Development of an Image Captioning Model to Assist The Activities of Visually Impaired Pedestrians in Urban Environments

Syahmi Sajid*[1], Agus Harjoko[2]
Master Program of Computer Science in Artificial Intelligence, FMIPA UGM, Yogyakarta, Indonesia[1]
Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia[2]
Yogyakarta, Indonesia
e-mail: syahmisajid@mail.ugm.ac.id[1], aharjoko@ugm.ac.id[2],

***Abstract*— Visual impairment is a global issue with significant impacts on the mobility and safety of individuals, especially in urban environments. Artificial intelligence solutions, such as image captioning, promise assistance for people with visual impairments to aid their daily activities. However, the field of image captioning in this context still has performance limitations. To address this, this study proposes a hybrid method combining image feature extraction from VGG16, ResNet50, and YOLO on the encoder side with LSTM and BiGRU on the decoder side to generate descriptions that have proven to enhance model performance on the Flickr8k dataset in previous research. By adapting this method to the Visual Assistance dataset and adding image augmentation processes and transfer learning to address the limitations of the dataset size, the study successfully improved the model's performance in aiding the activities of visually impaired pedestrians in urban environments. Evaluation results showed significant improvements in several evaluation metrics. Overall, this model has shown improvement compared to previous research. This study achieved a 60.53% increase in BLEU-4 score compared to previous research on the Visual Assistance dataset. Overall, this study provides a positive contribution to developing more effective and accurate solutions for visually impaired navigation users in urban environments.**

Keywords*— *Image captioning, Visually Impaired, Urban Environment, Hybrid Method, Image Augmentation, Transfer Learning

I. INTRODUCTION

Visual impairment is a significant global issue. In 2011, approximately 285 million people experienced visual impairment, and projections indicate that this number will increase to 550 million by 2050 [1]. This condition not only causes difficulties in daily mobility but also increases the risk of accidents, such as bumping into obstacles or falling [2]. Traditional aids like white canes are not always effective in detecting airborne obstacles that are not visible, such as awnings or tree branches [3]. The development of artificial intelligence has brought new solutions to assist people with visual impairments. Some solutions include using image classification techniques to identify sidewalk obstacles [4] and developing guidance systems using image segmentation [5]. However, these approaches sometimes do not provide sufficiently descriptive information to understand the surrounding environment well. One promising solution is image captioning, where text-based descriptions are generated based on information in images. This not only identifies objects but also provides richer contextual descriptions. Image captioning has the potential to improve accessibility and user experience.

Several studies have shown advancements in image captioning technology, particularly using deep learning approaches. The use of Convolutional Neural Networks (CNN) for image feature extraction and Long Short-Term Memory (LSTM) to generate descriptions has been a primary focus [6,7]. Visual attention-based approaches have also proven effective in enhancing the readability of image descriptions [6,8]. In efforts to improve the performance of models for activities of visually impaired pedestrians in urban environments, adaptation to relevant datasets such as visual assistance datasets is crucial. Previous research has shown that integration with appropriate datasets can enhance the accuracy and relevance of image descriptions, although there is still room for improvement as model performance remains subpar [9].

This research proposes using a hybrid method that has been proven to outperform previous methods on the Flickr8k dataset [10] to try to surpass previous research with the InceptionV3 – BiLSTM model using the same dataset, the Visual Assistance dataset [9]. The hybrid method to be used combines image feature extraction from VGG16, ResNet50, and YOLO in the encoder part, as well as LSTM and BiGRU in the decoder part [10]. Adapting and developing this method for the case of activities of visually impaired pedestrians in urban environments on the Visual Assistance dataset is expected to improve model performance on this dataset compared to previous research.

II. METHODS

The model design flow is shown in Figure 1, with eight main processes in this study: data collection to gather a relevant dataset for urban activities of visually impaired pedestrians, data splitting to ensure that the model can be evaluated with data it has not been trained on, image augmentation to increase dataset variation by applying techniques such as rotation, translation, and zoom on images. Text pre-processing includes cleaning and tokenizing text to ensure it is ready for use in the model, image feature extraction to extract important features from images. Modeling is carried out in two stages: training a pre-trained model with a general dataset and applying transfer learning on

data with a relevant domain. After the model is trained, generating descriptions for each test data is done to see how the model generates captions for images it has not seen before. The final stage is model evaluation, where the model's performance is assessed using specific metrics to evaluate how well the model generates accurate and relevant descriptions using several evaluation metrics such as BLEU, METEOR, ROUGE-L, and Cider.

