A Transformer-Based Multimodal Classification Network for Smart Tire

Tong Wu [1] **Jiahao Li** [1]

Abstract

Recent days have witnessed more and more interest in multimodal sensing and processing. With accurate and delicate sensors invented, data from different modalities can be captured efficiently. Meanwhile, numerous multimodal fusion methods have been proposed. However, in the realm of intelligent tires, the potential of multi-modal fusion and perception technology has not yet been explored. In this work, based on a smart tire that applies multi-modal visuotactile sensing, we propose to carry self and cross-attention mechanisms to fuse multiple modalities and investigate our method for terrain classification and tire breakage detection. We have collected over 3000 raw data for training and evaluation, demonstrating the effectiveness of our method and leading a new route for the application of smart tires. To the best of our knowledge, our current study is novel and unique.

1. Introduction

With the continuous advancement of multimodal visuotactile sensing technology, it is valuable for us to fuse multiple modalities of data to achieve more accurate recognition or detection. How to integrate this technology into vehicle tires to improve their intelligence, such as precise tire status detection, etc., has still not been studied in detail.

In modern intelligent systems, multimodal fusion methods are widely applied in various domains such as computer vision (Feng et al., 2017; Hou et al., 2018; Kumar et al., 2024), natural language processing (Gandhi et al., 2023), and robotics (Xue et al., 2020). The objective of multimodal fusion methods is to effectively integrate and fuse information from different sensors or modalities, enhancing the perceptual capabilities and decision-making performance of systems. In recent years, significant progress has been made in multimodal fusion methods, driven by advancements in technologies such as deep learning and neural networks.

Intelligent tires, also known as smart tires, have emerged as a remarkable advancement in the automotive industry. These tires incorporate advanced sensing and communication technologies to enhance vehicle performance, safety, and efficiency.

The increasing demand for improved safety, fuel efficiency, and vehicle performance drives the development of intelligent tires. Traditional tires have limited capabilities in monitoring and adapting to changing road conditions or detecting potential tire-related issues. Intelligent tires aim to overcome these limitations by utilizing cutting-edge technologies to enable better control, monitoring, and optimization of tire performance.

Intelligent tires incorporate a range of sensors such as pressure sensors, temperature sensors, accelerometers, and optical sensors, tactile sensors (Lee & Taheri, 2017). These sensors continuously monitor tire pressure, temperature, tread wear, and road conditions. The collected data is processed and analyzed in real-time, providing valuable insights to both the driver and the vehicle's control system. One of the key features of intelligent tires is their ability to provide tire condition monitoring. By continuously monitoring tire pressure and temperature, intelligent tires can alert drivers to potential issues such as underinflation or overheating, which can lead to reduced fuel efficiency and increased risk of tire failure.

By fusing and analyzing multimodal sensing data, smart tires can achieve numerous functionalities. For example, they can provide more accurate vehicle attitude estimation and motion control to enhance vehicle maneuverability and stability. Additionally, smart tires can enable real-time road condition monitoring, allowing drivers or vehicle systems to respond appropriately, thereby improving driving safety and comfort. What's more, the visuotactile sensing data collected also provides us the opportunity to detect tire breakage by achieving accurate terrain classification tasks, ensuring optimal traction and safety.

In summary, the technical contributions include:

- We propose a transformer-based multimodal classification network that leverages self-attention and cross-attention mechanisms for extracting multimodal features.

[1] Data Science and Information Technology Research Center, Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, China.

