossViT model (42.4 GFLOPS and 209.4M Parameters).

Table 1 Comparison with Different Methods

Model	Accuracy		FLOPs	Params
	CK+	M-CK+		
Handcrafted [6]	0.9503	0.6359	N/A	N/A
VGG19 [30]	0.9658	0.6847	19.7G	144M
ResNet50 [45]	0.9723	0.7003	25.7G	4.3M
ViT [1]	0.9830	0.744	17.6G	86.7M
Cross ViT [5]	0.9865	0.7532	21.2G	104.7M
Proposed	0.9856	0.7702	24.6G	125.8M

Table 2 shows more quantitative results on different datasets compared with state-of-the-art methods, based on either CNN or visual transofmer. On both lab datasets and wild datasets, our model consistently achieves good results

over other methods. We have the highest accuracy on M-JAFFE with a 70.59% acuraccy and 77.85% on M-FER-2013, both on scenarios where mask is randomly applied.

Table 2 Comparison with State-of-the-Art (Accuracy)

Model	M-JAFFE	M-RAF-DB	M-FER-2013
ResNet50 [45]	0.6602	0.6126	0.6978
ViT [1]	0.688	0.643	0.7519
Cross ViT [5]	0.6994	0.6479	0.7632
Ours	0.7059	0.6438	0.7785

4.2.2 Facial Mask Detection

We compare our performance of mask classification with existing methods. The performances of different methods compared with our proposed network can be found in Table 3. Generally, CNN-based networks outperform the traditional SVM method, while Transormer based methods outperform CNN methods. Our method showed slight improvements over the ViT network while sharing the underlying network with FER task.

Table 3 Evaluation Results on MMD-FMD (Accuracy)

Model	Accuracy
SVM RBF	0.8527
VGG19 [30]	0.9459
ResNet50 [45]	0.9675
ViT [1]	0.9774
Ours	0.9793

4.3 Ablation Studies

Different attention methods. We compare different settings of our network, substituting the cross-addition-attention fusion method with cross-attention method which uses dot-product attention. Table 4 shows that cross-addition-attention achieves higher accuracy on both M-CK+ and M-JAFFE dataset. Intuitively, additive attention module is a better fit for the scenario of feature aggregation and final prediction.

Importance of the cross-task fusion module. We also experiment with predicting from fusing the output directly from both L-Branch and S-Branch from phase 1 while training for both tasks. Result on Table 4 reveals that the model degrades without using the output from the second phase where cross attention is applied, which is 2% worse than our proposed architecture. This shows the effectiveness of the cross-task learning and prediction phase.

Table 4 Evaluation Result on Different Network Configurations

Model	Accuracy		FLOPs	Params
	M-CK+	M-JAFFE		
Ours with dot-product attention	0.7689	0.7033	24.6G	125.8M
Ours with output from first phase	0.7485	0.6859	21.2G	104.7M
Ours	0.7702	0.7059	24.6G	125.8M

5 Conclusion

In this paper, we present the Cross-Task Multi-Branch Vision Transformer, a novel multi-branch vision transformer for learning mult-tasks features, to improve the accuracy of facial expression recognition and mask detection tasks. We utilize the similarity of both classification tasks that are able to share a joint representation, and then by using the proposed cross task fusion stage, task specific representation can be obtained while exchanging information by cross attention modules. By conducting extensive experiments, we demonstrate our proposed model performs better than or on par with several recent works, in addition to other baseline methods. In future work, we aim to generalize the network architecture to provide a solution for more generalized and competitive tasks.

Acknowledgments

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Funding

Not applicable.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author Contributions

Not applicable.

About the Authors

ZHU, Armando

Armando Zhu obtained his Master of Science degree in Software Engineering from Carnegie Mellon University. His research interests include computer vision and machine learning.

LI, Keqin

Affiliation: AMA University, Philippines.

WU, Tong

Affiliation: University of Washington.

ZHAO, Peng

Affiliation: Microsoft, China.

HONG, Bo

Affiliation: Northern Arizona University.

References

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[2] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[3] Sun, Chen, et al. "Revisiting unreasonable effectiveness of data in deep learning era." Proceedings of the IEEE international conference on computer vision. 2017.