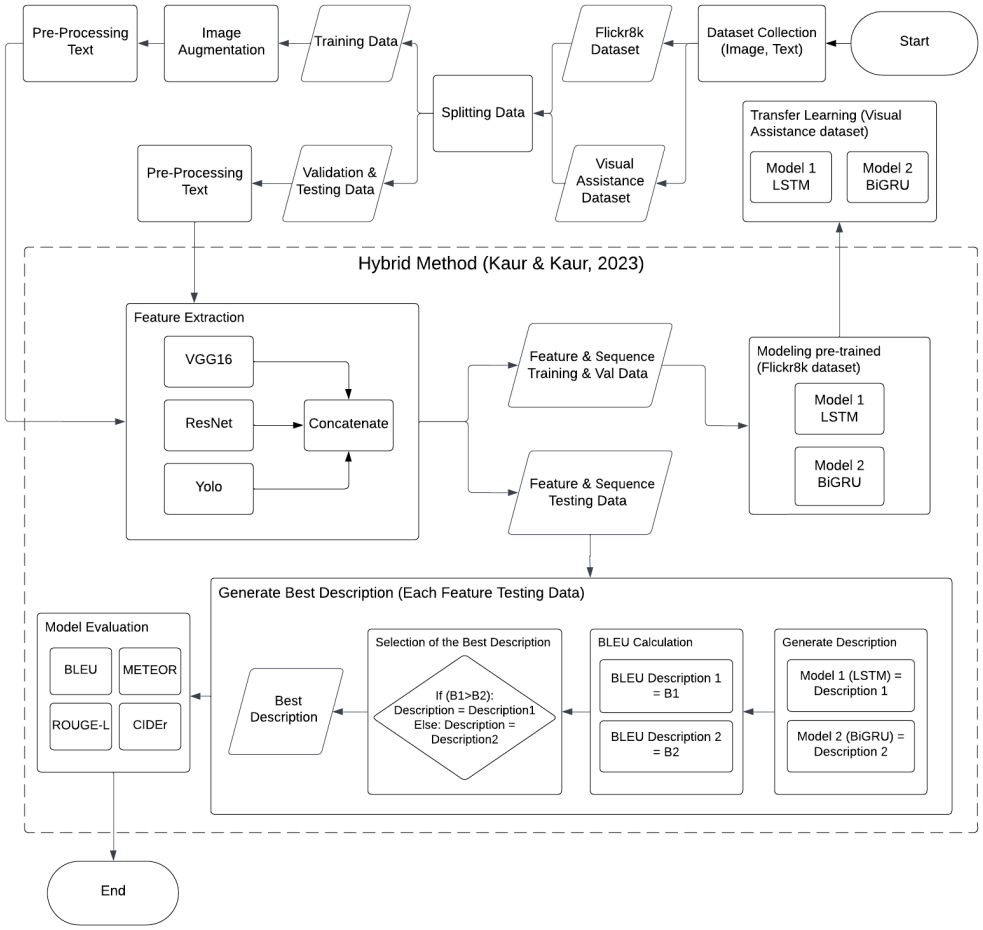

Fig. 1. Research Flowchart Flow

A. Collecting Data

This research utilizes two datasets. First, the dataset for the pre-trained model is Flickr8k from Kaggle, which consists of over 8,000 images with five description annotations per image generated by several human annotators. This aids in training the model to generate automatic descriptions. This dataset plays an important role in image captioning research, combining visual understanding with language expression. Additionally, the Visual Assistance dataset from Kaggle is used as the main dataset due to its relevance to the case of visually impaired pedestrians in urban environments. This dataset consists of 1,600 images in 21 categories, covering navigation needs, reading, object identification, and other relevant activities. Both datasets support the development of artificial intelligence systems and technologies to improve the quality of life for the visually impaired.

B. Splitting Data

Splitting the data into training, validation, and testing datasets is a crucial step in model development. The proportions for splitting the datasets are based on the research by previous researchers [9]. Previous researchers [9] conducted research using two separate datasets: the Flickr8k dataset and the Visual Assistance dataset. For the Flickr8k dataset, the data proportions used were: 75% for training data, 7.5% for validation data, and 7.5% for testing data. Assuming previous researchers [9] used the same data splitting proportions for both datasets, this serves as the reference for data splitting in this research. This study splits both the Flickr8k dataset and the Visual Assistance dataset using the same proportions as those used by previous researchers [9]. This process ensures that the developed model can learn from the training dataset, be tested with an independent testing dataset, and be validated with a validation dataset to objectively assess its performance. The results of the data splitting can be seen in Table I.

