Contents lists available at Indonesian Scholar Society
Indonesian Journal of Data Risk Research
Journal Homepage:
https://journal.solusiriset.com/index.php/ijdrr/index
e-ISSN: xxxx-xxxx

Traffic Accident Image Classification using CLIP with a Zero-Shot Approach

Della Septiani, Hartiti Fadilah, Josua Alfa Viando Panggabean, Anissa Luthfi Alifia, Syifa Firnanda, Pramudya Wibowo, Ade Lailani*

Data Science Study Program, Sumatra Institute of Technology, Indonesia

Corresponding Email: ade.lailani@sd.itera.ac.id

ABSTRACT

Accident image detection is a critical step in supporting rapid response systems during emergency situations. This study employs the Contrastive Language–Image Pre-training (CLIP) model to detect accident images using a zero-shot approach, without requiring retraining on a specific dataset. The CLIP model leverages multimodal embeddings from text and images, enabling detection based on textual descriptions. The experimental results, using the ViT base patch 32 model, show that this method achieves a Top-1 accuracy of 32% and a Top-5 accuracy of 95.86%. Although the Top-1 accuracy indicates that further optimization is needed, the high Top-5 accuracy demonstrates the significant potential of CLIP for efficient accident image detection. With further development, this technology can serve as a reliable solution for various emergency response scenarios, offering flexibility and efficiency in detecting accident-related images.

ARTICLE INFO

Received
02 Feb 202x
Revised
28 March 202x
Accepted
18 Mei 202x

Keywords: CLIP; Computer Vision; Crash Detection; Multimodal Model; Zero-Shot Learning.

I. INTRODUCTION

In the increasingly developing digital era, artificial intelligence (AI) technology has become an important component in various fields, including public safety [1]. One application that has attracted attention is the automatic detection of accident images using AI technology. This system has the potential to speed up the response to accidents by identifying incidents in real time from images or videos uploaded, whether through surveillance cameras, social media, or

https://doi.org/10.55749/ijdrr.v1.i1.185

other devices. However, the main challenge in developing this system is the need for a large and well-regarded dataset to train the detection model, which is often difficult to obtain [2].

To address this issue, zero-shot learning approaches offer an innovative solution. By leveraging models such as Contrastive Language–Image Pre-Training (CLIP), systems can recognize objects or situations without requiring retraining using specific datasets [3], [4]. CLIP combines natural language capabilities with image recognition, enabling models to more flexibly understand the relationship between text descriptions and visual content. This approach provides high efficiency, especially in situations that require rapid detection without additional training [5].

This research was conducted to examine the potential of using CLIP to detect accident images using a zero-shot learning approach. This approach is expected to enable the system to recognize various types of accidents without requiring extensive training data, thereby accelerating the technology's implementation in the real world. This could be a significant contribution to improving the effectiveness of accident management.

The objective of this research is to develop and test a CLIP-based accident image detection model using a zero-shot learning approach to measure its accuracy and effectiveness. Thus, this research is expected to open new opportunities in the development of AI-based technologies to support public safety and well-being.

II. METHODS

This research uses a Kaggle dataset from Charan Kumar entitled Accident Detection from CCTV Footage [6]. This dataset contains various CCTV recordings focused on detecting traffic accident incidents through image analysis. There were originally three folders in this dataset: training, testing, and validation. Each folder had two classes: accident and non-accident. In this research, only 1 folder of the dataset was used and not split, with a composition of 350 accident image data for the accident class and 350 normal traffic image data for the non-accident class.

Figure 1. (a) Samples from Accident Class and **(b)** Non-Accident Class

In addition to image data, the CLIP research requires multimodal data, namely image data and text data. The text data in this research consists of candidate descriptions for each class.

Ultimately, these candidate descriptions will be selected by the CLIP model, which is used to determine the candidate description relevant to its image by considering the highest similarity score. In Tables 1 and 2, the following candidate descriptions were generated manually.

