Improvement of Vision Transformer Using Word Patches

Ayato Takama[a], Sota Kato[b], Satoshi Kamiya[c] and Kazuhiro Hotta[d]
Meijo University, 1-501 Shiogamaguchi, Tempaku-ku, Nagoya 468-8502, Japan

Keywords: Classification, Vision Transformer, Visual Words, Word Patches, Trainable.

Abstract: Vision Transformer achieves higher accuracy on image classification than conventional convolutional neural networks. However, Vision Transformer requires more training images than conventional neural networks. Since there is no clear concept of words in images, we created Visual Words by cropping training images and clustering them using K-means like bag-of-visual words, and incorporated them into Vision Transformer as "Word Patches" to improve the accuracy. We also try trainable words instead of visual words by clustering. Experiments were conducted to confirm the effectiveness of the proposed method. When Word Patches are trainable parameters, the accuracy was much improved from 84.16% to 87.35% on the Food101 dataset.

1 INTRODUCTION

Transformer(Vaswani et al., 2017) was proposed in the field of natural language processing and outperforms conventional methods. By dividing sentences into words and using Multi-Head Attention, the method uses the relationship between words effectively. Vision Transformer(Dosovitskiy et al., 2020) uses Transformer instead of a convolutional neural network (CNN) which is commonly used in the field of image recognition. However, unlike natural language processing, we believe that the potential of Transformer is not used well in the current Vision Transformer because there is no concept of words in images.

Therefore, in this paper, we created Visual Words by cropping regions in images and clustering them using K-means like bag-of-visual words. If the visual words are used as the concept of words in the Vision Transformer well, the accuracy of the Vision Transformer is expected to be improved because it can learn and classify images using what patterns appear in the training samples and how similar they are to those patterns. The proposed method incorporates Visual Words into the network as "Word Patches".

We also evaluate the case that Word Patches themselves are trainable. By making Word Patches trainable parameters, it is possible to learn the model while taking into account the relationship between Word Patches and the outputs of conventional Vision Transformer. It is expected that Word Patches would adapt to the features that make it easier for Vision Transformer to learn.

In experiments, we compared the proposed method with the conventional Vision Transformer on the Food101 dataset(Bossard et al., 2014) and the CIFAR100 dataset(Bossard et al., 2014). Experimental results showed that the proposed method improved the accuracy in comparison with the original Vision Transformer. We also confirmed that the accuracy was improved when the Word Patches were used as trainable parameters.

The paper is organized as follows. Section 2 describes related works. Section 3 explains the proposed method. Section 4 presents experimental results. Finally, Section 5 describes our summary and future works.

2 RELATED WORKS

2.1 Vision Transformer

Vision Transformer uses Transformers instead of CNNs which are commonly used in image recognition, and achieves the state-of-the-art accuracy in image classification when we trained it on large amounts of training data. The overview of Vision Transformer is shown in Figure 1. To input an image to the Transformer, the Vision Transformer converts two-

[a] https://orcid.org/0000-0001-7255-1328
[b] https://orcid.org/0000-0003-0392-6426
[c] https://orcid.org/0000-0002-7057-3280
[d] https://orcid.org/0000-0002-5675-8713

Takama, A., Kato, S., Kamiya, S. and Hotta, K.
Improvement of Vision Transformer Using Word Patches.
DOI: 10.5220/0011732900003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 731-736
ISBN: 978-989-758-634-7; ISSN: 2184-4321