TABLE I Data Splitting Results

Dataset	Train	Validation	Test
Flickr8k	6000 Images & 30000 Descriptions (75%)	1000 Images & 5000 Descriptions (7.5%)	1000 Images & 5000 Descriptions (7.5%)
Visual Assistance dataset	1200 Images & 6000 Descriptions (75%)	200 Images & 1000 Descriptions (7.5%)	200 Images & 1000 Descriptions (7.5%)

C. Image Augmentation

This research employs several types of augmentation, including random translation in vertical and horizontal directions, random scaling up or down, and random rotation. By adjusting specified factors such as a translation factor of 2%, a scaling factor of 20%, and a rotation factor of 2%, these values were obtained from experiments conducted on each augmentation technique. Using image augmentation, the training dataset is significantly expanded, allowing the model to learn from a greater variety of the original data. Thus, the resulting model is expected to be more robust and capable of handling variations in the test data that it has not seen before. Examples of augmented images are shown in Figure 2.

Fig. 2. Example of Image Augmentation Results

D. Pre-processing Text

The initial step in text preprocessing in this research is to add the "<start>" token at the beginning of a sentence and the "<end>" token at the end of a sentence. This addition allows the system to start generating a sentence when given the "<start>" token and stop generating when it encounters the "<end>" token. The next stage is standardization, which includes several processes such as lowercasing, where the text is converted to lowercase to maintain consistency and uniform word recognition; punctuation removal, which involves deleting all punctuation marks from the text to simplify the data and reduce noise; number removal, which involves deleting all numbers from the text to focus on word-based content and eliminate numerical elements that may be irrelevant; and whitespace removal, which involves deleting excessive spaces and tidying up the text to ensure a consistent and readable format.

The tokenization process is then applied to break the text into smaller units such as words or phrases. Text to Sequence is performed to convert the text into a numerical representation that can be understood by machine learning or deep learning models. Finally, pad sequence is applied to convert the text into a numerical sequence with tokenization, indexing, and embedding, allowing machine learning models to process textual information effectively. The process flow diagram is shown in Figure 3

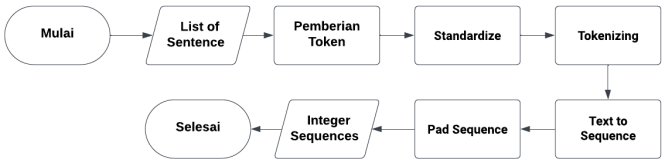

Figure 3. Flowchart of Text Pre-processing Process.

E. Image Feature Extraction

In the encoder stage, feature extraction is performed using transfer learning to generate high-quality image representations by applying pre-trained models such as VGG16, ResNet50, and YOLO separately using the pre-processed image input. Each model produces rich and informative feature representations from the images. The next step involves combining the extraction results from the three models using the Concatenate operation to create a more complete feature representation, as shown in Figure 4.

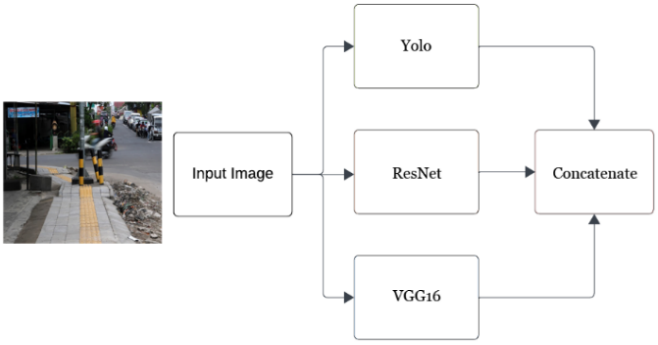

Fig. 4. Flowchart of Image Feature Extraction Process

The concatenate process allows for the merging of information from each model, creating a more comprehensive feature representation. With this step, the model can access different aspects of the image, resulting in a stronger and deeper representation. Consequently, the model becomes more effective at capturing the nuances of the image, enhancing its ability to recognize relevant patterns and produce more accurate results. The Concatenate process not only unifies information but also facilitates better integration between diverse elements, enriching the model's understanding of the image content as a whole.

F. Modeling

Image captioning in this research involves two stages: training a pre-trained model using the Flickr8k dataset and then applying transfer learning on the Visual Assistance dataset. The decoder stage is the captioning phase, which receives two inputs: the image feature extraction results and the pre-processed text descriptions. The process flow diagram for image captioning is shown in Figure 5.