Table 1. Accident Image Description Candidate

No.	Candidate Description
1	A car accident on a road
2	A car accident involving multiple vehicles
3	A motorcycle accident involving multiple vehicles
4	A damaged vehicle after a collision
5	An overturned car on the highway
6	A car was hit by a truck
7	A car was crushed by rocks
8	The car crashed into a pedestrian crossing
9	The car hit the road divider
10	The truck skidded on the road

Table 2. Non-Accident Image Description Candidate

No.	Candidate Description
1	A calm street with clear traffic
2	A mountain landscape with no traffic
3	A busy street with no accidents
4	Street view after rain with no accidents
5	A car parked on the side of the road
6	Many cars parked on the side of the road

1. CLIP (Contrastive Language Image Pre-training)

CLIP is a machine learning model that involves the use of more than one type of data (multimodal) or sources of information in its training and inference processes [7], [8]. CLIP was developed by OpenAI, which can understand the relationship between text and images [9].

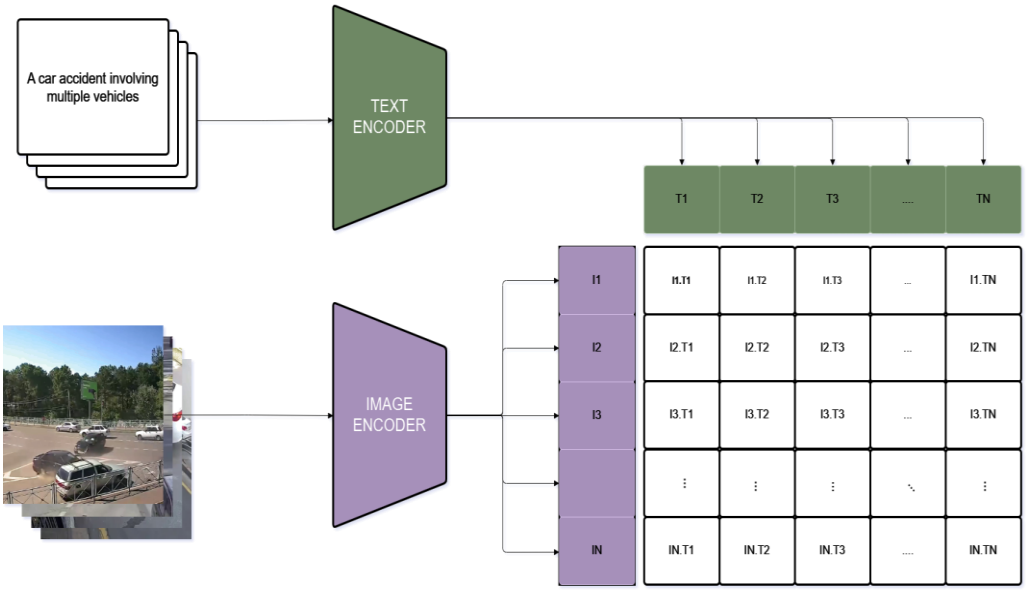

Figure 2. CLIP (Contrastive Language–Image Pre-training) Process

The characteristic of the CLIP method uses 2 types of data, namely images and text, requiring 2 main encoders in the process, namely [10]:

1. Image Encoder, which converts images into vector representations.
2. Text Encoder, which converts text into vector representations.

The CLIP method is a contrastive learning approach that uses a positive approach (when images and text are related) and a negative approach (when images and text are not related) [11]. CLIP is trained to map images and text descriptions into a shared vector space with the goal of bringing relevant image-text pairs closer together in a positive approach while making irrelevant pairs more distant [9]. After training, CLIP can be used for various tasks, as CLIP embeds images and text in the same space, allowing it to handle image retrieval and image classification without image retrieval by looking at the similarity between text and images.

2. Zero-Shot Learning Approach

Zero-Shot Learning in CLIP refers to the ability of the CLIP model to perform certain tasks without the need for additional training (fine-tuning) or specific adjustments for that task [12]. In other words, CLIP can perform various new tasks just by being given a relevant text description without needing to see examples of data or images related to that task before [13]. CLIP works by converting dataset labels or candidate descriptions and comparing them with image representations through cosine similarity [14]. Predictions are made based on the highest similarity score between the image and text with a softmax function (softmax probability) [3], [15].