[4] Li, Panfeng, Youzuo Lin, and Emily Schultz-Fellenz. "Contextual hourglass network for semantic segmentation of high resolution aerial imagery." arXiv preprint arXiv:1810.12813 (2018).

[5] Chen, Chun-Fu Richard, Quanfu Fan, and Rameswar Panda. "Crossvit: Cross-attention multi-scale vision transformer for image classification." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[6] Turan, Cigdem, and Kin-Man Lam. "Region-based feature fusion for facial-expression recognition." 2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2014.

[7] Zhu, Ziwei, and Wenjing Zhou. "Taming heavy-tailed features by shrinkage." International Conference on Artificial Intelligence and Statistics. PMLR, 2021.

[8] Farzaneh, Amir Hossein, and Xiaojun Qi. "Facial expression recognition in the wild via deep attentive center loss." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021.

[9] Wang, Kai, et al. "Region attention networks for pose and occlusion robust facial expression recognition." IEEE Transactions on Image Processing 29 (2020): 4057-4069.

[10] Shi, Ge, Jason Smucny, and Ian Davidson. "Deep learning for prognosis using task-fmri: A novel architecture and training scheme." Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022.

[11] Ma, Fuyan, Bin Sun, and Shutao Li. "Robust facial expression recognition with convolutional visual transformers." arXiv preprint arXiv:2103.16854 2.6 (2021): 7.

[12] Ding, Wenhao, et al. "Vehicle pose and shape estimation through multiple monocular vision." 2018 IEEE International Conference on Robotics and Biomimetics (ROBIO). IEEE, 2018.

[13] Osherov, Elad, and Michael Lindenbaum. "Increasing cnn robustness to occlusions by reducing filter support." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[14] Weng, Yijie, Jianhao, Wu. "Fortifying the global data fortress: a multidimensional examination of cyber security indexes and data protection measures across 193 nations". International Journal of Frontiers in Engineering Technology 6. 2(2024).

[15] Jagadeeswari, C., and M. Uday Theja. "Performance evaluation of intelligent face mask detection system with various deep learning classifiers." International Journal of Advanced Science and Technology 29.11s (2020): 3074-3082.

[16] Yao, Jiawei, et al. "Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space." 2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE

Computer Society, 2023.

[17] Yao, Jiawei, et al. "Building lane-level maps from aerial images." ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.

[18] Ge, Shiming, et al. "Detecting masked faces in the wild with lle-cnns." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[19] Tabassum, Tarafder Elmi, et al. "Integrating GRU with a Kalman Filter to Enhance Visual Inertial Odometry Performance in Complex Environments." Aerospace 10.11 (2023): 923.

[20] Read, Andrew J., et al. "Prediction of Gastrointestinal Tract Cancers Using Longitudinal Electronic Health Record Data." Cancers 15.5 (2023): 1399.

[21] Zhao, Peng, et al. "HTN planning with uncontrollable durations for emergency decision-making." Journal of Intelligent & Fuzzy Systems 33.1 (2017): 255-267.

[22] Wang, Wenhai, et al. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[23] Li, Keqin, et al. "The application of Augmented Reality (AR) in Remote Work and Education." arXiv preprint arXiv:2404.10579 (2024).

[24] Goodfellow, Ian J., et al. "Challenges in representation learning: A report on three machine learning contests." Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20. Springer berlin heidelberg, 2013.

[25] Ru, Jingyu, et al. "A Bounded Near-Bottom Cruise Trajectory Planning Algorithm for Underwater Vehicles." Journal of Marine Science and Engineering 11.1 (2022): 7.

[26] Liu, Tianrui, Qi, Cai, Changxin, Xu, Bo, Hong, Jize, Xiong, Yuxin, Qiao, Tsungwei, Yang. "Image Captioning in News Report Scenario". Academic Journal of Science and Technology 10. 1(2024): 284–289.

[27] Zhao, Peng, Chao Qi, and Dian Liu. "Resource-constrained Hierarchical Task Network planning under uncontrollable durations for emergency decision-making." Journal of Intelligent & Fuzzy Systems 33.6 (2017): 3819-3834.