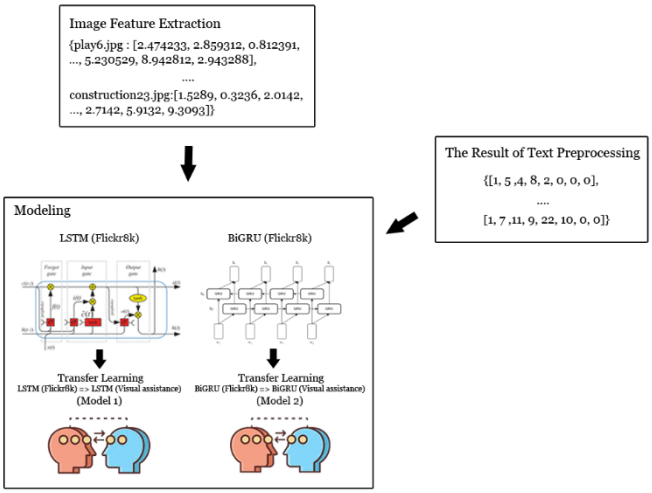

Fig. 5. Model Training Process Flow

This stage uses two main architectures in Recurrent Neural Networks (RNN), namely Bidirectional Gated Recurrent Units (BiGRU) and Long Short-Term Memory (LSTM). The BiGRU architecture distinguishes itself by using two Gated Recurrent Units (GRU) that operate simultaneously in two directions, forward and backward. This approach allows the model to respond to and understand contextual information from the text holistically, enhancing the comprehension and representation of sentences as a whole. LSTM (Long Short-Term Memory) plays a crucial role in addressing the challenges of understanding long sequences. This architecture is equipped with a feedback connection mechanism that allows it to store long-term information and recognize sequence patterns in a broader context. The combination of BiGRU and LSTM in the captioning phase is expected to create a model that not only understands the context of words continuously but also can recognize and represent the more complex relationships between objects in the image.

1) Training the Pre-trained Model: Training the pre-trained model using the Flickr8k dataset aims to "sharpen" the knowledge of the target model by transferring knowledge from the previous model to a smaller and more domain-specific dataset by updating its internal parameters. This allows the model to become more specific and accurate in specific tasks within that domain. Thus, this approach helps accelerate the learning process and improve the model's performance in completing specific tasks, according to the desired domain context.

2) Transfer Learning: In the transfer learning process, a pre-trained model created using the Flickr8k dataset is used, where the embedding layers employed to process text input are transferred to the visual assistance model specifically designed for visually impaired individuals on roadways. The illustration of the transfer learning process in this study can be seen in Figure 6.

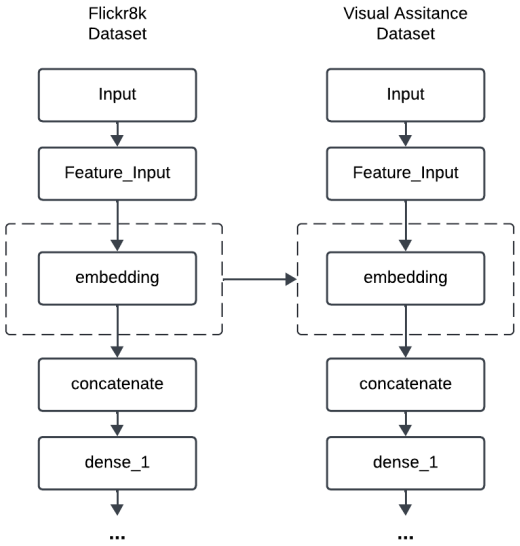

Fig. 6. Transfer Learning Illustration

Transfer learning is conducted by integrating the target domain dataset into a model that already possesses knowledge from previous learning, which is the source domain. This step allows the model to recognize and adapt to the specific context frequently encountered by visually impaired individuals in road environments, resulting in more accurate and responsive representations to their unique needs. With this transfer learning approach, it is expected that the BiGRU and LSTM models will generate more accurate and meaningful image captions, positively impacting the mobility and environmental awareness of visually impaired users on roadways.

G. Generate Descriptions for Testing Data

The description generation stage is essential for evaluating the model, with the goal of creating descriptions for each image feature in the testing data using the previously trained model. First, for each image feature in the testing dataset, generate descriptions using both previously trained models, namely model 1 LSTM and model 2 BiGRU. After successfully creating descriptions for each model, BLEU calculation is performed on both generated descriptions, named B1 for the BLEU score of model 1 and B2 for the BLEU score of model 2. Finally, the two BLEU scores are compared, and the best description is selected based on the highest BLEU score. This process is carried out for each image feature in the test data and incorporated into the best description dataset. The flow of the description generation stage is illustrated in Figure 7.