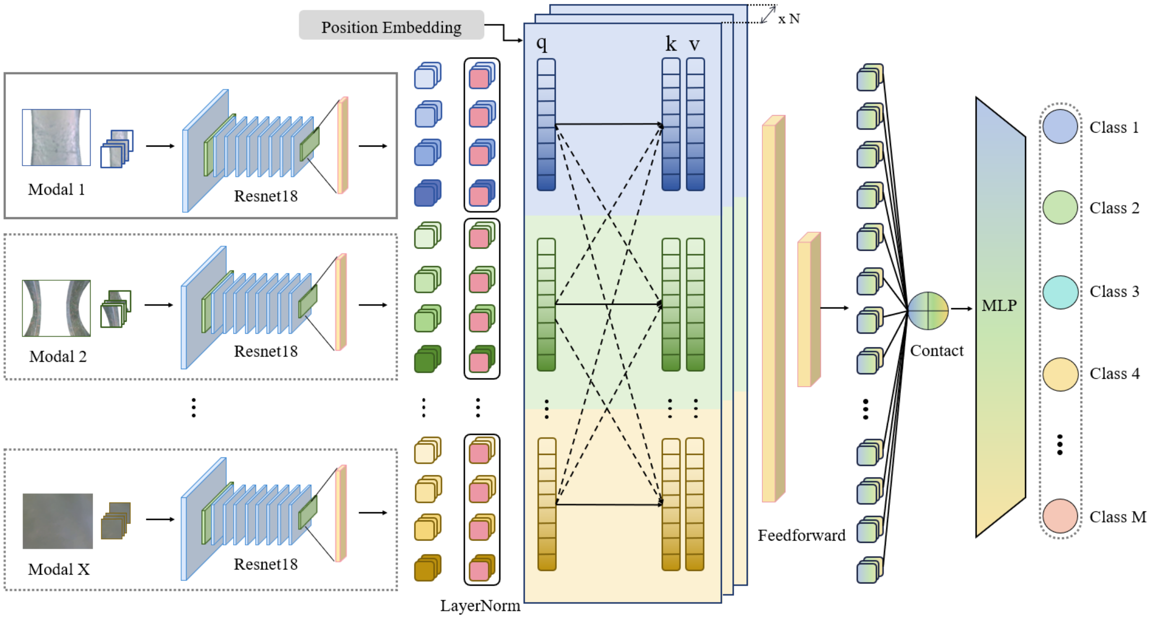

Figure 1. Transformer-based multimodal classification algorithm.

- We evaluate our method on the terrain classification task in complex environments. The results demonstrate that our approach outperforms baseline methods and effectively handles multimodal inputs.
- We also apply our method to tire fault detection, which aids in identifying abnormalities in tires.

2. Related Work

Multimodal Fusing: Currently, researchers have proposed various multimodal fusion methods. For instance, some methods employ fusion strategies based on images and speech (Feng et al., 2017; Hou et al., 2018; Kumar et al., 2024), training and fusing image and speech features to achieve more accurate object recognition and speech understanding. Other approaches combine visual and motion sensor data (Luvizon et al., 2020; Hu et al., 2024), enabling joint analysis of visual and motion information for tasks like human action recognition and pose estimation. Additionally, some methods (Talmor et al., 2021; Ilievski & Feng, 2017) fuse visual and language information for tasks such as image captioning and image question answering. By

Intelligent Tire: By integrating various sensors and data processing capabilities, intelligent tires are capable of providing real-time information about tire conditions, road conditions, and vehicle dynamics. Current methods used in intelligent tires can be divided into three aspects: Estimation based on Acceleration Measurement, Strain Measurement, and Global Deflection Measurement (Lee & Taheri, 2017). Xu et al. (Xu et al., 2020) use machine learning for tire force estimation and propose an intelligent tire system with a three-axis acceleration sensor. khaleghian et al. (Khaleghian & Taheri, 2017) developed a fuzzy logic algorithm that was developed and used for terrain classification, where all different surfaces are classified into four main categories; asphalt, concrete, grass, and sand. Optical sensors also develop in a rapid manner, huber et al. (Huber et al., 2022) present the concept of TireEye, which is an optical device mounted inside the wheel well and facing the road.

3. Method

Given a range of distinct modalities $m_i, i = 0, 1, ..., M-1$, which encompass optical images of the terrain, tactile feedback from the underlying surface, and even partially observable states like dark or smoky images, we employ a set of encoders $\{E_i\}$ tailored specifically for each modality to extract its respective features. Rather than extracting a single global feature for each modality, we divide each modality into K_i distinct fragments and leverage the encoder E_i to extract K_i fragmented features. Since modalities can vary significantly in their data distributions, this can pose challenges for modality fusion and the training process. To alleviate these issues, we apply Layer Normalization (LayerNorm) (Ba et al., 2016) to each modality. This feature extraction process can be formulated as follows:

$$\{f_k\}^i = \text{LayerNorm}(E_i(\text{Seg}(m_i))). \tag{1}$$

To further exploit the information within each individual modality as well as across different modalities, we incorporate self and cross-attention mechanisms (Chen et al., 2022) into our framework for fusing the extracted features. After obtaining the fragmented features $\{f_k\}^i$ for each modality i and fragment k_i, we first concatenate these features in the fragment dimension, effectively grouping all fragments from the same modality together. Next, we concatenate the modality-grouped features in the modality dimension, thereby combining features from all modalities into a single feature map. To preserve the sequential and spatial information within the concatenated feature map, we add position embeddings. This step ensures that the model can distinguish between different positions in the feature map, enabling it to attend to relevant features based on their location. The resulting tokenized feature representation F encapsulates the information from all modalities and fragments:

$$F = \begin{bmatrix} F_0 & F_1 & \cdots & F_{M-1} \end{bmatrix}. \tag{2}$$

To compute the attention, we pass F through linear layers to derive the queries $\{Q_j\}$, keys $\{K_j\}$ and values $\{V_j\}$, for each attention head $j = 1, 2, ..., N$:

$$Q_j = W_j^Q F = \begin{bmatrix} Q_{j,0} & Q_{j,1} & \cdots & Q_{j,M-1} \end{bmatrix}, \tag{3}$$

$$K_j = W_j^K F = \begin{bmatrix} K_{j,0} & K_{j,1} & \cdots & K_{j,M-1} \end{bmatrix}, \tag{4}$$

$$V_j = W_j^V F = \begin{bmatrix} V_{j,0} & V_{j,1} & \cdots & V_{j,M-1} \end{bmatrix}, \tag{5}$$

These queries, keys, and values facilitate the model's ability to attend to relevant features within the same modality (self-attention) and across different modalities (cross-attention). Then we use the following equation to calculate the attention at head j:

$$A_j = V_j \times \text{softmax}((K_j)^T Q_j / \sqrt{d}), \tag{6}$$

where d is the dimension of the multimodal features and the matrix form of $(K_j)^T Q_j$ can be denoted as follows:

$$(K_j)^T Q_j = \begin{bmatrix} (\mathbf{K_{j,0}})^\mathbf{T} \mathbf{Q_{j,0}} & (K_{j,0})^T Q_{j,1} & \cdots \\ (K_{j,1})^T Q_{j,0} & (\mathbf{K_{j,1}})^\mathbf{T} \mathbf{Q_{j,1}} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}, \tag{7}$$

where diagonal elements represent self-attention which uses self-generated queries to draw self-generated keys, while non-diagonal elements represent cross-attention which uses queries and keys generated from different modalities to extract features. Afterward, we pass the attention through a feedforward network and concatenate all tokens into a fused feature.

By leveraging these attention mechanisms, our model is able to effectively extract and utilize information from both individual modalities and the interplay between modalities, enabling it to make more informed decisions and predictions in complex, multimodal environments.

4. Results

4.1. Bimodal Terrain Classification (VTire Bimodal Data)

We collect tactile and visual data from the tire's contact with 16 different terrains to validate the effectiveness of bimodal tires and multimodal sensing networks. We capture 150 images for each terrain, as shown in Fig. 2. These terrains contain rubber tracks, painted roads, brick roads, lawns, and gravel roads made of different colored and sized stones. To better demonstrate the effectiveness of the system, we design different comparison experiments.

First, to test the effectiveness of the smart tires, we compare the classification accuracy in different modalities. We process and divide the collected dataset into the following sets: 1) Tactile data only (TO): raw data is segmented to focus solely on the tactile region; 2) Visual data only (VO): raw data is segmented to focus solely on the transparent and visible region; 3) Raw VisuoTactile data (RVT): raw data encompass both the tactile region and visible region; 4) Segmented VisuoTactile data (SVT): raw data is segmented into tactile region and visible region. For all cases, we split the dataset into 70% for training and 30% for evaluation. To simulate noise caused by mud on the transparent region, we add salt-and-pepper noise to the visual modality. We train our network on an Intel(R) Xeon(R) Gold 5218 with a single GeForce RTX A6000 for 80 epochs. The learning rate is 2e-5, and we repeated the experiment on 3 different seeds. The training results and related indicators are shown in Table 1 and Fig. 2. It can be seen from the results that the classification method with bimodal fusion has higher accuracy, and SVT also achieves a better classification performance than RVT.