[28] Zhao, Peng, Chao Qi, and Dian Liu. "Resource-constrained Hierarchical Task Network planning under uncontrollable durations for emergency decision-making." Journal of Intelligent & Fuzzy Systems 33.6 (2017): 3819-3834.

[29] Qi, Chao, et al. "Hierarchical task network planning with resources and temporal constraints." Knowledge-Based Systems 133 (2017): 17-32.

[30] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[31] Liu, Dian, et al. "Hierarchical task network-based emergency task planning with incomplete information, concurrency and uncertain duration." Knowledge-Based Systems 112 (2016): 67-79.

[32] Li, Panfeng, Mohamed Abouelenien, and Rada Mihalcea. "Deception Detection from Linguistic and Physiological Data Streams Using Bimodal Convolutional Neural Networks." arXiv preprint arXiv:2311.10944 (2023).

[33] Atulya Shree, Kai Jia, Zhiyao Xiong, Siu Fai Chow, Raymond Phan, Panfeng Li, & Domenico Curro. (2022). Image analysis.

[34] Jin Wang, JinFei Wang, Shuying Dai, Jiqiang Yu, Keqin Li. "Research on emotionally intelligent dialogue generation based on automatic dialogue system." arXiv preprint arXiv:2402.11447 (2024).

[35] Levi, Gil, and Tal Hassner. "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns." Proceedings of the 2015 ACM on international conference on multimodal interaction. 2015.

[36] Xin, Yi, et al. "MmAP: Multi-modal Alignment Prompt for Cross-domain Multi-task Learning." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 14. 2024.

[37] Xin, Yi, et al. "Parameter-Efficient Fine-Tuning for Pre-Trained Vision Models: A Survey." arXiv preprint arXiv:2402.02242 (2024).

[38] Wang, Jun, et al. "Facex-zoo: A pytorch toolbox for face recognition." Proceedings of the 29th ACM international conference on multimedia. 2021.

[39] Liu, Hao, et al. "Deep Reinforcement Learning for Mobile Robot Path Planning." arXiv preprint arXiv:2404.06974 (2024).

[40] Wang, Xiaosong, et al. "Advanced Network Intrusion Detection with TabTransformer." Journal of Theory and Practice of Engineering Science 4.03 (2024): 191-198.

[41] Liu, Tianrui, et al. "News recommendation with attention mechanism." arXiv preprint arXiv:2402.07422 (2024).

[42] Li, Shan, Weihong Deng, and JunPing Du. "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[43] Yuan, Li, et al. "Tokens-to-token vit: Training vision transformers from scratch on imagenet." Proceedings of

the IEEE/CVF international conference on computer vision. 2021.

[44] Wang, Hong-Wei, et al. "Review on hierarchical task network planning under uncertainty." Acta Autom. Sin 42 (2016): 655-667.

[45] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[46] Lyons, Michael, et al. "Coding facial expressions with gabor wavelets." Proceedings Third IEEE international conference on automatic face and gesture recognition. IEEE, 1998.

[47] Liu, Tianrui, Changxin, Xu, Yuxin, Qiao, Chufeng, Jiang, Jiqiang, Yu. "Particle Filter SLAM for Vehicle Localization". Journal of Industrial Engineering and Applied Science 2. 1(2024): 27–31.

[48] Castellano, Giovanna, Berardina De Carolis, and Nicola Macchiarulo. "Automatic emotion recognition from facial expressions when wearing a mask." Proceedings of the 14th Biannual Conference of the Italian SIGCHI Chapter. 2021.

[49] Su, Jing, et al. "Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review." arXiv preprint arXiv:2402.10350 (2024).

[50] Loey, Mohamed, et al. "Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection." Sustainable cities and society 65 (2021): 102600.

[51] Liu, Tianrui, Qi, Cai, Changxin, Xu, Bo, Hong, Fanghao, Ni, Yuxin, Qiao, Tsungwei, Yang. "Rumor Detection with A Novel Graph Neural Network Approach". Academic Journal of Science and Technology 10. 1(2024): 305–310.

[52] Lucey, Patrick, et al. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." 2010 ieee computer society conference on computer vision and pattern recognition-workshops. IEEE, 2010.