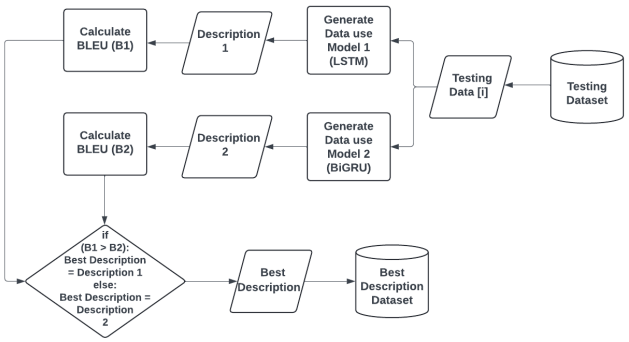

Fig. 7. Flowchart of Best Model Prediction Results on Testing Dataset

H. Model Evaluation

The best description dataset that has been collected previously is evaluated for this research by comparing it with the actual descriptions. Four main metrics will be used: BLEU, METEOR, Rouge-L, and CIDEr. BLEU (Bilingual Evaluation Understudy), a commonly used metric in machine translation evaluation, will provide an overview of how well the captions generated by the model capture important information from human references. METEOR, originally designed to evaluate the quality of machine translation, will help measure the extent to which the captions generated by the model correspond to human references, considering word, phrase, and word order matching. Rouge-L, a metric often used in natural language processing, will help evaluate the longest word sequence similarity between the model's captions and references, providing deeper insights into content similarity aspects. Meanwhile, CIDEr, specifically developed for image captioning evaluation, will enable the assessment of consistency and diversity in image descriptions, considering the consensus among multiple captions generated by humans. By using the combination of these four metrics, it is hoped to obtain a more comprehensive insight into the quality of captions generated by the model and identify areas where the model can be improved to enhance the navigation experience for visually impaired users in the highway environment.

III. Results And Discussion

A. Model Inference Results

The results of the model training will yield model weights, training configurations, and tokenizer layers stored in local storage. Model weights include the updated weights learned during the training process, enabling the model to recognize patterns in the data. Training configurations include the parameters used during training, such as learning rate, number of epochs, and batch size. The tokenizer layer is responsible for converting text into a format that can be processed by the model. Once the model is well-trained, the results of the model training are used to generate description predictions with input images. This inference process involves providing input images to the model, which then outputs text descriptions based on the patterns learned during training. In other words, the model is capable of analyzing image content and providing relevant and accurate textual explanations. Examples of prediction results using test data can be seen in Figure 8.

Fig. 8. Example of Model Prediction Results on Test Data

B. Comparison of Evaluation Results among Models

The first experiment involves comparing the performance of various models by evaluating each individual model and combinations of multiple models. In this model comparison experiment, these models do not use image augmentation and transfer learning processes, as the goal is to evaluate the best performance of each individual model and combination of models. By comparing these models, the research can identify the strengths and weaknesses of each approach and determine whether the hybrid method [10] performs better than using individual models or combinations of only 2 methods. The best performing model will be used for the proposed model.

The worst performance was observed with the YOLOv4 – BiGRU and YOLOv4 – LSTM models. This is because YOLOv4 is designed for object detection, and YOLOv4 tends to produce features focused on detected objects in images. According to previous researchers [11], using class probabilities to detect objects and bounding boxes as initial vectors, without allowing the LSTM model to review the image information, results in poor text description performance. This is mainly because the information in the last fully connected layer of the YOLO model only represents class probabilities for detected objects and their bounding boxes, which is not rich enough. Object detection-based YOLO models perform better when combined with convolutional feature-based models, as shown in the study by previous researchers [12], which combined YOLOv4 and Xception for feature extraction.

The ResNet50 – BiGRU and VGG16 – BiGRU models show similar results, indicating that feature extraction using ResNet50 and VGG16 has almost the same capability in image captioning tasks. The ResNet50 – LSTM and VGG16 – LSTM models achieved better scores compared to using BiGRU, suggesting that using LSTM as the language model can provide better results in generating image descriptions. The second-best results were obtained from the combination of encoder models VGG16 and ResNet50, with BLEU-4 of 0.3622, METEOR of 0.8326, Rouge-L of 0.5883, and CIDEr of 2.3422. The performance of the ResNet and YOLOv4 encoder combination model is also quite good and has similar scores to the VGG16 and ResNet50 combination, which ranks second. This shows that the poor performance of YOLOv4 in individual model evaluation does not affect its effectiveness when combined with convolutional models, as explained by previous researchers [12]. Overall, the performance results of model combinations are better than using individual models alone, indicating that combining models can achieve better performance than using just one method.