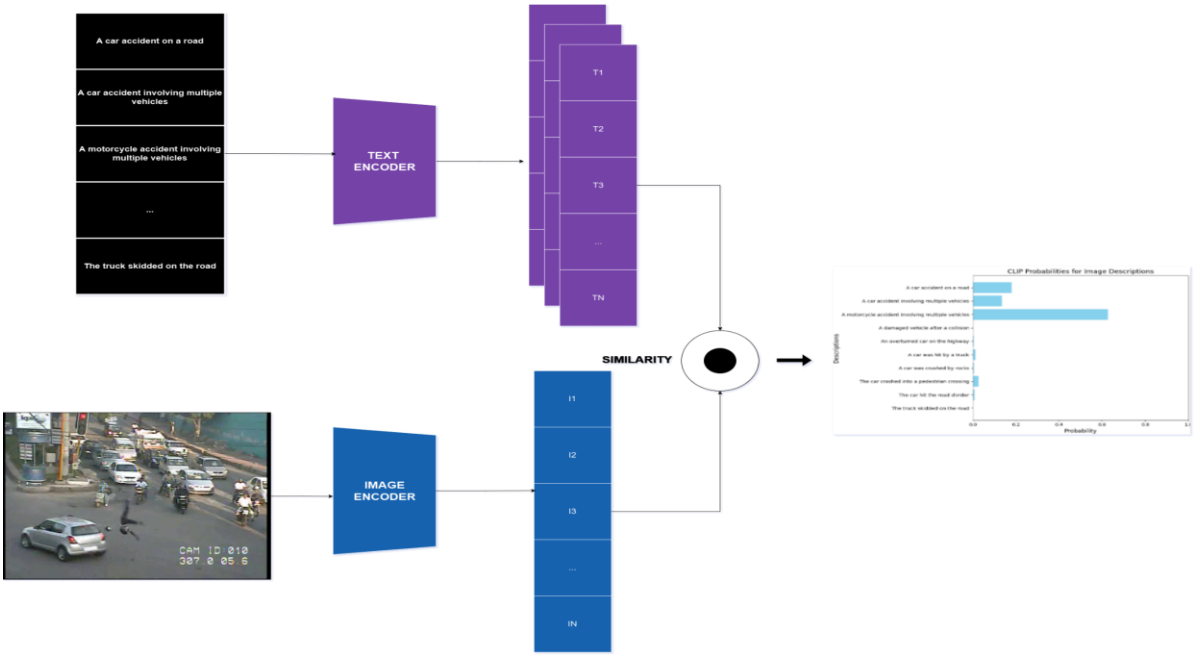

Figure 3. Zero-Shot Process in CLIP

This method allows CLIP to excel in tasks such as zero-shot image classification in multimodal data [14].

$$cosine\ similarity(v_i, t_j) = \frac{v_i \,.\, t_j}{||v_i||\ ||t_j||} \tag{1}$$

Description:
$v_i \,.\, t_j$: The dot product between vectors v and t
$||v_i||$: The length of vector v
$||t_j||$: The length of vector j

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{2}$$

Description:
x_i : the i-th element of the vector

The research flowchart can be seen in Figure 4.