Second, to further validate the effectiveness of our proposed network, we compare it with current classical classification algorithms. Specifically, we benchmark our method against two baselines: **1) ResNet** (Koonce & Koonce, 2021): We utilize a pre-trained ResNet50 to extract global features for each modality. These global features are then concatenated and passed through a classification head; **2) LSTM** (Shi et al., 2015): We employ a ResNet18 to extract patched features, which are subsequently processed by an LSTM to achieve a fused feature. The fused feature is passed through a classification head to produce the final output. Our method, **MultiModal VisoTactile Transformer (MMVTT)**, is conceptually similar to the LSTM approach but replaces the recurrent structure with an attention block. To ensure a fair comparison, we design the classification heads of these networks to be as similar as possible. The parameters of MMVTT, LSTM, and ResNet are approximately 14M, 13M, and 18M, respectively. All three networks are trained on the multimodal dataset with a learning rate of 2e-5 for 80 epochs. Additionally, we conduct the experiments using

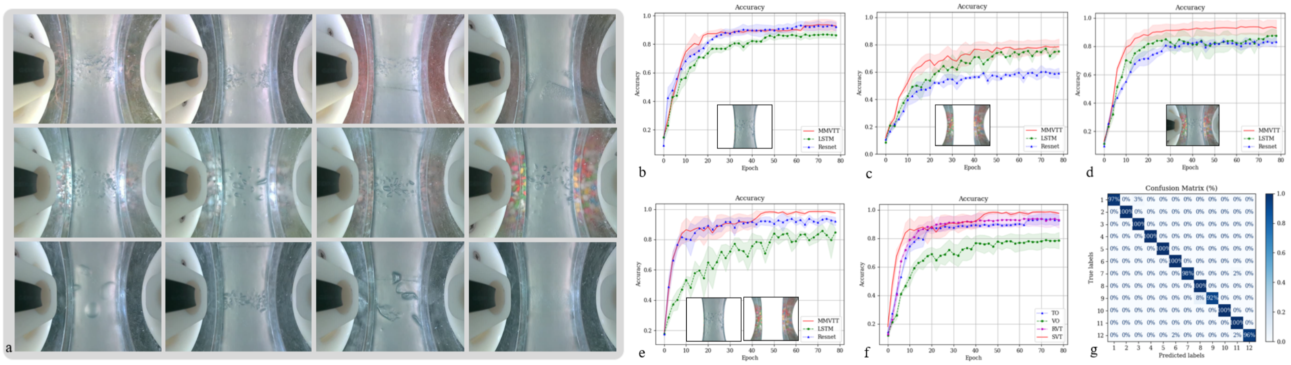

Figure 2. Bimodal terrain classification. (a) Raw visuotactile data in different terrains; (b) The classification result of tactile data only (TO); (c) The classification result of the sensor's visual data only (VO); (d) The classification result of the raw visuotactile data (RVT); (e) The classification result of segmented visuotactile data (SVT); (f) The classification result of our proposed network in different modalities; (g) The confusion matrix of our proposed network for segmented visual image data.

Table 1. Test results under different modal and network conditions

	TO	**VO**	**RVT**	**SVT**
ResNet	**92.7%**/93.4%	59.0%/60.4%	82.3%/83.6%	92.4%/94.0%
LSTM	86.3%/86.7%	74.4%/77.2%	84.8%/87.5%	82.4%/85.7%
MMVTT	92.6%/**93.8%**	**77.9%/77.9%**	**93.4%/93.9%**	**98.1%/98.7%**

three different random seeds to minimize the likelihood of incidental results.

The training results and related indicators are shown in Table 1 and Fig. 2. From the results, we can see that our proposed network outperforms ResNet and LSTM in both last-10-epoch-average and maximum classification accuracy in most cases, especially in bimodal classification, proving our proposed network's effectiveness in dealing with multimodal data.

4.2. Multimodal Terrain Classification (VTire Bimodal Data + External Visual Data)

The most common method for terrain recognition is visual processing because the visual information has a greater detection distance and range. Still, for some smoke, darkness, and other scenes, the visual detection effect will receive a great impact, but the tactile information has better stability. To prove this, we collect visual information from 16 different scenes of normal, smoke, and darkness and compare it with the effect of terrain classification of smart tires. Among them, the data for the smoke scene was collected using a lens wrapped in a semi-permeable film. We capture 150 images for each terrain, as shown in Fig. 3a.

In this experiment, we consider three modalities of inputs: 1) External Visual Only (EVO): only external vision is used for classification (Although the smart tire itself has visual perception capabilities, its perception is often fuzzy with a limited viewing angle. Therefore, we consider adding external vision to enhance detection accuracy further.); 2) External Visual data + Tactile data (EVT): both external vision and segmented tactile region data are employed; 3) External Visual data + segmented VisuoTactile data (EVVT): all available modalities are utilized. The multimodal data was randomly split into training and validation datasets with a 7:3 ratio. We applied MMVTT to the three input modes mentioned above. Each configuration was run with 3 random initializations for 80 epochs, using a learning rate of 2e-5. As illustrated in Table 2 and Fig. 3, the results demonstrate that the EVVT configuration achieves the highest last-10-epoch-average and maximum accuracy with efficient modality fusion.

Furthermore, we compare our method with the previously mentioned baselines: ResNet and LSTM. We utilized all modalities as input, corresponding to the EVVT input mode, and split the data into a training ratio of 0.7. The learning rate is set to 2e-5, and the training process is repeated three times. Table 2 and Fig. 3 present the training and evaluation results. Notably, our network outperforms the other baselines and achieves an accuracy exceeding 99%.

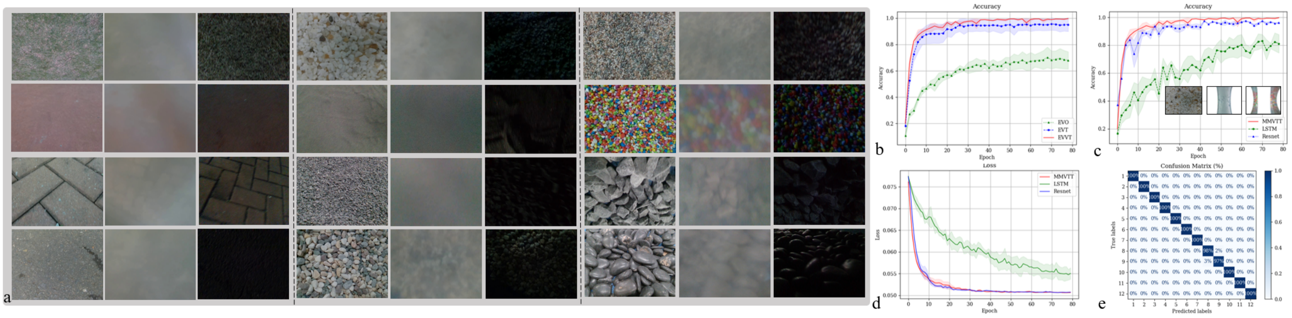

Figure 3. Multimodal terrain classification. (a) Visual images detected by an external camera under sunny, smoky, and dark conditions; (b) The classification result of our proposed network in different modalities; (EVO: external visual only; EVT: external visual data + tactile data; EVVT: external visual data + segmented visual data + tactile data.) (c) The classification result of different networks in segmented visuotactile data with EVVT; (d) The loss of different networks in segmented visuotactile data with EVVT; (e) The confusion matrix of our proposed network for EVVT.

Table 2. Test results under different modal and network conditions

	EVO	**EVT**	**EVVT**
ResNet	-	-	96.1%/97.5%
LSTM	-	-	79.5%/83.0%
MMVTT	68.5%/70.2%	95.1%/95.6%	**99.2%/99.7%**

4.3. Tire Damage Detection Experiment

Tires are susceptible to damage due to contact with sharp objects or long-distance driving while the vehicle is driving. Compared with traditional smart tires, VTire can detect damage in real-time, such as crack, abrasion, and nail penetration, thanks to high-resolution tactile data.

We experiment with different kinds of discrimination to evaluate the capability of detecting damage. We consider 3 types of damage: 1) cracks, 2) irregular wear, 3) punctures, and a normal state. We collect 120 images for each state of tile and split 70% for training. We also add pepper and salt noise to mimic the possible effects of the real environment. MMVTT, LSTM, and ResNet are employed to learn from the training data. We choose a learning rate of 2e-5 and run the experiment on 3 random seeds for 80 epochs. Fig. 4 illustrates the training results and related metrics. We find that MMVTT can detect the damage accurately and correctly classify the type of damage in more than 97% of cases.

5. Conclusions & Discussion

In this work, we propose a transformer-based multimodal classification algorithm based on a bimodal smart tire with high elasticity, high transparency, and high toughness, which overcomes the problems of the low resolution of tactile sensing and small sensing area of traditional smart tires. Firstly, we realize real-time classification of different terrains with multimodal sensing. Furthermore, we achieve robust terrain detection in complex environments with external visual information. Finally, We realize accurate tire fault detection with raw visuotactile inputs.