The best performance was achieved using the hybrid model combination proposed by previous researchers [10] which integrates the three models: VGG16, ResNet50, and YOLOv4, combined with LSTM and BiGRU. This model provided the best performance compared to other model combinations, achieving BLEU-4 of 0.3649, METEOR of 0.8253, Rouge-L of 0.5875, and CIDEr of 2.3822. These values indicate that this model can generate more accurate and relevant descriptions for the given images. The comparison of evaluation results among individual models is shown in Table II.

TABLE II Comparison of Evaluation Results between Models

Model	Matriks Evaluasi						
	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**Rouge-L**	**CIDer**
YOLOv4 – BiGRU	0.2348	0.1531	0.0894	0.0497	0.2714	0.2166	0.3812
YOLOv4 – LSTM	0.2334	0.1578	0.1076	0.0765	0.2866	0.2386	0.4400
ResNet50 – BiGRU	0.4455	0.3793	0.3143	0.2394	0.6203	0.4427	1.6717
VGG16 – BiGRU	0.4519	0.3900	0.3270	0.2484	0.6489	0.4492	1.6410
VGG16 – LSTM	0.5611	0.4967	0.4250	0.3377	0.7849	0.5786	2.1291
ResNet50 – LSTM	0.5776	0.5165	0.4456	0.3506	0.8215	0.5773	2.3166
(VGG16+ Resnet50) - (LSTM+BiGRU)	0.5858	0.5299	0.4578	0.3622	0.8326	0.5883	2.3403
(VGG16+ YOLOv4) - (LSTM+ BiGRU)	0.5587	0.5003	0.4296	0.3385	0.7958	0.5576	2.2569
(Resnet50+ YOLOv4) - (LSTM+ BiGRU)	0.5826	0.5300	0.4572	0.3592	0.8243	0.5841	2.3332
Hybrid (Kaur & Kaur, 2023) (VGG16+ Resnet50+ YOLOv4) - (LSTM+ BiGRU)	**0.5928**	**0.5302**	**0.4582**	**0.3649**	**0.8253**	**0.5875**	**2.3822**

C. Comparison Results of Model Evaluation using Augmentation

The second experiment in this research involves adding image augmentation processes to the hybrid method to enrich the data in the Visually Assistance dataset. This experiment only uses the Visually Assistance dataset without transfer learning from the Flickr8k dataset, as it aims to see whether adding image augmentation processes can improve model performance and to choose the best combination of image augmentations to implement in the final proposed model. The best result among the use of individual augmentations was obtained with zooming, achieving a BLEU-4 score of 0.3743, compared to translation with a BLEU-4 score of 0.3704 and rotation with a BLEU-4 score of 0.3674. Augmentation techniques that only use translation and rotation can cause noise such as redundancy due to image shifts or rotations, as shown in Figure 2. According to previous researchers [13], augmentation techniques like translation and rotation can generate data variations that might not be fully representative of the original data, potentially adding "noise" and interfering with the model's ability to recognize relevant patterns.

Translation and rotation augmentation techniques are better combined with techniques like zooming to avoid noise, as the image will be enlarged after shifting or rotating. This can be seen in the combination results of augmentations (translation + zooming) with a BLEU-4 score of 0.3767 and (translation + rotation) with a BLEU-4 score of 0.3765, which show performance improvements compared to using translation or rotation alone. However, the combination of (translation + rotation) produced the worst results due to increased noise without image enlargement, achieving a BLEU-4 score of 0.3442. The combination of all three augmentations resulted in a slightly worse outcome than the combination of two augmentations using zooming, with a BLEU-4 score of 0.3743. This is because zooming alone was not sufficient to avoid noise from both translation and rotation simultaneously. Overall, image augmentation techniques successfully improved the model's performance on this dataset. The best image augmentations identified in this experiment are the combinations of (Rotation + Zooming) and (Translation + Zooming), as they produced similar results with a difference of 0.002. The comparison of experimental results using various augmentation techniques is shown in Table II.

TABLE III Comparison of Experimental Results using Various Augmentation Techniques

Model	Matriks Evaluasi						
	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**Rouge-L**	**CIDer**
Hybrid	0.5928	0.5302	0.4582	0.3649	0.8253	0.5875	2.3822
Hybrid + Augmentasi (Translasi)	0.5783	0.5246	0.4585	0.3704	0.8207	0.5775	2.3132
Hybrid + Augmentasi (*Zoom*)	0.5938	0.5369	0.4677	0.3743	0.8243	0.5782	2.3668
Hybrid + Augmentasi (Rotasi)	0.5767	0.5188	0.4533	0.3674	0.8064	0.5747	2.3468
Hybrid + Augmentasi (Translasi + *Zoom*)	**0.5815**	**0.5288**	**0.4632**	**0.3767**	**0.8396**	**0.5909**	**2.3813**
Hybrid + Augmentasi (Translasi + Rotasi)	0.5665	0.5101	0.4391	0.3442	0.8178	0.5770	2.2924
Hybrid + Augmentasi (Rotasi + *Zoom*)	**0.5907**	**0.5363**	**0.4728**	**0.3765**	**0.8326**	**0.5667**	**2.3238**
Hybrid + Augmentasi (Translasi + *Zoom* + Rotasi)	0.5840	0.5294	0.4606	0.3743	0.8476	0.5731	2.4022