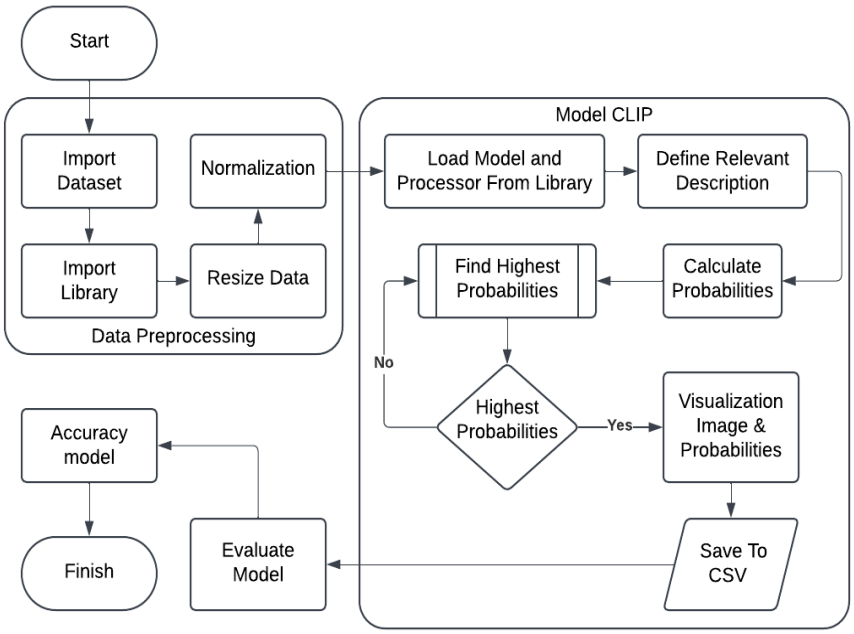

Figure 4. Research Flowchart

Figure 4 is a flowchart where all processes will be visualized in a chart sequentially according to the research process. In the preprocessing stage, dataset collection is carried out, which, in this research, was downloaded from the open-source website Kaggle [6]. Next, we proceed with importing libraries, where we will download and call the libraries we use. After that, the data is resized to standardize the input size [16], followed by normalization, which is the process of transforming image data to have uniform values on a certain scale [17]. After the preprocessing has been carried out, it is followed by the CLIP process, where the researcher must input the model and processor used.

In this research, two models are used for the main comparison by looking at efficiency in terms of time and cost. The first model used is ViT base patch 32, where the model has 12 encoder layers and 12 attention heads, and each input will be cut into patches of size 32x32 pixels and converted into embedding vectors [18]. The second model used is Vit Large patch 14, where the model has 24 transformer layers and 16 multi-head attention, and each input will be cut into patches of size 14x14 pixels [19]. Next, we define relevant description candidates, namely 10 candidates for the accident class (Table 1) and 6 candidates for the non-accident class (Table 2). The model will select the most relevant description by considering the highest similarity value produced by the model, and this value will be saved in a CSV file.

III. RESULTS AND DISCUSSION

The image data used in this research was divided into two classes, namely, accident and non-accident, with 350 images per class. Before being used for model training, the images underwent two preprocessing stages: resizing and normalization. Resizing was performed due to the

varying sizes of the input images, necessitating adjustment of the image dimensions to 224×224 pixels. This size is commonly used and can help maximize model performance by ensuring consistency of input dimensions. Next, the image data was normalized to the range $[-1, 1]$ with a mean and standard deviation of 0.5. This normalization is important to change the image pixel values that initially have a certain range so that no single pixel dominates the model's calculations during the training process. Images that have gone through the preprocessing stage are then saved in a pre-prepared storage folder.

Figure 5 shows the difference between the original and preprocessed images. The original images vary in size and resolution, while after resizing, all images are adjusted to 224×224 pixels for input consistency. Furthermore, normalization ensures that no single pixel dominates the computation during model training.

(a) (b)

Figure 5. (a) Original Data and **(b)** Data After Preprocessing

Table 5. Comparison of CLIP Models

Model	Average Score	Time (s)
ViT Base Patch32	29.58	102.64
ViT Large Patch14	23.79	1194.54

After the image data is normalized, the next step is to define the model and processor used in this research, namely ViT base patch 32 and ViT large patch 14. In its application, CLIP receives two inputs, images and text, and produces an output in the form of a similarity score or probability that indicates the extent of their compatibility. CLIP has two main encoders: a Vision Encoder based on Vision Transformer (ViT), which converts images into vector representations, and a Text Encoder based on Transformer, which converts text into vectors that can be compared with image representations. The ViT model used here is the "base" variant with 12 layers and 32x32 pixel patches, which allows image processing through the attention mechanism, and the "large" variant with 24 layers and 16 multi-head attentions with 14x14 pixel patches. In Table 5, the average scores of the two compared models are obtained, derived from the similarity score calculation process between image embeddings and text embeddings. It appears that the first

model excels in average score and execution time. In this study, only the first model was continued.

In addition to images, categorical data is needed in the form of image descriptions that serve to provide relevant labels or categories for each image in the dataset. Image descriptions such as "A car accident on a road", "A calm street with clear traffic", and others. This description data is very important in the zero-shot approach applied by CLIP, because the model needs text descriptions to calculate the similarity between images and labels. Some descriptions used for images in the accident category include "A car accident involving multiple vehicles" and "An overturned car on the highway," while for the non-accident category, descriptions such as "A calm street with clear traffic" and "A busy street with no accidents" are used. These descriptions help the model determine the relevance between images and text to achieve accurate results.

Highest Similarity Candidate

The detection process begins with image processing and the calculation of similarity between the image and relevant text descriptions. The text descriptions used include various conditions that describe the categories of images, such as "A car accident on a road" for accident images and "A calm street with clear traffic" for non-accident images. The similarity between the image and the description is calculated using cosine similarity.

For example, for Figure 6, the description with the highest probability is "A car accident on a road," which indicates that the image most closely fits the "Accident" category. Similarly, for Figure 8, the description with the highest probability is "A calm street with clear traffic," which indicates that the image most closely fits the “Non-Accident” category. The highest probability results are obtained from the calculation of similarity between the image and text representations in the CLIP latent space, with higher probabilities indicating a greater degree of match between the image and the text description.

Figure 6. Accident Input Image Sample

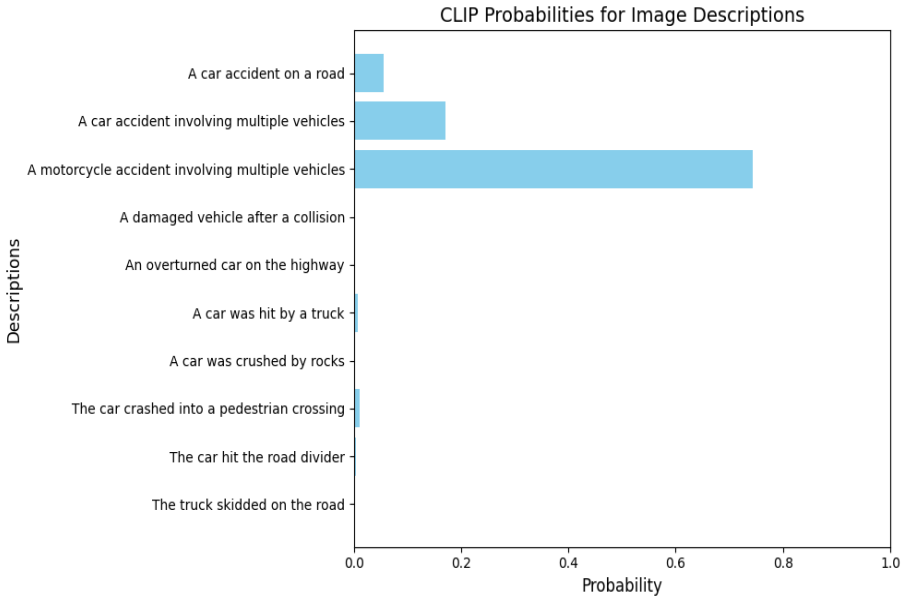

Figure 7. Softmax Probability Results for Figure 6

Figure 8. Non-Accident Input Image Sample

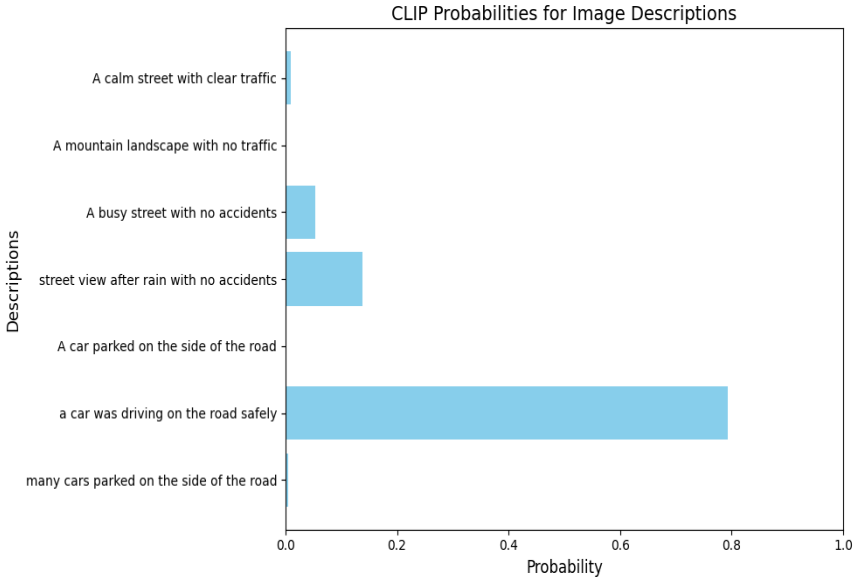

Figure 9. Softmax Probability Results for Figure 8

Cosine similarity measures the linear relationship between two vectors, in this case, between the representation of images and text in the CLIP latent space. The cosine similarity value ranges from -1 to 1, with higher values indicating greater similarity between the image and text. The similarity value is relative to a single description, not considering the similarity of the image with other descriptions.

On the other hand, softmax probability converts similarity scores into probabilities [0, 1], which facilitates model training and evaluation using the loss function. This probability indicates the model's confidence in the relevance of a specific description to the image, while considering its relationship with all existing descriptions.

Table 3. Comparison of Cosine and Softmax Probabilities for Figure 10

Description	Cosine	Softmax Probability
A car accident on a road	0.2811	0.1524
A car accident involving multiple vehicles	0.2806	0.0617
A motorcycle accident involving multiple vehicles	0.2935	0.0725
A damaged vehicle after a collision	0.2436	0.0002
An overturned car on the highway	0.2544	0.0143
A car was hit by a truck	0.2642	0.1214
A car was crushed by rocks	0.2561	0.0149
The car crashed into a pedestrian crossing	0.263	0.5484
The car hit the road divider	0.2535	0.0009
The truck skidded on the road	0.2178	0.0013

For example, Table 3 shows the results of the cosine similarity and softmax probability calculations for Figure 10. The description "The car crashed into a pedestrian crossing" has the

highest probability (0.5484), indicating that the image is most compatible with this description in the context of an accident. Conversely, other descriptions, such as "A damaged vehicle after a collision", have lower scores in both cosine similarity (0.2436) and probability (0.0002), indicating a lower compatibility with the image.

As another example, Table 4 shows the results of the cosine similarity and softmax probability calculations for Figure 11. The description "A car was driving on the road safely" has the highest probability (0.6219), indicating that the image is most compatible with this description in a non-accident context. Conversely, other descriptions, such as "A mountain landscape with no traffic", have lower scores in both cosine similarity (0.1604) and softmax probability (0.0001), indicating a lower compatibility with the image.

Figure 10. Accident Data

Table 4. Comparison of Cosine and Softmax Probabilities for Figure 10

Description	Cosine	Softmax Probability
A calm street with clear traffic	0.218	0.0241
A mountain landscape with no traffic	0.1604	0.0001
A busy street with no accidents	0.2406	0.2307
Street view after rain with no accidents	0.2197	0.0285
A car parked on the side of the road	0.2262	0.0547
A car was driving on the road safely	0.2505	0.6219
Many cars parked on the side of the road	0.2231	0.04

Figure 11. Non-Accident Data

After calculating the similarity score and softmax probability, the description with the highest probability is considered the most suitable to represent the image. The final result of this process is stored in CSV format with three columns: image path, Best Description, and Highest Probability. This allows for further evaluation and analysis of the image category detection results based on the given description.

Table 5. CLIP Vit Base Patch32 Model Accuracy Value

No.	Top	Accuracy (%)
1	Top 1	32.00
2	Top 5	95.86

The evaluation results of the CLIP model with a zero-shot classification approach in Table 5 show that the Top-1 accuracy reaches 32.00%, while the Top-5 accuracy reaches 95.86%. Top-1 accuracy reflects how often the model's top prediction matches the actual label; in this case, the model still struggles to select the correct prediction as the main answer. Meanwhile, Top-5 accuracy indicates that in most cases (almost all), the actual label appears among the top 5 predictions given by the model. This indicates that the CLIP model is quite good at understanding the relationship between text and images in general.

IV. CONCLUSION

The image processing process for classifying "Accident" and "Non-Accident" uses a CLIP-based model approach. After going through preprocessing stages, such as resizing and normalizing image data, the results show that the CLIP ViT base patch 32 model can achieve a Top 1 accuracy of 32% and a Top 5 accuracy of 95.86%. Although the Top 1 accuracy still requires further optimization, the Top 5 results indicate that the model can consistently include the correct answer within the top five predictions.

In the next research, it is recommended to optimize the dataset or fine-tune the model to improve accuracy. Then, adding more data from various accident detections can also improve the model's generalization for real-world applications. In addition, it is also possible to combine several other methods to achieve better results.

V. AUTHOR CONTRIBUTION

The authors contributed equally to the formulation of the research concept, data collection, data analysis, manuscript writing, and approval of the final manuscript for publication.

VI. CONFLIC OF INTEREST

The authors declare that there are no potential conflicts of interest related to the research, writing, or publication of this article.

ACKNOWLEDGMENT

With deep gratitude and appreciation, we would like to extend our heartfelt thanks to the various parties who have contributed to the research and writing of this article. This research would not have been possible without the help and support of many parties. We would like to express our gratitude to the Institute of Technology Sumatra and the Data Science Study Program for their adequate research facility support.

REFERENCES

[1] S. M. Prasetiyo, A. Rahmayani, and A. Melania, "Artificial Intellegence dalam kesehatan dan keselamatan kerja di bidang kelistrikan," *OKTAL: Jurnal Ilmu Komputer dan Science*, vol. 2, no. 8, 2023.

[2] A. Rezky, A. Bagir, D. Pamerean, and F. Makhrus, "Deteksi kecelakaan lalu lintas otomatis pada rekaman CCTV Indonesia menggunakan Deep Learning," *Buletin Pagelaran Mahasiswa Nasional Bidang Teknologi Informasi dan Komunikasi*, vol. 1, no. 1, 2023.

[3] M. M. Al Rahhal, Y. Bazi, H. Elgibreen, and M. Zuair, "Vision-Language models for Zero-Shot classification of remote sensing images," *Applied Sciences (Switzerland)*, vol. 13, no. 22, 2023, doi: 10.3390/app132212462.

[4] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," Feb. 2021.

[5] A. Syarwani, "Deteksi kecelakaan berbasis perubahan pola berkendara menggunakan parameter arah dan kecepatan kendaraan," Universitas Hasanuddin, 2023.

[6] C. Kay, "Accident detection from CCTV footage." [Online]. Available: https://www.kaggle.com/datasets/ckay16/accident-detection-from-cctv-footage

[7] G. Arya *et al.*, "Multimodal hate speech detection in memes using Contrastive Language-Image Pre-Training," *IEEE Access*, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3361322.

[8] Y. N. Nabuasa, "Pengolahan citra digital perbandingan metode histogram equalization dan spesification pada citra abu-abu," *J-Icon : Jurnal Komputer dan Informatika*, vol. 7, no. 1, 2019.

[9] OpenAI, "CLIP: Connecting text and images." [Online]. Available: https://openai.com/research/clip

[10] X. Pan, T. Ye, D. Han, S. Song, and G. Huang, "Contrastive Language-Image Pre-Training with knowledge graphs," in *Advances in Neural Information Processing Systems*, 2022.

[11] W. Tu, W. Deng, and T. Gedeon, "A closer look at the robustness of Contrastive Language-Image Pre-Training (CLIP)," in *Advances in Neural Information Processing Systems*, 2023.

[12] Y. Wei *et al.*, "iCLIP: Bridging image classification and Contrastive Language-Image Pre-training for visual recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2023. doi: 10.1109/CVPR52729.2023.00272.

[13] F. Pourpanah *et al.*, "A review of generalized Zero-Shot learning methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–20, 2022, doi: 10.1109/TPAMI.2022.3191696.

[14] Z. Han, Z. Fu, S. Chen, and J. Yang, "Contrastive embedding for generalized Zero-Shot learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021. doi: 10.1109/CVPR46437.2021.00240.

[15] L. Choi and R. Greer, "Evaluating vision-language models for Zero-Shot detection, classification, and association of motorcycles, passengers, and helmets," in *IEEE Vehicular Technology Conference*, 2024. doi: 10.1109/VTC2024-Fall63153.2024.10757944.

[16] N. M. Abdi and S. Aisyah, "Peningkatan kualitas citra digital menggunakan metode super resolusi pada domain spasial," *Jurnal Rekayasa Elektrika*, vol. 9, no. 3, 2011.

[17] N. R. Hanifan and A. Rahmatulloh, "Optimasi histogram equalization untuk peningkatan kualitas citra digital," *Jurnal Rekayasa Informatika*, vol. 2, no. 2, 2025.

[18] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," Jun. 2021.

[19] OpenAI, "CLIP ViT-Large-Patch14 (336px): A vision transformer model." [Online]. Available: https://huggingface.co/openai/clip-vit-large-patch14-336