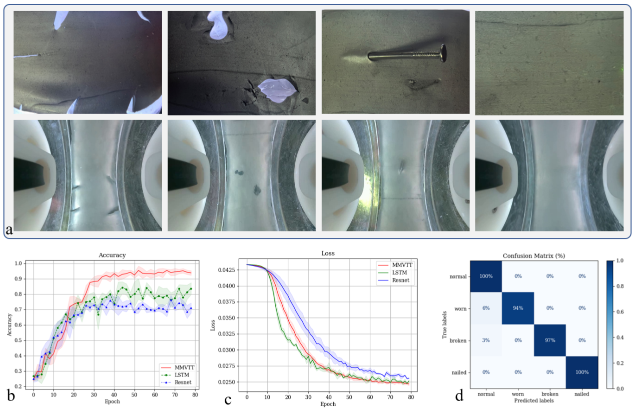

Figure 4. Damage detection. (a) Common tire damage: cracks, irregular wear, punctures, and normal tires (from left to right); (b) Classification accuracy of different networks; (c) Loss curve of different networks; (d) Confusion matrix for the classification result of our proposed network.

There are still some limitations to address in this project. Firstly, more types of modalities should be investigated. Currently, we only consider visual information, using ResNet as the backbone of the encoder. However, if we encounter other modalities, such as sequential data, different encoders like RNNs should be considered. Secondly, more complex tasks should be explored. In this project, we only investigate classification tasks, which are relatively simple. We should consider more challenging tasks, such as autonomous driving. Additionally, we should aim for a more general model. Concurrently, a specific model is trained for each task or even for each dataset. A possible direction is to integrate all the models into a single but more general one, similar to a brain. Knowledge from one domain could potentially benefit other domains.

6. Author Contribution & Acknowledgement

Tong Wu built the model primarily and ran the terrain classification experiments. Jiahao Li conducted the literature survey and ran the tire damage detection experiments. This report is written and edited under the collaboration of both authors. Extremely thankful for Shoujie Li's contribution to the design and deployment of the Smart Tire. He also collected the data for analysis and processing and had many inspiring discussions with us.

References

Ba, J., Kiros, J. R., and Hinton, G. E. Layer normalization. *ArXiv*, abs/1607.06450, 2016. URL https://api.semanticscholar.org/CorpusID:8236317.

Chen, Y., Sipos, A., der Merwe, M. V., and Fazeli, N. Visuo-tactile transformers for manipulation. In *Conference on Robot Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:252682998.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Feng, W., Guan, N., Li, Y., Zhang, X., and Luo, Z. Audio visual speech recognition with multimodal recurrent neural networks. In *2017 International Joint Conference on neural networks (IJCNN)*, pp. 681–688. IEEE, 2017.

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., and Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023.

Hou, J.-C., Wang, S.-S., Lai, Y.-H., Tsao, Y., Chang, H.-W., and Wang, H.-M. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):117–128, 2018.

Hu, Z., Xiao, J., Li, L., Liu, C., and Ji, G. Human-centric multimodal fusion network for robust action recognition. *Expert Systems with Applications*, 239:122314, 2024.

Huber, S., Preindl, P., and Betz, J. Tireeye: Optical on-board tire wear detection. In *Annual Conference of the PHM Society*, volume 14, 2022.

Ilievski, I. and Feng, J. Multimodal learning and reasoning for visual question answering. *Advances in neural information processing systems*, 30, 2017.

Khaleghian, S. and Taheri, S. Terrain classification using intelligent tire. *Journal of Terramechanics*, 71:15–24, 2017.

Koonce, B. and Koonce, B. Resnet 50. *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pp. 63–72, 2021.

Kumar, P., Malik, S., and Raman, B. Interpretable multimodal emotion recognition using hybrid fusion of speech and image data. *Multimedia Tools and Applications*, 83 (10):28373–28394, 2024.

Lee, H. and Taheri, S. Intelligent tires? a review of tire characterization literature. *IEEE Intelligent Transportation Systems Magazine*, 9(2):114–135, 2017.

Luvizon, D. C., Picard, D., and Tabia, H. Multi-task deep learning for real-time 3d human pose estimation and action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2752–2764, 2020.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

Talmor, A., Yoran, O., Catav, A., Lahav, D., Wang, Y., Asai, A., Ilharco, G., Hajishirzi, H., and Berant, J. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*, 2021.

Xu, N., Askari, H., Huang, Y., Zhou, J., and Khajepour, A. Tire force estimation in intelligent tires using machine learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(4):3565–3574, 2020.

Xue, T., Wang, W., Ma, J., Liu, W., Pan, Z., and Han, M. Progress and prospects of multimodal fusion methods in physical human–robot interaction: A review. *IEEE Sensors Journal*, 20(18):10355–10370, 2020.