D. Comparison Results of Model Evaluation using Transfer Learning

The third experiment in this research involves adding a transfer learning process by utilizing embedding weights from a previously trained model (pre-trained model). This experiment does not use image augmentation processes because the aim is to see whether the addition of transfer learning techniques can improve the model's performance on this dataset by comparing the performance of models without and with the transfer learning process. In this study, the pre-trained model uses the Flickr8k dataset as a foundation to train the new model on a specific domain, namely the case of activities of visually impaired pedestrians in urban environments. Transfer learning is a commonly used technique in machine learning, involving the use of a model previously trained on a large dataset. The weights from the pre-trained model are used to train the new model on a smaller and more specific domain dataset.

By using transfer learning, the model can "fine-tune" the knowledge from the previous model on the smaller dataset and specific domain by updating its internal parameters. This allows the model to become more specific and accurate in tasks specific to that domain. In the context of this research, the comparison of model evaluation results using transfer learning can provide insights into how effective this approach is in improving the model's performance in recognizing and understanding images from the Flickr8k dataset.

The results of the second experiment in this research prove that adding the transfer learning process can improve the model's performance. The model using transfer learning achieved a BLEU-4 score of 0.3823, while the model without transfer learning only achieved a BLEU-4 score of 0.3649. Using transfer learning techniques can enhance model performance on the target small dataset domain by leveraging the knowledge contained in the source domain with a larger dataset [14]. The comparison of models with and without the transfer learning process is shown in Table IV.

TABLE IV Comparison of Experimental Results Using Various Augmentation Techniques

Model	**Matriks Evaluasi**						
	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**Rouge-L**	**CIDer**
Hybrid	0.5928	0.5302	0.4582	0.3649	0.8253	0.5875	2.3822
Hybrid + Transfer Learning	**0.6012**	**0.5479**	**0.4815**	**0.3823**	**0.8529**	**0.6014**	**2.5108**

E. Comparison with Previous Research

Based on the previous 3 experiments, the final experiment is to compare the proposed model, which is the hybrid method enhanced with image augmentation (translation/rotation + zoom) and transfer learning, with previous research conducted by previous researchers [9] which used the InceptionV3-BiLSTM architecture enhanced with adaptive attention. Their study achieved a BLEU-1 score of 0.702 and a BLEU-4 score of 0.266. Meanwhile, in this study, the proposed hybrid model (Kaur & Kaur, 2023) with added image augmentation and transfer learning has proven to improve model performance for this case. The proposed model achieved a highest BLEU-4 score of 0.427, showing a 60.53% improvement from the previous study, although it obtained a BLEU-1 score of 0.632, which is still lower than the previous study's BLEU-1 score of 0.702.

This discrepancy occurs because BLEU-1 measures the similarity between unigrams (single words) in the predicted text and the reference text, while BLEU-4 measures the similarity between n-grams (sequences of n words, in this case, 4 words) in the predicted text and the reference text. A high BLEU-1 score but a low BLEU-4 score can occur when the model overfits on unigram data. In this architecture, the addition of image augmentation and transfer learning helps prevent overfitting on unigram data.

The best evaluation metric for BLEU is using the maximum n-gram sequence up to 4 (BLEU-4) [15]. Thus, overall, the proposed model is superior compared to previous research because it outperforms in BLEU-4, BLEU-3, and BLEU-2, assuming the same pre-processing steps as the previous study. The comparison results with previous research can be seen in Table V.

TABLE V Comparison of Experimental Results using Various Augmentation Techniques

Model	**Matriks Evaluasi**						
	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**Rouge-L**	**CIDer**
Sharma *et al* (2022) InceptionV3-BiLSTM + *Adaptive Attention*	**0.702**	0.491	0.359	0.266	-	-	-
Hybrid (Kaur & Kaur, 2023)	0.593	0.530	0.458	0.365	0.825	0.588	2.382
Hybrid (Kaur & Kaur, 2023) + Augmentasi Gambar + *Transfer learning*	0.632	**0.585**	**0.523**	**0.427**	**0.895**	**0.657**	**2.751**

IV. CONCLUSIONS

This research successfully developed the hybrid method proposed by previous researchers [10] with adjustments for the case of visually impaired pedestrian activities in urban environments, using a dataset specifically designed for the visually impaired, namely the Visual Assistance dataset. The implementation of image augmentation and transfer learning contributed to the improvement of model performance and addressed the limitations of dataset size by enriching the training data and leveraging pre-trained models already trained on larger datasets.