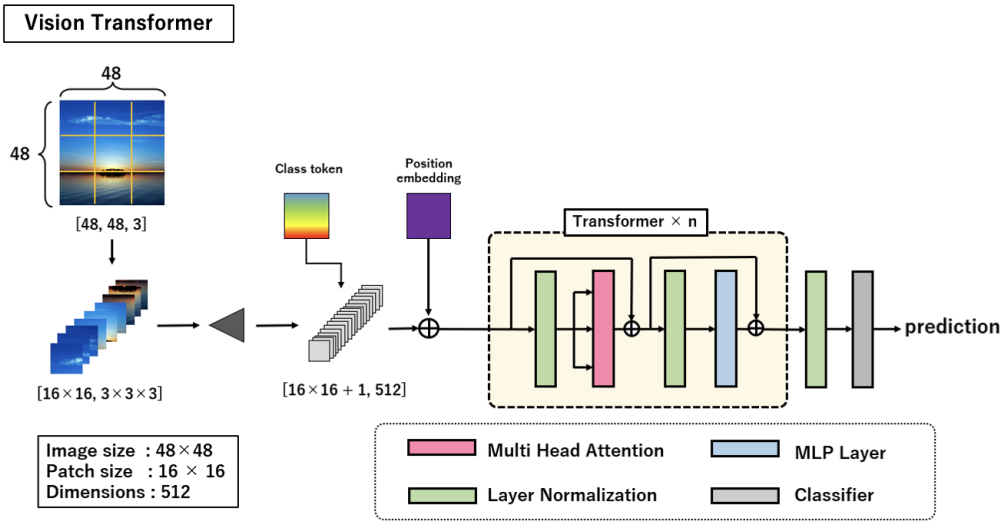

Figure 1: The architecture of Vision Transformer. To input an image to the Vision Transformer, the image is divided into patches, vectorized, linearly projected, and then positional embeddings are added. Vectors are then fed into the Transformer Encoder, which is composed of Multi-Head Attention and Layer Normalization. Classification is performed through MLP of class tokens.

dimensional images into the set of one-dimensional sequence data and performs linear projection. A class token is then introduced to the beginning of the sequence data to enable image classification. Next, position information is added to identify where the patches are located in the image. The feature vectors are then fed into the Transformer Encoder which is composed of Multi-Head Attention and Layer Normalization, and the MLP of class token is used to classify the image. The problem is that it is difficult to obtain high accuracy when a large amount of image data is not used in training process. In recent years, many methods that improve the Vision Transformer have been studied. For example, SegFormer(Xie et al., 2021), Swin Transformer(Liu et al., 2021), MetaFormer(Yu et al., 2022), and TransUnet(Chen et al., 2021) are well-known for semantic segmentation, TrackFormer(Meinhardt et al., 2022), MOTR(Zeng et al., 2022), and TransMOT(Chu et al., 2021) for object tracking, and DETR(Carion et al., 2020) for object detection.

In recent years, there has been a lot of methods (Guibas et al., 2021; Sethi et al., 2021; Tan et al., 2021) that improve token mixing, but there has been no method focusing on the words. We consider that words are important because original transformer was proposed in natural language processing and used words effectively. Thus, we use Visual Words that obtained by clustering or trainable parameters in the Vision Transformer. If the proposed Visual Words are used effectively, we would use the potential performance of the Transformer, and the classification accuracy would be improved. This is because it can learn and classify by using the information what patterns appear in each class and how similar they are to each other. This paper aims to improve the classification accuracy by incorporating Visual Words in Vision Transformer.

2.2 Visual Words

Before deep learning, image classification methods mainly used Visual Words obtained by clustering of local features in images[7]. Since images do not have the concept of ”words” like natural language processing, we made ”visual words” from image patches in training images through clustering. The overview to obtain visual words is shown in Figure 2. (1) Images are divided into patches as in the conventional Vision Transformer. (2) All patches are clustered by K-means. (3) The average of each cluster is used as ”Visual Words”.

2.3 Object Queries

DETR is the first object detection method using the Transformer. The object queries in the DETR are used as queries for Multi-Head Attention in the Transformer decoder to improve accuracy. In the proposed method, Visual Words are fed into the Vision Transformer like Object Queries. But, unlike DETR, the outputs of the original Vision Transformer are used as a queries in the Multi-Head Attention of the Mix Transformer. The output of Word Patches Trans-

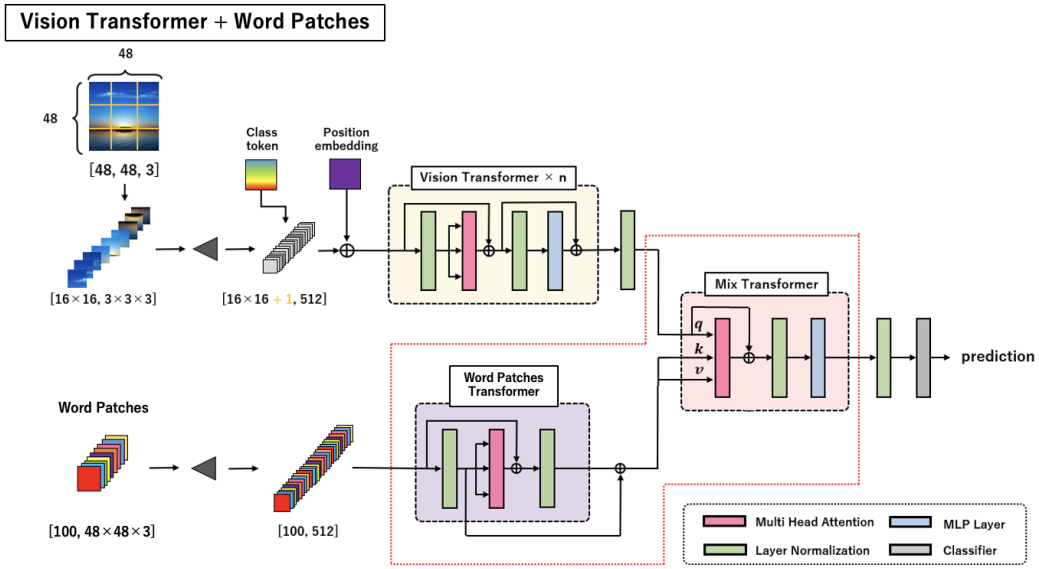

Figure 2: The architecture of the proposed method. Visual Words are fed into the Word Patches Transformer, and the outputs are mixed with the Encoder's outputs of the conventional Vision Transformer in Mix Transformer. The output of the Mix Transformer was then used for classification using Class tokens.

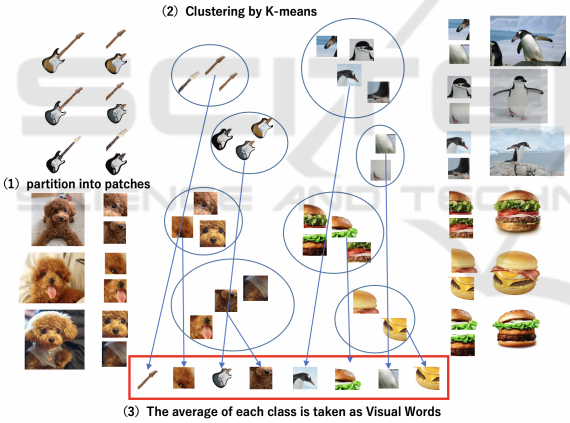

Figure 3: Visual Words. The training image is divided into patches, clustered by K-means, and the average vector of each cluster is used as a Visual Word

former, that Visual Words are the inputs, are used as keys and values in the Multi-Head Attention of the Mix Transformer. This allows the proposed method to proceed learning while taking into account the relationship between Word Patches and the outputs of conventional Vision Transformer. We expect that the relationship with Word Patches improves the classification accuracy.

3 PROPOSED METHOD

The proposed method uses the original Vision Transformer as a baseline, and we add Word Patches Transformer and Mix Transformer to the original one newly. The relationship between Word Patches and patches in an input image are used to improve the classification accuracy.

Figure 2 shows the overview of the proposed method. The upper network shows the conventional Vision Transformer, and the processing of the network is the same as conventional one. The lower network shows the Word Patches Transformer added newly. The Visual Words, that obtained by K-means or are trainable parameters, are used as Word Patches in Word Patches Transformer. Word Patches Transformer is also consists of Multi-Head Attention and Layer Normalization. The outputs of both Transformers are fed into Mix Transformer which also consists of Multi-Head Attention, Layer Normalization and MLP. The outputs of the conventional Vision Transformer are used as the queries and the outputs of the Word Patches Transformer as the keys and values. This is because the outputs of the conventional Vision Transformer are from the patches in an input image and the outputs of Word Patch Transformer are the support to improve the accuracy.

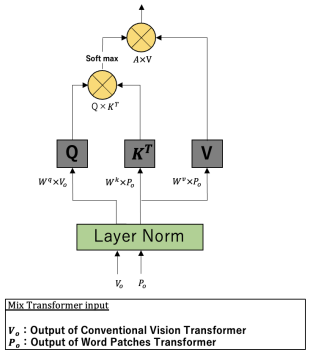

Figure 4: Mix Transformer.

Figure 4 and Equation (1) shows the Mix Transformer.

$$\begin{aligned} Z(Q,K,V) &= Concat(head_1,...,head_{Nh}) \\ head_i &= softmax\left[\frac{Q_i(K_i)^T}{\sqrt{Ch}}\right]V_i \\ &= HEREHEREHEREA_iV_i \end{aligned} \quad (1)$$

where $Z(Q,K,V)$ is the output of Multi Head Attention in Mix Transformer, A_i is the attention map. Query, Key, Value are computed as

$$\begin{aligned} Q &= W^qV_o \\ K &= W^kP_o \\ V &= W^vP_o \end{aligned} \quad (2)$$

where V_o indicates the output of the conventional Vision Transformer and P_o indicates the output of the Word Patches Transformer. W^q, W^k and W^v are the 1×1 convolution. Thus, attention map is generated from the similarity between Query(Q) from Conventional Vision Transformer and Key(K) from Word Patches Transformer.

By using the Mix Transformer, we can use the relationship between the Transformer's outputs of patches in an input image and Transformer's outputs of Word Patches. It is possible to train the model by using what features appear in each class and how similar they are. Finally, the class token which is the output of the Mix Transformer is used for classification through MLP.

4 EXPERIMENTS

In experiments, we use the Food101 dataset and the CIFAR100 dataset. In section 4.1, we explains both datasets and implementation details. Experimental results on Food101 dataset and CIFAR100 dataset are shown in section 4.2 and 4.3, respectivelly.

4.1 Dataset and Implementation Details

The Food101 dataset contains 101,000 images of 101 classes, 75,750 for training and 25,250 for testing. In this paper, the training set is divided into 70,700 images for training (700 images for each class) and 5050 images for validation (50 images for each class) because there is no validation set. The original 25,250 test images were used for evaluation. Although original image size is random, we resize it to 384x384 pixels.

CIFAR100 dataset consists of 60,000 images of 100 classes. 50,000 images are used for training and 10,000 images are used for test. Although original image size is 32x32 pixels, we resize it to 224x224 pixels.

We used Vision Transformer pre-trained on ImageNet1k dataset. Word Patches Transformer and Mix Transformer were not pre-trained. In this paper, the number of Word Patches was set to 101 for Food101 and 100 for CIFAR100 to match the number of classes.

Minibatch size was set to 16 and we set the number of epochs to 50. AdamW($weight_d = 0.3$) and Cosine scheduler were used in optimization. Classification accuracy was used as a evaluation measure.

Table 1: Results of classification on the Food101 dataset. "baseline" is the conventional Vision Transformer, ours is the proposed method with Word Patches by clustering, and ours(trainable) is the result when Word Patches is the trainable parameter.

	Food101	
	top-1	top-5
baseline	84.16	95.62
ours	85.99	96.53
ours(trainable)	87.35	97.07

4.2 Results on Food101 Dataset

Table 1 shows the comparison results of the proposed method with the conventional Vision Transformer on Food101 dataset. Baseline indicates a conventional Vision Transformer, ours indicates the proposed method with Patch Words by clustering, and ours(trainable) indicates the proposed method with trainable Word Patches. The top-1 shows the standard accuracy whether class with the highest probability is correct. The top-5 is the accuracy if the true class is included in the top five classes with the highest probability. As we can see from the Table, the usage of Word Patches brings higher accuracy than the conventional Vision Transformer. We also evaluated our method when Word Patches were used as trainable pa-

rameters. As shown in Table 1, we confirmed that the best accuracy was obtained when trainable Word Patches are used.

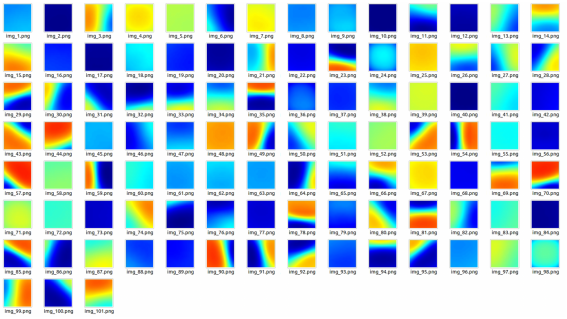

Figure 5: Word Patches by clustering for Food101 dataset.

Figure 5 shows the Word Patches created from training images. We see that variety of Word Patches are included. However, since they are average vectors of clusters, they are rough patterns rather than detailed patterns. The accuracy improvement suggests that the usage of these Word Patches have provided hints for classifying classes.

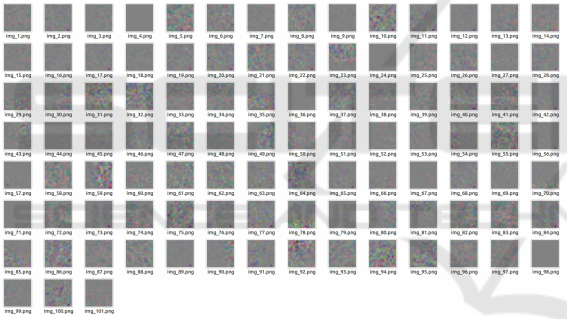

Figure 6: Trainable Word Patches for Food101 dataset.

Figure 6 shows the visualization results of the trainable Word Patches. We see that detailed features are captured though Word Patches obtained by clustering were only rough patterns. This improved the accuracy.

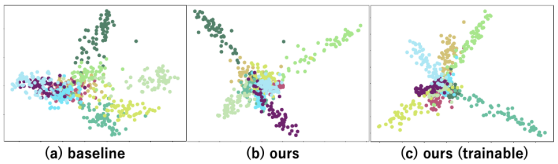

Figure 7: The distribution of class tokens for 10 classes in the Food101 dataset. The different classes are represented by different colors.

Distribution of class tokens is shown in Figure 7. This shows the distribution of 10 classes in the Food101 dataset because the distribution of 101 classes are too crowd. The different classes are represented with different colors. (a) shows the distribution of baseline, (b) shows our method with Word Patches obtained by clustering and (c) shows our method with trainable Word Patches.

Figure 7 shows that (a) has many overlap between classes. In contrast, the proposed methods shown as (b) and (c) have less within-class variation, which allows us to judge the classification correctly. Therefore, the proposed method is more accurate than standard Vision Transformer. When we compare (b) with (c), the distribution among classes in (c) is wider than that in (b). This allows us to improve the accuracy when Word Patches are trainable.

4.3 Results on CIFAR100 dataset

Table 2: Results of classification on the CIFAR100 dataset. baseline is the conventional Vision Transformer, ours is the proposed method with Word Patches by clustering, and ours(trainable) is the result when Word Patches is the trainable parameter.

	CIFAR100	
	top-1	top-5
baseline	92.53	98.91
ours	92.97	98.97
ours(trainable)	93.27	99.04

Table 2 compares the results of the proposed method with the conventional Vision Transformer on the CIFAR100 dataset. The proposed method also achieved higher accuracy than the conventional Vision Transformer by using Word Patches. In addition, we conducted the experiment when Word Patches were used as trainable parameters. We confirmed that the accuracy was also improved by using trainable parameters.

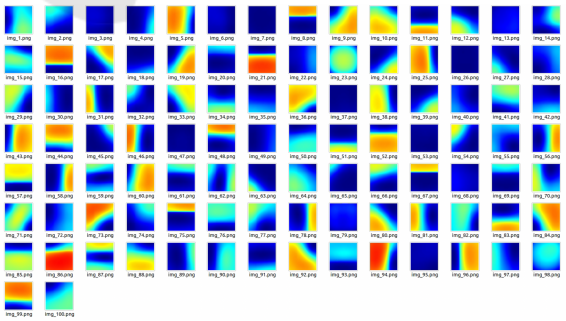

Figure 8: Word Patches by clustering for CIFAR100 dataset.

Figure 8 shows the Word Patches generated from training images. As can be seen from the Figures, Word Patches also contain a variety of patterns.

Figure 9 shows the visualization results of trainable Word Patches. Similarly with the Word Patches

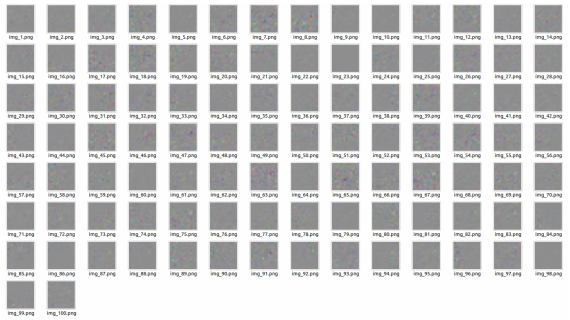

Figure 9: Trainable Word Patches for CIFAR100 dataset.

in previous section, it seems to capture more detailed features than the Word Patches obtained by clustering. This may have resulted in improved accuracy.

5 CONCLUSIONS

We proposed a method to use Word Patches in the Vision Transformer. Experimental results on the Food101 and CIFAR100 datasets showed that the accuracy of the Vision Transformer was improved. In addition, by using trainable Word Patches that fine patterns are generated automatically, the classification accuracy was improved further. The improvement of the Vision Transformer using Word Patches will lead to advances in recent researches using the Transformer.

Although the accuracy was improved by our method, we are not sure that the proposed method is the best way for creating Word Patches. In the future, we would like to find a new method for creating adaptive Word Patches according to an input image.

ACKNOWLEDGEMENTS

This work is supported by JSPS KAKENHI Grant Number 21K11971.

REFERENCES

Bossard, L., Guillaumin, M., and Gool, L. V. (2014). Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chu, P., Wang, J., You, Q., Ling, H., and Liu, Z. (2021). Transmot: Spatial-temporal graph transformer for multiple object tracking. *arXiv preprint arXiv:2104.00194*.

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

Guibas, J., Mardani, M., Li, Z., Tao, A., Anandkumar, A., and Catanzaro, B. (2021). Efficient token mixing for transformers via adaptive fourier neural operators. In *International Conference on Learning Representations*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Meinhardt, T., Kirillov, A., Leal-Taixe, L., and Feichtenhofer, C. (2022). Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8844–8854.

Sethi, A. et al. (2021). Wavemix: Multi-resolution token mixing for images.

Tan, C.-H., Chen, Q., Wang, W., Zhang, Q., Zheng, S., and Ling, Z.-H. (2021). Ponet: Pooling network for efficient token mixing in long sequences. *arXiv preprint arXiv:2110.02442*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090.

Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., and Yan, S. (2022). Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10819–10829.

Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., and Wei, Y. (2022). Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pages 659–675. Springer.