LEVERAGING EFFICIENT TRAINING AND FEATURE FUSION IN TRANSFORMERS FOR MULTIMODAL CLASSIFICATION

Kenan Emir Ak, Gwang-Gook Lee, Yan Xu, Mingwei Shen

{*kenanea, gglee, yanxuml, mingweis*}*@amazon.com*
Amazon Inc.

ABSTRACT

People navigate a world that involves many different modalities and make decision on what they observe. Many of the classification problems that we face in the modern digital world are also multimodal in nature, where textual information on the web rarely occurs alone, and is often accompanied by images, sounds, or videos. The use of transformers in deep learning tasks has proven to be highly effective. However, the relationship between different modalities remains unclear. This paper investigates ways to simultaneously utilize self-attention over both text and vision modalities. We propose a novel architecture that combines the strengths of both modalities. We show that combining a text model with a fixed image model leads to the best classification performance. Additionally, we incorporate a late fusion technique to enhance the architecture's ability to capture multiple modalities. Our experiments demonstrate that our proposed method outperforms state-of-the-art baselines on Food101, MM-IMDB, and FashionGen datasets.

***Index Terms*—** Multimodal, classification, transformers, feature fusion, efficient training.

1. INTRODUCTION

With the introduction of transformers [1], there have been significant advancements in various domains such as language processing [2, 3], computer vision [4, 5]. Recently, it was demonstrated that web-sourced paired image-text data can be used to pre-train strong models for zero-shot transfer [6]. These recent advances, where a large pre-trained model is utilized, can be used to achieve competitive performance with a fine-tuning step for different tasks.

Image and text can be utilized in tandem to facilitate understanding and interpretation of information. There have been several works in computer vision-language understanding that involve pre-training multiple modalities [7, 8], which can then be applied to specific tasks such as visual question answering [9] and visual reasoning with language [10]. Other works focus on how to utilize these large pre-trained models for specific classification tasks [11, 12]. The advantage of these two works is that they can be directly used to fine-tune any unimodal model, without requiring multimodal retraining, which is costly and typically involves complex settings.

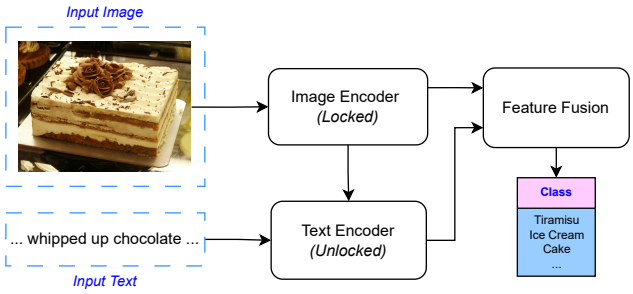

Fig. 1. Given a pair of image and text data, the proposed method utilizes image and text encoders to make the prediction. The proposed method utilizes locked image tuning and feature fusion. The text encoder is fine-tuned to connect language and vision, along with a feature fusion technique.

A commonly used approach to handle multimodal data involves using two separate pre-trained models to represent both image and textual data, followed by finetuning step for downstream tasks [13]. However, this approach may overlook the connections that exist between different modalities and can lead to issues when the gap between the two modalities is significant. The CLIP model [6] addresses this issue by making use of multimodal data and reducing the gap between different modalities, as demonstrated in [10, 14]. Despite its effectiveness, this approach does not provide inter-modal connections, which could be critical for classification tasks. Additionally, it remains unclear how to effectively leverage the capabilities of transformers for multimodal tasks and reduce the gap between different modalities.

To address these challenges, we present a multimodal architecture that enhances the inter-modal connection between text and vision. Our approach explores new multimodal models to efficiently train large, complex models while reducing the gap between modalities. Inspired by the MultiModal Bi-Transformers (MMBT) [12], we also utilize Bidirectional Encoder Representations (BERT) model [15] in our multimodal architecture as a backbone. In order to extract image features,

we utilize pre-trained CLIP vision model [6] to extract image tokens from image patches. Our findings, based on [16], indicate that locking the weights of the image model leads to the best configuration. Additionally, we employ a feature fusion layer TIRG [17] to fuse image and text features. As a result, our architecture employs both early and late fusion techniques, which maximize interaction between modalities. The design of our proposed method is shown in Figure 1.

We evaluate the performance of our proposed method by comparing it with different baselines using three datasets: UPMC Food-101 [18], Multimodal IMDB (MM-IMDB) [19], and FashionGen [20]. Our results show that we attain significant improvement in terms of compared to the baselines without sacrificing run-time efficiency. Additionally, we conduct ablation studies to examine the effect of each component of the proposed method. In conclusion, our contributions are:

1. We propose a multimodal transformer architecture that effectively integrates the intermodal connection between text and vision.

2. Our model leverages a locked image tuning technique to minimize the gap between image and text modalities and improve training stability.

3. Furthermore, we introduce a feature fusion layer in later stages of the proposed architecture to improve utilization of the image modality.

The paper is organized as follows: In Section 2, we review related methods. Section 3 presents the details of our proposed approach. The experimental results are reported in Section 4, and the paper concludes in Section 5 with a summary and future work.

2. RELATED WORK

Multimodal learning is a crucial research area as data may come from various modalities. This section will briefly go over current advancements in multimodal vision-language models and multimodal transformers.

The pre-training scheme in transformers has been a major advancement in various deep learning problems, including language processing [21], computer vision [22]. BERT [15] and Vision Transformers (ViT) [4] have become strong baselines for natural language processing and computer vision tasks, respectively. To enable transformers in computer vision, ViT employs patch projection embeddings. These pre-trained models can be used for various downstream tasks.

Several multimodal methods such as ALIGN [23] and CLIP [6] have been successful in handling multimodal data when trained for contrastive learning [24]. They perform well on various tasks, but lack intermodal interaction, making them insufficient for multimodal classification. ViLBERT [7], which is trained on images and captions in a multimodal manner, aims to utilize the BERT framework, but uses Faster RCNN-extracted bounding boxes, which are costly similar to VinVL [25]. On the other hand, MultiModal BiTransformers (MMBT) [12] is a BERT-like architecture that can embed multimodal data by projecting image embeddings extracted from a ResNet [26] architecture to the text token space. More recently introduced Vision-and-Language Transformer (ViLT) [14] is a compact model that processes visual inputs directly in its transformer architecture in a convolution-free manner. ViLT performs similarly to other multimodal models [7, 27] but with faster inference time. However, these models are trained on smaller datasets due to limited multimodal data and do not fully benefit from larger pre-training schemes.

Multimodal learning has gained popularity in recent years, with vision-language models being used in a wide range of applications, such as image captioning [28, 29, 30], text-to-image retrieval [31], image synthesis [32, 33, 34] and multimodal classification [12, 11]. An important focus of multimodal learning is how to effectively combine multiple modalities. Effective combination of multiple modalities is a key focus of this field. Traditional methods extract image and text features using separate models and then combine the features using various techniques (e.g., [35]). In our proposed architecture, we incorporate a TIRG layer to enforce greater involvement of visual features, leading to improved results.

3. PROPOSED APPROACH

Despite the remarkable performance of transformers in deep learning, it remains unclear how to effectively utilize them for handling multimodal data. In order to utilize their power, we propose a novel approach that tackles the multimodal classification problem through efficient training and feature fusion. Our goal is to correctly identify the label associated with a given input pair of modalities.

3.1. Architecture

We illustrate the proposed architecture in Figure 2. The transformer encoder is based on the Multimodal Transformer, which adopts BERT as the backbone. Given an image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$, we extract its hidden features from ViT-B/32 that is pretrained on ImageNet, using 32×32 visual patches, which results to feature dimension of $(HW/32^2 \times N)$ where H and W features sizes and N is the dimension of image features. We additionally incorporate a linear layer to project images to the space of text features. The text sequence $\boldsymbol{t}$ is converted to a textual embedding of $(T \times D)$ with a text tokenizer, where T is the number of words in the sequence and D is the width of the transformer. Both image and text features are then organized into tokens, along with token position and modal-type embeddings.

The integration of vision and text information is accomplished by merging their respective embeddings, which also

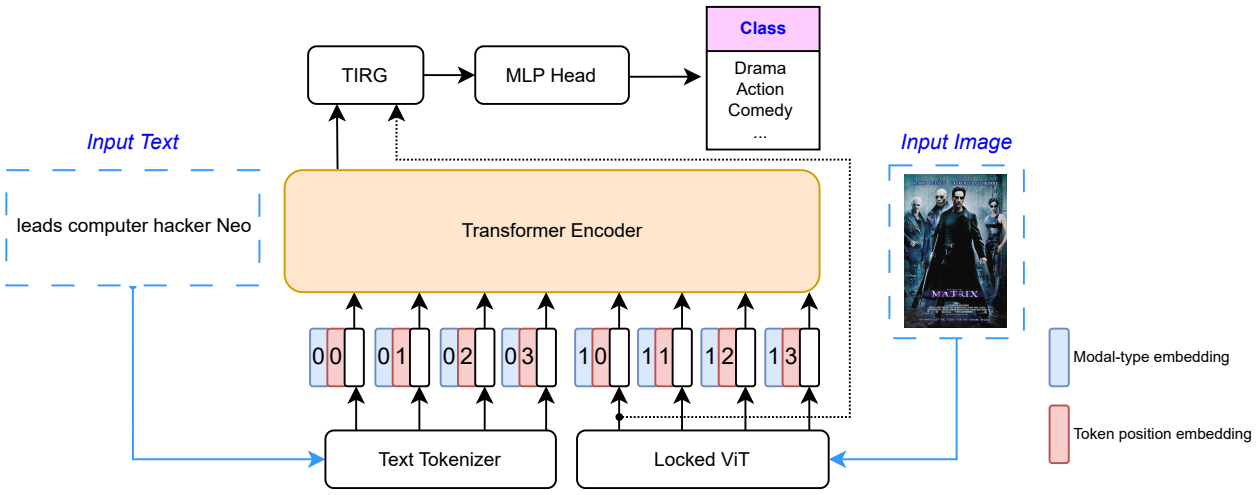

Fig. 2. The proposed architecture for multimodal classification.

enables self-attention over different modalities. A bidirectional transformer with pre-trained BERT serves as the base model for this purpose. To ensure that the classifier utilizes the image modality effectively, we employ a feature fusion technique called Text Image Residual Gating (TIRG) [17]. The output of the TIRG features are the connected to a Multi-Layer Perceptron (MLP) for performing the classification.

3.2. Locked-Image Tuning

Zhai et al. [16] investigated the performance of dual text and vision models, such as CLIP, for pre-training and found that the best performance in contrastive-tuning of transformer is achieved by locking pre-trained image models and unlocking text models. Their proposed method, called "Locked-Image Tuning" (LiT), trains a text model to extract meaningful representations from a pre-trained image model for new tasks. We argue that the same idea can be applied to finetuning for downstream tasks.

Inspired by the findings in [16], we propose to lock the ViT model in our architecture and tune the text model using image-text data for the classification task. Our approach involves mapping image embeddings to the text model's token space using a set of randomly initialized mappings. Then, we fine-tune the proposed architecture with cross-entropy loss.

3.3. Feature Fusion

The proposed architecture utilizes BERT-based transformer to make prediction, which may result in an imbalance between image and text features as the transformer would be more familiar with the text data. In order to put more emphasis on image features, we incorporate a feature fusion layer before the MLP layer. Fusion layer aims to combine pooled transformer output f and pooled image encoder output g.

This layer blends the outputs of the BERT-based transformer (represented as f) and the image encoder (represented as g) using Text Image Residual Gating (TIRG) [17]. The goal of this fusion is to adjust the image features based on the text features, rather than creating a new feature space.

4. EXPERIMENTS

This section provides the details of our experimental setup, including the datasets used, training and implementation of our proposed architecture. The experiments are performed on three datasets for multimodal classification: UPMC Food 101 [36], MM-IMDB [19], and FashionGen [37]. To ensure consistency across all experiments and improve run-time inference, we set the maximum text sequence length to 256 and the maximum image sequence length to 49. Therefore, we re-train all methods using the same settings; for instance, the original MMBT model uses max text sequence of 512 compared to 256.

UPMC Food 101 dataset connects food labels to given recipe descriptions and food images. Each data recipe was obtained from web pages, and each recipe is accompanied by one image. This dataset has 101 categories, with 60,101 text-image pairs used for training and 21,695 pairs for testing.

Multimodal IMDB (MM-IMDB) dataset consists of movie plots, poster images, and other metadata, and the task is to classify the movie genre based on the movie plots and posters. This dataset is a multi-label problem and consists of 23 labels. The dataset consists of 15,552 image-text pairs used for training and 7,799 for testing.

FashionGen dataset consists of images and their descriptions, labeled by stylists. This dataset has 260,480 images in the training set and 32,528 images in the test set. Each fashion item is photographed from multiple angles and paired with a text description. Following the convention from previ-

Table 1. Quantitative results on three publicly available datasets. Along with baseline methods, we conduct an ablation study to see the effect of each novelty. sl-40 indicates that maximum text sequence length is limited to 40.

	Food 101		MM-IMDB		FashionGen		
	Accuracy	Macro-AP	Micro-F1	Macro-F1	Mirco-F1	Macro-F1	Inference-time data/ms
Text	78.7	83.3	63.3	52.5	76.1	52.1	29.6
Text-sl:40	43.1	47.0	49.1	38.8	74.4	49.1	24.0
Image	73.3	78.0	48.6	31.6	31.8	20.4	**18.0**
MMBT	86.5	91.1	64.7	55.3	81.8	62.2	38.4
ViLT-sl:40	80.9	85.8	55.4	42.3	82.3	61.5	55.2
Ours w/o TIRG, LiT	80.3	84.5	64.1	55.0	82.3	63.5	43.2
Ours w/o TIRG	91.0	94.1	67.4	57.4	82.6	**64.9**	43.2
Ours	**92.2**	**95.3**	**68.6**	**58.7**	**83.3**	64.3	48.7

ous studies [38, 39], we use 121 sub-categories for our experiments. However, in many samples of FashionGen, the first sentence directly specifies the item category, which makes it too easy for models to rely on the text input alone. Therefore, to increase the challenge, we remove the first sentence from the descriptions.

4.1. Implementation Details

We use a pre-trained 12-layer, 768-width BERT model [15] as the backbone of our architecture (transformer encoder). The image encoder is implemented as a 12-layer, 768-width ViT-B/32 CLIP image encoder [6]. To make use of the expert image encoder network, we found it to be effective to freeze it completely. To mitigate the potential dominance of the text modality, we incorporate a feature fusion layer before the MLP head in the architecture.

For training, we set a learning rate of $5e-5$, batch size to 128 and limit the maximum text sequence length to 256, while the image token length is fixed at 49 as all images are resized to 224×224. All models are trained for 30 epochs. To ensure a fair comparison, we apply the same token limitations to all other methods used in our experiments. It's worth noting that we used the original ViLT implementation [14] in our experiments, which has a limitation of a maximum text sequence length of 40.

4.2. Results

We present the results of our experiments along with ablation studies in Table 1. The table includes evaluations based on text modality only, as well as using state-of-the-art multimodal methods (*MMBT* and *ViLT*). Our results indicate that using only text modality outperforms the image modality for unimodal data. However, when compared to state-of-the-art methods, our proposed architecture (denoted as *Ours*) shows significant improvements in all evaluation metrics and datasets. The main reason for this significant training boost is the better use of visual features via improved image encoder, efficient training and feature fusion.

Furthermore, we perform ablation studies to understand the impact of each component of our architecture. *Ours w/o TIRG, LiT* denotes the proposed architecture without TIRG and locked-image tuning, while *Ours w/o TIRG* denotes the proposed architecture without TIRG feature fusion. The results show that *Ours w/o TIRG, LiT* performs similarly or worse than *MMBT*. This is mostly due to the complexity introduced by the ViT image encoder compared to ResNet features. To overcome this, we introduce the locked-image tuning method (*Ours w/o TIRG*) which improves the performance and outperforms both *MMBT* and *ViLT* by a large margin. Adding the TIRG feature fusion method (*Ours*) results in the best performance in all metrics and datasets. Moreover, we also show that the inference time per data, which does not increase much compared to MMBT and text-based methods.

5. CONCLUSION

This paper proposes a solution to the problem of multimodal classification by introducing techniques for efficient training, improved image encoding, and effective feature fusion. By using self-attention to simultaneously process both text and image modalities, we aim to maximize the intermodal interaction. Our experiments demonstrate that combining a text model with a fixed image model yields the best classification performance, and that the late fusion technique enhances the architecture's ability to capture both modalities. Our proposed method outperforms the state-of-the-art in all datasets.

6. REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[2] Andrew M Dai and Quoc V Le, "Semi-supervised sequence learning," *NeurIPS*, vol. 28, 2015.

[3] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton, "Big self-supervised models are strong semi-supervised learners," *NeurIPS*, vol. 33, pp. 22243–22255, 2020.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICML*, 2021.

[5] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al., "Image as a foreign language: Beit pre-training for all vision and vision-language tasks," *arXiv preprint arXiv:2208.10442*, 2022.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[7] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *NeurIPS*, vol. 32, 2019.

[8] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela, "Flava: A foundational language and vision alignment model," in *CVPR*, 2022, pp. 15638–15650.

[9] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus, "Simple baseline for visual question answering," *arXiv preprint arXiv:1512.02167*, 2015.

[10] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi, "A corpus of natural language for visual reasoning," in *ACL*, 2017, pp. 217–223.

[11] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng, "Are multimodal transformers robust to missing modality?," in *CVPR*, 2022, pp. 18177–18186.

[12] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine, "Supervised multimodal bitransformers for classifying images and text," *arXiv preprint arXiv:1909.02950*, 2019.

[13] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei, "Trocr: Transformer-based optical character recognition with pre-trained models," *arXiv preprint arXiv:2109.10282*, 2021.

[14] Wonjae Kim, Bokyung Son, and Ildoo Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *ICML*, 2021, pp. 5583–5594.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer, "Lit: Zero-shot transfer with locked-image text tuning," in *CVPR*, 2022, pp. 18123–18133.

[17] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays, "Composing text and image for image retrieval-an empirical odyssey," in *CVPR*, 2019, pp. 6439–6448.

[18] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso, "Recipe recognition with large multimodal food dataset," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2015, pp. 1–6.

[19] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.

[20] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal, "Fashion-gen: The generative fashion dataset and challenge," *arXiv preprint arXiv:1806.08317*, 2018.

[21] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al., "Unilmv2: Pseudo-masked language models for unified language model pre-training," in *ICML*, 2020, pp. 642–652.

[22] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid, "Vivit: A video vision transformer," in *ICCV*, 2021, pp. 6836–6846.

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021, pp. 4904–4916.

[24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *NeurIPS*, vol. 33, pp. 18661–18673, 2020.

[25] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao, "Vinvl: Revisiting visual representations in vision-language models," in *CVPR*, 2021, pp. 5579–5588.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[27] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu, "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers," *ArXiv*, 2020.

[28] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*. PMLR, 2022, pp. 12888–12900.

[30] Kenan E Ak, Ying Sun, and Joo Hwee Lim, "Learning by imagination: A joint framework for text-based image manipulation and change captioning," *IEEE Transactions on Multimedia*, 2022.

[31] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen, "Dual-path convolutional image-text embeddings with instance loss," *TOMM*, vol. 16, no. 2, pp. 1–23, 2020.

[32] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee, "Generative adversarial text to image synthesis," *arXiv:1605.05396*, 2016.

[33] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim, "Semantically consistent text to fashion image synthesis with an enhanced attentional generative adversarial network," *PRL*, 2020.

[34] Heqing Zou, Kenan E Ak, and Ashraf A Kassim, "Edge-gan: Edge conditioned multi-view face image generation," in *ICIP*. IEEE, 2020, pp. 2401–2405.

[35] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 2016, pp. 439–448.

[36] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool, "Food-101–mining discriminative components with random forests," in *ECCV*, 2014, pp. 446–461.

[37] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal, "Fashion-Gen: The Generative Fashion Dataset and Challenge," *ArXiv e-prints*, June 2018.

[38] Jinmiao Fu, Shaoyuan Xu, Huidong Liu, Yang Liu, Ning Xie, Chien-Chih Wang, Jia Liu, Yi Sun, and Bryan Wang, "Cma-clip: Cross-modality attention clip for text-image classification," in *ICIP*, 2022, pp. 2846–2850.

[39] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao, "Kaleido-bert: Vision-language pre-training on fashion domain," in *CVPR*, 2021, pp. 12647–12657.