Using the Visual Assistance dataset, this research achieved an improved BLEU-4 score of 0.427, higher than previous researchers [9] with a BLEU-4 score of 0.266. Overall, this

research shows a satisfactory performance on the Visual Assistance dataset with the following evaluation results: BLEU-1 of 0.632, BLEU-2 of 0.585, BLEU-3 of 0.523, BLEU-4 of 0.427, METEOR of 0.895, ROUGE-L of 0.657, and CIDEr of 2.751.

V. Future Work

Based on the results of this research, several recommendations can be made for future studies and further development, First, it is recommended to use a larger dataset beyond Flickr8k, such as MSCOCO or Flickr30k, as the foundation for the model. Larger datasets not only provide more data for training the model but also offer a broader variety of images and annotations, which can ultimately enhance the model's generalization capabilities.

Second, exploring more advanced image augmentation techniques, such as geometric transformations and color blurring, should be considered. More complex image augmentation techniques can help the model become more robust to variations in input data and reduce the likelihood of overfitting. Refining the model architecture through experimentation with more advanced models like Transformers can improve performance. The use of Transformer models, which have proven highly effective in various natural language processing and computer vision tasks, can aid in capturing more complex relationships between elements in images and text. Finally, further research could include exploring more efficient training methods, such as transfer learning or multi-task learning, which can leverage knowledge from related tasks to enhance the model's performance in image description tasks.

References

[1] Bourne, R, R, A., Flaxman, S, R., Braithwaite, T., Cicinelli, M, V., Das, A., 2017, Vision Loss Expert Group. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: A systematic review and meta-analysis. Lancet Glob. Health 2017, 5, e888–e897.

[2] Manduchi, R., PhD* Kurniawan, S. 2011, Mobility-Related Accidents Experienced by People with Visual Impairment, University of Santa Cruz Santa Cruz, CA, pp.10

[3] Sáez, J.M., Escolano, F., Lozano, M.A. Aerial Obstacle Detection With 3-D Mobile Devices. IEEE J. Biomed. Health Inform. 2015, 19, 74–80.

[4] Ahmed, F., and M. Yeasin, 2017, Optimization and Evaluation of deep architectures for ambient awareness on a sidewalk, in Neural Networks (IJCNN), 2017 International Joint Conference on. IEEE, pp. 2692–2697.

[5] Liu, L., Wang, Y., and Zhao, H., 2019, "An Image Segmentation Method for the blind sidewalks recognition by using the convolutional neural network U-net," 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Chongqing, China, pp. 1-4, doi: 10.1109/ICSIDP47821.2019.9172970.

[6] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio. Y., 2015, Show, attend and tell: Neural Image caption generation with visual attention, in: Proceedings of the International conference on machine learning, pp. 2048–2057.

[7] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T., 2017, SCA-CNN: Spatial and channel-wise attention in convolutional networks for Image captioning, in: Proceedings of the IEEE Conference on Computer vision and Pattern Recognition (CVPR), pp. 6298–6306.

[8] You, Q., Jin, H., Wang, Z., Fang, C., Luo, J 2016 Image captioning with semantic attention. In proceedings of the IEEE conference on computer vision and pattern recognition, 4651–4659

[9] Sharma, D., Dhiman, C., and Kumar, D., 2022, "Automated Image Caption Generation Framework using Adaptive Attention and Bi-LSTM," 2022 IEEE Delhi Section Conference (DELCON), New Delhi, India, pp. 1-5.

[10] Kaur, M., Kaur, H., 2023, An efficient deep learning based hybrid model for image caption generation, International Journal of Advanced Computer Science and Applications, 14(3)

[11] Han, M., Chen, W., & Moges, A.D., 2019, Fast image captioning using LSTM. Cluster Comput 22 (Suppl 3), 6143–6155. https://doi.org/10.1007/s10586-018-1885-9

[12] Al-Malla, M.A., Jafar, A., & Ghneim, N., 2022, Image captioning model using attention and object features to mimic human image understanding. J Big Data 9, 20. https://doi.org/10.1186/s40537-022-00571-w.

[13] Shorten, C., & Khoshgoftaar, T. M. 2019, A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 6(1).

[14] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. 2020, A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1), 43-76.

[15] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J., 2001, BLEU, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia.