Paired Cross-Modal Data Augmentation for Fine-Grained Image-to-Text Retrieval

Hao Wang
Nanyang Technological University
Singapore
hao005@ntu.edu.sg

Guosheng Lin*
Nanyang Technological University
Singapore
gslin@ntu.edu.sg

Steven C. H. Hoi
Singapore Management University
Salesforce Research Asia
Singapore
chhoi@smu.edu.sg

Chunyan Miao*
Nanyang Technological University
Singapore
ascymiao@ntu.edu.sg

ABSTRACT

This paper investigates an open research problem of generating text-image pairs to improve the training of fine-grained image-to-text cross-modal retrieval task, and proposes a novel framework for paired data augmentation by uncovering the hidden semantic information of StyleGAN2 model. Specifically, we first train a StyleGAN2 model on the given dataset. We then project the real images back to the latent space of StyleGAN2 to obtain the latent codes. To make the generated images manipulable, we further introduce a latent space alignment module to learn the alignment between StyleGAN2 latent codes and the corresponding textual caption features. When we do online paired data augmentation, we first generate augmented text through random token replacement, then pass the augmented text into the latent space alignment module to output the latent codes, which are finally fed to StyleGAN2 to generate the augmented images. We evaluate the efficacy of our augmented data approach on two public cross-modal retrieval datasets, in which the promising experimental results demonstrate the augmented text-image pair data can be trained together with the original data to boost the image-to-text cross-modal retrieval performance.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Image-to-text retrieval

*Corresponding authors

1 INTRODUCTION

It is often difficult or expensive to collect large amounts of data annotations for deep learning based model training, thus automated data augmentation has been widely used as a practical technique to boost the model performance. Existing methods [13, 35, 51] are mainly restricted within the single-modality data augmentation. Specifically, Rand-Augment [7] and RandomErasing [57] are adopted on the vision transformers [42] to boost the model performance. In the cross-modal domain of image-text data augmentation, Mahajan et al.[32] replace the object tokens in the captions with a placeholder to describe the context of the image, then they generate pseudo captions with similar semantic meanings to supervise captioning model learning. Wang et al. [49] use regions of high overlap with the ground truth regions as the augmented positive regions, such that the augmented region-phrase pairs give better matching results at evaluation. However, these prevailing data augmentation methods fail to change the semantic contents of the original data and cannot generate data with large diversity. Moreover, there are few works on the data augmentation techniques on the paired cross-modal text-image data. If we want to do the data augmentation for text and image both and construct useful text-image pairs, the challenge appears to be: how can we generate the augmented text-image pairs with the same semantic information?

To tackle the problem of paired data augmentation for the image-to-text cross-modal retrieval task, we propose a novel paired text-image data augmentation algorithm, which can be used together with the prevailing single-modality data augmentation strategy and is easily plugged into the existing retrieval methods. To be specific, since text is formed by the combinations of various word tokens, increasing the number of semantic word combinations can be a simple yet effective way to produce more diverse text features for the robust inference. That means we can randomly replace part of the words in the textual captions, to construct the augmented text. To further match the semantic consistency between the augmented text and images, we generate the augmented images from the augmented text.

StyleGAN   **latent space alignment**
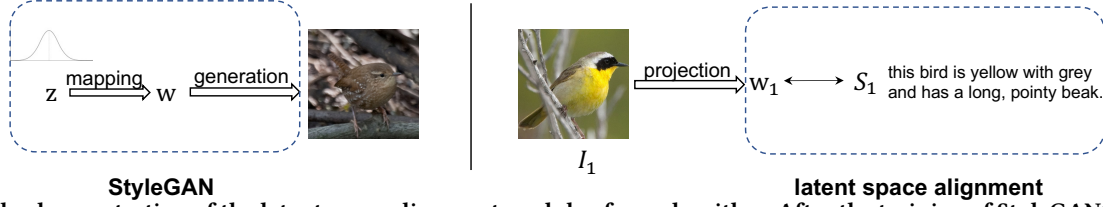
Figure 1: The demonstration of the latent space alignment module of our algorithm. After the training of StyleGAN2, we project the real images back and get the latent codes w that can be used to reconstruct the given images. To manipulate the generated images based on the textual captions, we align the text features to the paired w. With the learned latent space alignment module, we can generate images with the text semantic information, by giving the corresponding latent codes w.

We take the advantage of the StyleGAN2 [22] model, which can generate images with high quality and diversity. It is notable that we train the StyleGAN2 without the conditional text input, otherwise the model generation performance would be restricted by the limited text-image pairs [46]. Since the latent space $\mathcal{W}$ of StyleGAN has been proven to be disentangled to the semantic contents [21, 38], the disentanglability of StyleGAN2 enables the effective semantic manipulation on the generated images. To this end, we first project the real images back to the latent space $\mathcal{W}$ of the trained StyleGAN2, where we can obtain the latent codes **w** that can be used to reconstruct the given images [20]. With the projected **w** and the corresponding text captions, we map the text features to the space $\mathcal{W}$ and learn the alignment model between paired **w**-text feature representations. The procedure is presented in Figure 1. We feed the augmented text into the trained latent space alignment module, and the output can be used as the latent codes **w** for StyleGAN2 to generate the augmented images. Thus we can get the paired augmented text-image data with consistent semantics.

To evaluate the efficacy of the augmented paired data, we experiment on the fine-grained image-to-text cross-modal retrieval task. Limited by the generation capacity of existing generative models, we can hardly generate images with multiple objects, such as images of COCO dataset [30]. The recently proposed XMC-GAN [53] adopts complex architecture to produce decent COCO images, which is infeasible to online generate the augmented images from text during the retrieval training. Therefore, here we focus on the datasets with single-object images.

In the image-to-text cross-modal retrieval task [26, 31, 36], given a sample from one modality (e.g., text), the model is required to find the corresponding data samples from another modality (e.g., image), or vice versa. Since our cross-modal data augmentation method can give unlimited raw text-image pairs in an online fashion, it can be used on top of the existing data augmentation methods and retrieval models. We experiment with multiple settings and model backbones, the experiments demonstrate models using our method can boost the original performance on two public datasets. Finally, we also present qualitative results of the augmented data.

Our contributions. In this paper, we develop a novel framework to generate new text-image data pairs to address the problem of cross-modal data augmentation. Our model consists of several novel contributions: (i) we present a method to tackle the challenge of semantic consistency between the generated text and images, which can be achieved through the projected latent codes of the StyleGAN2 model; (ii) we construct the augmented text by random

token replacement, then pass the augmented text into the latent space alignment module to give the latent codes, which are fed into StyleGAN2 to generate the augmented images; and (iii) we adapt the proposed algorithm to the image-to-text retrieval task and boost the benchmark model performance.

2 RELATED WORK

Data augmentation. For image augmentation, random rotation and flipping are some classic methods to obtain highly generalized deep networks [14, 17, 39]. To increase the robustness to adversarial examples, Zhang et al. propose mixup [52] to use the convex combinations of pairs of examples for model training, it can also benefit some downstream tasks such as the long-tailed recognition [56]. Other than giving new samples to increase the data diversity, it is shown that abandoning certain information of the training images can also improve the model performance, for instance Cutout [8] randomly cuts a region out the original image for augmentation. OnlineAugment [41] can adapt online to the learner state throughout the entire training and learn a wide variety of local and global geometric and photometric transformations. For text augmentation, there are also mixup methods. Guo et al. [13] combine the word or sentence embeddings in a certain portion to perform the augmentation on the text embeddings. Training with adversarial examples [2, 55, 59] can also be used to improve the generalization of language models. Belinkov et al. [2] manipulate each word in the sentences with natural noise or synthetic and further increase model robustness. Zhuang et al. [60] use generative models to improve the fMRI classification performance. Wang et al. [45] augment the LiDAR Points and camera images for the 3D object detection task.

Cross-modal retrieval. The cross-modal retrieval task aims to retrieve the corresponding instance of different modalities based on the given query. The general idea is to map the cross-modal data to a common space, such that the heterogeneous data can be correlated. To capture the complex non-linear correlations between text-image data, most existing methods [18, 26, 29, 31, 47] are based on the deep model architecture. Specifically, to learn the fine-grained correspondence across different semantic levels, Liu et al. [31] construct textual and visual graphs, where they achieve the cross-modal correlation through the node-level and structure-level matching. PVSE [40] learns the one-to-many mappings, which produces multiple candidate samples for a given input. PCME [6] uses the probabilistic model while PVSE uses deterministic model, PCME yields better performance than PVSE. To make use of the
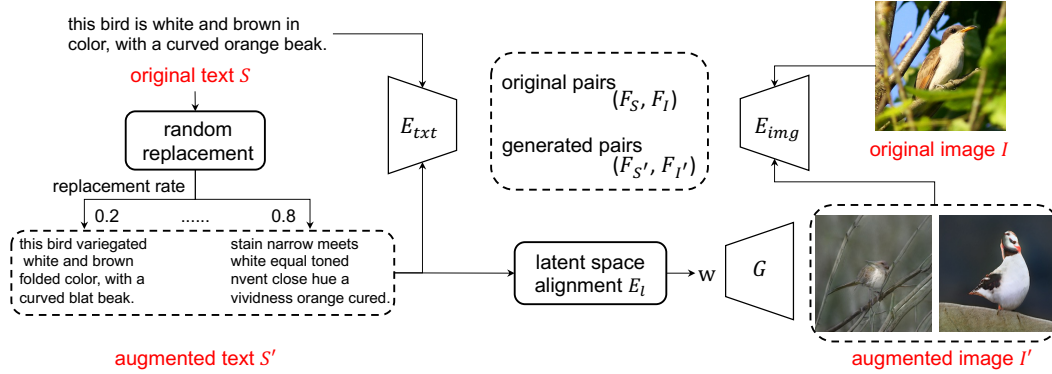
Figure 2: The generation flow of our proposed paired cross-modal data augmentation. We first replace part of the text tokens randomly, which can be controlled by different replacement rates. To further generate the augmented images based on the augmented text, we learn a latent space alignment module E_l. In this module, we align the StyleGAN2 latent codes w with the paired text features. The trained alignment module E_l is used to interpret the augmented text semantic information as w, which is fed into the StyleGAN2 to produce the augmented images. $E_{txt}(\cdot)$ and $E_{img}(\cdot)$ are the text and image encoders respectively. We incorporate both the original data representations (F_S, F_I) and the online generated text-image data representations $(F_{S'}, F_{I'})$ into the retrieval training.

unpaired data in the given dataset, Jing et al. [18] adopt autoencoders to obtain the modality-invariant representations, which can be achieved by the dual alignment at the distribution and the semantic levels. The triplet loss is widely used [25, 31, 47] as the objective for retrieval training.

3 METHOD

Our proposed paired cross-modal data augmentation method is shown in Figure 2. The whole pipeline can be summarized as the following three-stage training scheme:

- **Stage 1.** We train a StyleGAN2 [22] model with the images only. The StyleGAN2 model maps the random noise space $\mathcal{Z}$ to the style latent space $\mathcal{W}$, which is disentangled and helps generate images with high quality and diversity.
- **Stage 2.** We project the real images back to the latent space $\mathcal{W}$ and obtain the latent codes w for given images. We then learn a latent space alignment module E_l, where the text features are mapped to align with the corresponding latent codes w.
- **Stage 3.** We do the cross-modal data augmentation in an online fashion. The augmented text is constructed by random token replacement. Then we feed the augmented text into the trained alignment module E_l, whose output can be used as the latent codes w for StyleGAN2 to generate the augmented images.

We present more technical details for each stage in the following sections.

3.1 Image projection to latent space

The StyleGAN2 model can be denoted as $G(\cdot) : \mathcal{Z} \to \mathcal{X}$, where the model uses Multi-Layer Perceptron (MLP) to map the initial noise space $\mathcal{Z}$ to the style latent space $\mathcal{W}$. Then the StyleGAN2 generates images based on latent codes w of the disentangled space $\mathcal{W}$. After the training of StyleGAN2 model on the given datasets,

we follow [22] to project the real images back to the latent space $\mathcal{W}$.

In this module, we take the latent codes $w \in \mathcal{W}$ for optimization. Specifically, we first run 10,000 random noise input z to produce mapped latent codes $w = \text{MLP}(z)$. We use the average $\mu_w = \mathbb{E}_z \text{MLP}(z)$ as the initialization of w, and the approximated scale of $\mathcal{W}$ can be set as $\sigma_w^2 = \mathbb{E}_z ||\text{MLP}(z) - \mu_w||_2^2$, which is the average square Euclidean distance to the center. We adopt $\tilde{w} = w + \mathcal{N}(0, 0.05\sigma_w k^2)$ as the input to generate images, where k goes from one to zero gradually. This is an empirical formulation used in [22]. The usage of Gaussian noise on w adds stochasticity to the optimization process and makes the finding of the global optimum stabilized.

Our goal is to make the reconstructed images from the projected latent codes $x' = G(\tilde{w})$ identical to the original real images x. To this end, we adopt the perceptual loss [19] as the optimization objective, which can be denoted as

$$\min_w \mathcal{L}_p = ||F(x) - F(G(\tilde{w}))||_2^2, \quad (1)$$

where the $F(\cdot)$ denotes the VGG [39] feature extraction model. The optimized results w_{opt} are the closest latent codes that can reconstruct the given real images.

3.2 Latent space alignment

The projection of images to latent space $\mathcal{W}$ gives the corresponding latent codes w_{opt} for images x. As the latent space $\mathcal{W}$ of StyleGAN2 has been proven to be disentangled to the semantic contents [21, 38], which has the same property as text feature space. We can map the text representations into the same space as $\mathcal{W}$. Moreover, since the text representations can be semantically changed with the raw text input, when we can achieve the multi-modal alignment between text feature space and StyleGAN2 latent space $\mathcal{W}$, the generated images can be manipulated with the given text.

To be specific, we adopt an LSTM E_l to encode the textual captions S and output the text representations $t = E_l(S)$, where t has

the same feature dimensions as $\mathbf{w}$. Since the paired relationships between images and latent codes as well as that between images and text are available, we can adopt the pairwise ranking loss to learn the alignment between $\mathbf{t}$ and $\mathbf{w}_{opt}$, which can be trained with

$$\min_{\Theta_{E_l}} \mathcal{L}_{E_l} = ||\mathbf{w}_{opt} - \mathbf{t}||_2^2, \quad (2)$$

where the Θ_{E_l} denotes the parameters of the text encoder E_l, which is latent space alignment module, and $\mathbf{w}_{opt}$ is fixed during training.

This is a simple yet effective way to learn the alignment between text encoder E_l feature space and StyleGAN2 latent space $\mathcal{W}$, which can leverage the disentanglability of StyleGAN2 and uncover the hidden semantic structures of latent space $\mathcal{W}$, such that the generated images can be manipulate-able through the text. After the training of the latent space alignment model, we input the text data into the trained E_l, the output text representations can be regarded as the latent codes $\mathbf{w}$ for StyleGAN2 to generate images. We present generation visualizations in Section 4.7.

3.3 Online paired data generation

Based on the captions, we obtain the vocabulary V consisting of all the existing words of the given datasets. Moreover, we adopt the spaCy [16] library to give Part-of-Speech (POS) tagging, such as adj. and noun., for each word token in the captions. Then we also collect a POS vocabulary V_{pos}, where a set of words can be retrieved from the POS tagging.

Given a list of caption word tokens $S = \{s_1, \ldots, s_N\}$, we first select part of the tokens in S based on the replacement rate r, and then randomly pick other tokens in V or V_{pos} of the given dataset to replace the selected original tokens. The token list after the random replacement can be denoted as the augmented text data S'. We then feed the augmented text S' into the trained latent space alignment module E_l, the output $E_l(S')$ can be used as the latent codes for StyleGAN2 to generate augmented images $I' = G(E_l(S'))$. It is notable that the replacement rate r and augmentation strategy (e.g. using V or V_{pos}) are the hyper-parameters, the full procedure is presented in Algorithm 1.

In one mini-batch, we have the original text-image pairs $D = (S, I)$ and the augmented pairs $D' = (S', I')$ for training. Here we use the proposed online paired data augmentation method for the cross-modal retrieval task. $E_{txt}(\cdot)$ and $E_{img}(\cdot)$ represent text and image encoders respectively. $F_D = (E_{txt}(S), E_{img}(I))$ and $F_{D'} = (E_{txt}(S'), E_{img}(I'))$ denote the extracted feature sets of the original and augmented data in a mini-batch.

We adopt the triplet loss to learn the similarity between text and image data as follows

$$\mathcal{L}_{tri} = \sum_{D, D'} \left[d_{ap} - d_{an} + m \right]_+. \quad (3)$$

where m is the margin, d_{ap} and d_{an} denote the anchor-positive sample distances and anchor-negative distances respectively. Specifically, the form of the constructed triplet could be $< I_a, S_p, S_n >$ or $< I_a, S_p, I_n >$. That means when we take the image I_a as the anchor sample, the paired text S_p is used as the positive sample. We then select a text S_n or image I_n sample from different pairs as the negative sample. The triplets of text data as the anchor can be constructed in a similar way. The summation symbol means

---

Algorithm 1 Paired cross-modal data augmentation strategy

Require: The trained StyleGAN2 model G and latent space alignment module E_l;

Input: Textual caption data S, replacement rate r, whole vocabulary V and POS vocabulary V_{pos};

1: Initialize replacement strategy $\theta \in \{random, pos\}$;
2: **for** $S = \{s_1, \ldots, s_N\} \in$ minibatch **do**
3:     Give POS tagging $\{pos_{s_1}, \ldots, pos_{s_N}\}$ for S;
4:     Randomly select part of tokens $S_r = \{s_j, \ldots, s_k\}$ from S to be replaced based on r;
5:     **if** θ is random **then**
6:         Randomly select new tokens $\{s'_j, \ldots, s'_k\}$ from V;
7:     **if** θ is pos **then**
8:         For $s \in S_r$, select random tokens s' from $V_{pos}(pos_s)$ to replace s;
9:     Augmented captions $S' \leftarrow \{s_i, s'_j, \ldots, s'_k\}$;
10:     Augmented images $I' \leftarrow G(E_l(S'))$;
11: **return** paired data (S', I')

---

that we construct triplets and do the training for all the text and image instances of the mini-batch, including both original data D and augmented data D'. To improve the effectiveness of training, we adopt the hard sample mining method used in [15].

4 EXPERIMENTS

4.1 Dataset and experiment settings

We conduct experiments on two public datasets, i.e. CUB 200-2011 [44] and Recipe1M [36] datasets.

CUB. CUB dataset consists of 11,788 images from 200 bird classes, where each image has 10 English textual descriptions. We evaluate our proposed method on CUB dataset with two types of cross-modal retrieval tasks, i.e. instance-level and class-level retrieval. For instance-level retrieval training, we only take the image and its corresponding captions as the matched pairs. We follow the official splits [44], where there are 5,994 and 5,794 pairs for training and test respectively. While in the class-level retrieval, we regard all the image and text instances from the same class as the positive samples. We follow the class splits used in [6], we use 150 classes for training and validation, where 80% is used for training and 20% is used for validation. The remaining 50 classes are used for the test.

Recipe1M. Recipe1M is a large-scale cross-modal food dataset, collected from various cooking websites. It contains user-uploaded food image and recipe data, which includes the food titles, ingredients and cooking instructions. There are in total 238,999 training food image-recipe pairs, 51,119 and 51,303 pairs for validation and test respectively. We follow previous works [35, 36, 47] to do instance-level retrieval. To be specific, we are interested in finding the corresponding recipe for the given food image query, and vice versa.

Evaluation metrics. We follow previous works [36, 54] to use Recall@{1, 5, 10} to do the performance evaluation. In CUB and Recipe1M instance-level retrieval, we randomly sample 1000 pairs

Table 1: R@K (%) of CUB instance-level retrieval models trained with random replacement strategy. We show model evaluation results of various replacement rates r, where $r = 0$ denotes the baseline and $r = 0.7$ achieves the best performance.

r	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
R@1	13.0	13.9	14.2	15.0	15.3	14.4	15.3	**15.8**	15.6	15.0	14.2
R@5	34.8	37.5	39.6	39.2	40.9	40.2	40.3	**41.1**	40.1	40.0	39.1
R@10	49.5	50.5	52.6	52.3	52.6	52.8	52.3	**54.0**	53.3	51.5	51.3

Table 2: R@K (%) of CUB instance-level retrieval models trained with POS replacement strategy. The headings indicate the tokens with certain POS tagging are replaced. Using the random replacement strategy yields the best results.

$r = 0.7$	random	all tagging	adj.	noun.	adj.+nonn.
R@1	**15.8**	15.3	14.6	14.8	14.9
R@5	**41.1**	39.6	40.1	39.5	39.8
R@10	**54.0**	51.9	52.3	52.5	52.0

for 10 times and report the averaged metrics results. In CUB class-level retrieval, we adopt the R-Precision (R-P) metric as an alternative to do evaluation on the whole test set. R-P is used in [6, 33], it computes the ratio of positive items in the top-R retrieved items, where R is the number of references in the same class as the query. This metric can get the best scores only if it retrieves all the positive items on the top-R retrieved results.

4.2 Implementation details

We use the PyTorch implementation of StyelGAN2-ADA [20] with the auto settings to train the image generation model on CUB and Recipe1M datasets, with the default training setting. We use a one-layer bi-directional LSTM as the latent space alignment module, the output dimension is identical with the dimension of the StyleGAN2 latent codes $\mathbf{w}$, which is 512.

In cross-modal retrieval training, for CUB, we use a pretrained ResNet-50 [14] model and a DistilBERT [37] model as the image encoder and the text encoder respectively, where the output feature dimension is 512. For Recipe1M, we implement the proposed cross-modal data augmentation method on the state-of-the-art work [35]. They [35] adopt transformers [43] to encode the recipes and ResNet-50 [14] for images, where the output dimension is 1024. We train the model with batch size of 32 (resp. 128) for CUB (resp. Recipe1M). For both datasets, we use the Adam [24] optimizer with the initial learning rate of 0.0001 and the learning rate decays 0.1 after 30 epochs. We set the margin m as 0.3. We use the model having the best R@1 on the validation set for testing.

We first use the augmented cross-modal data to pretrain the retrieval model for 100 (5) epochs for CUB (resp. Recipe1M) dataset. Then we finetune the pretrained backbone on the real data. We run the experiments on a single Tesla V100 GPU. The pretraining (resp. finetuning) process costs 5 hours (resp. 2 hours) and 1 day (resp. 3 days) for CUB and Recipe1M respectively.

4.3 Results of different replacement strategies

Here we choose to use CUB instance-level image-to-text retrieval to show the results of the ablation studies. In CUB, the unpaired caption and image data of the same class, can vary due to the differences on the bird poses, shapes and colours, hence the instance-level retrieval is more challenging than the class-level one and can reflect more fine-grained information from the model. We take the same training method across all the experiments, where we first train the model with the adopted data augmentation techniques, then we finetune the trained model with the original training set.

We have two text replacement strategies, i.e. random replacement and POS replacement. The difference between the two strategies is that, random replacement randomly selects other words from

the whole vocabulary to replace, while POS replacement considers the POS tagging of the replaced word and retrieve another random word with the same POS tagging to replace.

Random replacement strategy. In Table 1, we show the evaluation performance of models trained with random replacement strategy. To be specific, we experiment with various replacement rates r to construct the augmented data, $r = 0$ denotes the baseline performance, where we do not perform the proposed augmentation method. We observe $r = 0.7$ yields the best performance, it can improve the baseline performance by 21.5% and 9.1% in R@1 and R@10 respectively, which demonstrates the effectiveness of the augmented cross-modal data for metric learning. When $r = 1$, although the generated text would be completely random, it still contains some semantic words, hence we can see $r = 1$ also outperforms the baseline model ($r = 0$) without using the augmented data. This suggests our method can increase the diversity of the paired data and produce more meaningful text-image pairs compared to the traditional single-modality data augmentation methods.

POS replacement strategy. In Table 2, we present results of the POS replacement strategy, where the replacement rate $r = 0.7$ is used for all the following experiments. We can see the results also outperform the baseline. However, there is a performance gap between POS and random replacement, where the model trained with POS replacement has over 2% drop on R@10 compared to that trained with random replacement. The reason can be when we select new word tokens from the vocabulary of certain POS tagging only, the diversity is decreased. It might suggest the importance of the diverse combinations of the cross-modal data for retrieval training.

4.4 Ablation studies

We further demonstrate results of more augmentation settings in Table 3.

Effect of different augmented data scale. Our default setting is to do the online augmentation, which means the augmented data has the same size as the real data. Here we keep $r = 0.7$ and set the augmented dataset size as 90% and 110% of the real data. We observe scaling the augmented data size as 90% of the real data size gets the best results on R@5 and R@10.

Using text-conditioned StyleGAN2 for augmentation. We use the CLIP [34] model to give text embeddings and implement the conditional StyleGAN2, where we directly input text into Style-GAN2 and obtain the generated images. Specifically, we observe the FID score of conditional generation is **5.52**, while the FID of our proposed method is **4.54** (lower is better). It indicates our generated images have better quality. The quantitative results of Table 3 also show our method has better performance on cross-modal

Table 3: R@K (%) of CUB instance-level retrieval models trained with different settings.

Method	R@1	R@5	R@10
baseline ($r = 0$)	13.0	34.8	49.5
best ($r = 0.7$)	**15.8**	41.1	54.0
Effect of augmented data scale			
90% of real data	15.7	**42.1**	**54.6**
110% of real data	15.2	40.8	53.4
Text-conditioned StyleGAN2			
$r = 0.7$	14.6	40.4	52.9
Augmented data as noise			
$r = 0.2$	13.9	37.7	49.7
$r = 0.3$	14.2	37.4	49.9
LSTM as text encoder			
no augmentation	5.4	18.0	26.7
augmentation	7.5	22.0	31.5
Single-modality data augmentation			
augmented text only	14.2	38.3	50.0
unpaired image and text	14.3	39.0	51.4
Prevailing data augmentation methods			
VSE∞ [4]	13.7	36.9	49.1
Rand-Augment	14.2	40.2	52.3
RandomErasing	14.0	38.6	51.3
Rand-Augment + RandomErasing	14.8	38.3	50.7

data augmentation than text-conditioned StyleGAN2. We give more discussions on Section 5.

Using augmented data as noise. We further experiment with small replacement rate $r \le 0.3$ in Table 3, where we assume the semantic contents of augmented data D' only has slight change compared with the original data D. Hence we set D' to remain as the same pair with D. In this case, D' can be regarded as the noisy image-text pairs, as they are not totally matched with D. We can see that this way also boosts the model performance, since adding the noisy pairs to the training process enables the retrieval model to be more noise-resistant.

Results of LSTM as the text encoder. When we change the text encoder from the pretrained DistilBERT to a plain LSTM, we can see the model performance drops heavily, where the R@10 score decreases over 46%. However, the model trained with our proposed method can still improve the LSTM baseline performance across all the evaluation metrics. It suggests the generalization of our algorithm.

Single-modality data augmentation. To evaluate if the latent space alignment module can give meaningful pair information, we run the experiments with the augmented text S' only, we also assign unpaired labels to the augmented text S' and image I' and train the model. In this experiment, we use the random replacement strategy and r is set as 0.7, which are the best setting in Table 1. It can be seen from Table 3, when we adopt S' or unpaired S' and I' for training, both methods yield inferior performance compared to the models trained with the paired augmented data.

Results of using prevailing data augmentation methods. VSE∞ [4] propose to mask partial input data as the data augmentation

Table 4: Evaluation of the performance of our proposed method compared against various CUB class-level retrieval benchmarks. The models are evaluated with R-Precision (%) and R@K (%). In baseline model, we use the pretrained ResNet-50 and DistilBERT without our augmented data.

Method	Image-to-text		Text-to-image	
	R-P	R@1	R-P	R@1
VSE0 [10]	22.4	44.2	22.6	32.7
PVSE [40]	18.4	47.8	19.9	34.4
PCME [6]	26.3	46.9	26.8	35.2
+ Ours	27.3	48.4	27.6	37.3
baseline	27.4	46.4	27.9	37.2
+ Ours	**30.0**	**52.7**	**29.7**	**40.6**

strategy, where they do not produce paired augmented image-text data but use masked data only, hence they have inferior results than ours. We also follow [42] to implement two prevailing data augmentation methods for images, i.e. Rand-Augment [7] and RandomErasing [57]. We set $N = 1, M = 2$ and erasing probability $p = 0.25$ in Rand-Augment and RandomErasing respectively, which are the empirical settings in [42]. We use the augmented data to pretrain the model, and then finetune the model without the Rand-Augment or RandomErasing augmented data. We observe these image data augmentation methods both can improve the baseline performance, while they still have performance gap with our proposed paired data augmentation method.

4.5 Results against existing methods

Here we take the augmentation settings with the best validated performance in Section 4.3, where we adopt $r = 0.7$ and the random text replacement strategy, to do the following experiments.

The existing cross-modal retrieval works on CUB [6, 10, 40] mainly focus on the class-level retrieval setting, we compare our proposed method against various models in Table 4, where the results of [6, 10, 40] are taken from [6]. We follow the prevailing batch construction method [6, 10, 25, 27], where the positive and the negative samples are from the different modality of the anchor samples. We train our baseline model by the triplet loss, where we use a pretrained ResNet-50 and DistilBERT model as image and text encoders respectively. We add the proposed paired data augmentation method on PCME [6] and our baseline model, to demonstrate the efficacy of our method. The model trained with our augmented data boosts the baseline performance across all the metrics. The augmented baseline model improves 9% and 14% on image-to-text R-Precision and R@1 scores respectively.

We show results on Recipe1M in Table 5. Recipe1M contains large amounts of recipe data and food images in the wild, hence it is more challenging than the CUB dataset. Existing methods mainly focus on how to improve the aligned feature representations of images and recipes. For example, Wang et al. [47] adopt the adversarial training and cycle consistency to learn the cross-modal embeddings. The current state-of-the-art (SOTA) method [35] introduce to use transformer as the recipe encoder, they apply the triplet loss not only on the paired images and recipes, but also on the recipe components, so that the semantic relationships within recipes can

Table 5: Evaluation of the performance of our proposed method compared against various Recipe1M instacne-level retrieval benchmarks. The models are evaluated on the basis of R@K (%).

Method	1k						10k					
	image-to-recipe			recipe-to-image			image-to-recipe			recipe-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Salvador et al. [36]	24.0	51.0	65.0	25.0	52.0	65.0	-	-	-	-	-	-
Chen et al. [5]	25.6	53.7	66.9	25.7	53.9	67.1	7.2	19.2	27.6	7.0	19.4	27.8
Carvalho et al. [3]	39.8	69.0	77.4	40.2	68.1	78.7	14.9	35.3	45.2	14.8	34.6	46.1
R2GAN [58]	39.1	71.0	81.7	40.6	72.6	83.3	13.5	33.5	44.9	14.2	35.0	46.8
MCEN [12]	48.2	75.8	83.6	48.4	76.1	83.7	20.3	43.3	54.4	21.4	44.3	55.2
ACME [47]	51.8	80.2	87.5	52.8	80.2	87.6	22.9	46.8	57.9	24.4	47.9	59.0
SCAN [48]	54.0	81.7	88.8	54.9	81.9	89.0	23.7	49.3	60.6	25.3	50.6	61.6
DaC [11]	55.9	82.4	88.7	-	-	-	26.5	51.8	62.6	-	-	-
SOTA [35]	59.1	86.9	92.3	59.1	87.0	92.7	27.3	55.4	67.3	27.8	55.6	67.3
+ Ours	**60.6**	**87.7**	**92.8**	**61.3**	**87.7**	**93.2**	**28.6**	**57.1**	**68.6**	**29.3**	**57.3**	**69.0**

Table 6: The evaluation on the test split of the COCO dataset.

	Method	image-to-text			text-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10
ResNet-50 [14]	baseline	49.2	80.1	89.7	43.4	79.6	89.9
	+ Ours	49.9	80.4	89.8	44.5	79.6	89.5
Beit [1]	baseline	58.6	84.7	91.4	47.3	81.9	**90.8**
	+ Ours	**60.2**	**85.3**	**92.3**	**49.0**	**82.1**	90.3

be leveraged. It can be seen that our augmented model further improves the performance of SOTA in the 1k and 10k retrieval settings, yielding the best results across all the metrics. Specifically, given that Recipe1M already provides abundant data with large diversity, the improvements over the SOTA at Recipe1M demonstrates the efficacy of our proposed data augmentation algorithm on the large and complex cross-modal datasets.

4.6 Results on COCO dataset

Our proposed paired data augmentation framework is mainly applicable to image-text datasets where image data has single objects, as the StyleGAN2 model can hardly produce high-quality images containing complex objects and scenarios. Nevertheless, here we show the experiment results on the COCO [30] dataset to see the efficacy of our proposed method on the dataset with complex scenarios.

The COCO dataset contain 123,287 images in total, where each image has 5 captions. The dataset is split into 113,287 images for training, 5,000 images for validation and 5,000 images for testing. We follow previous evaluation practice [9, 25], and report results of the average over 5 folds of 1K test images.

It is notable that most existing state-of-the-art cross-modal retrieval frameworks [9, 23, 25, 28] on COCO are based on the extracted object features from images, while we generate the augmented images online, it is unrealistic to adopt the object detection model to online extract the object features from the augmented images with limited computation resources. Hence we use a relatively simple end-to-end training framework to conduct COCO experiments, which is the same as the structure in Figure 2.

Specifically, we experiment with two types of image encoders, i.e. the pretrained ResNet-50 [14] and the vision transformer Beit [1]. We adopt the pretrained DistilBERT [37] model as the text encoder.

We present the evaluation results in Table 6. It is observed that our proposed paired data augmentation method boosts the baseline performance greatly at the evaluation metrics of R@1, while may not have great performance improvement at R@10. It might indicate that the augmented data helps more on identifying the top-ranked retrieved samples, since the data augmentation gives more data variety for the training process.
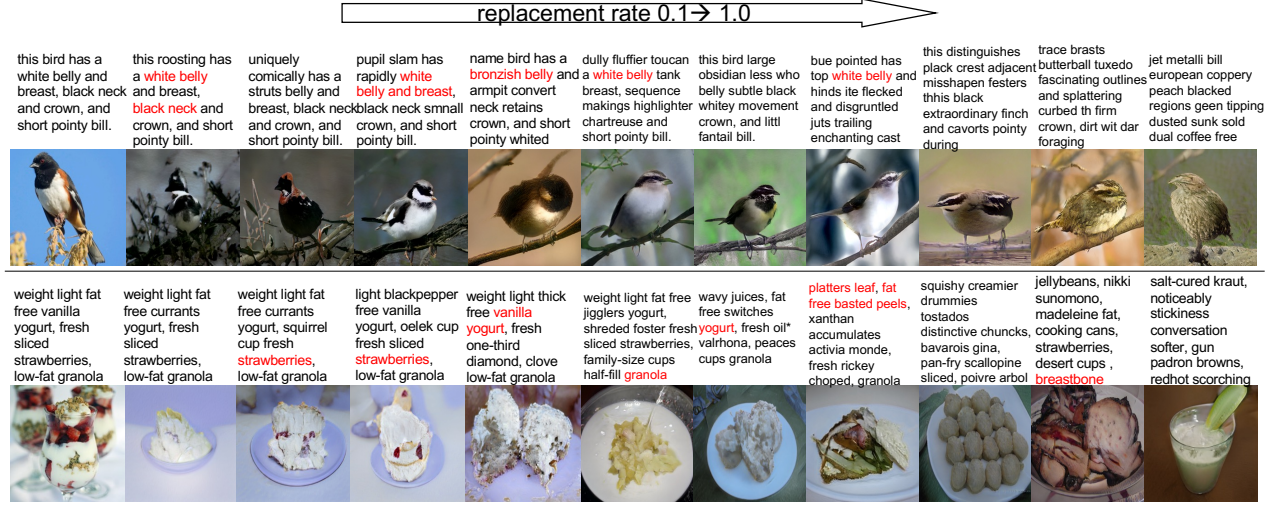
4.7 Qualitative results

We show the qualitative results of paired augmented cross-modal data in Figure 3. It is notable that, in Recipe1M dataset the text (recipe) data includes the food titles, ingredients and cooking instructions, we do random replacement on all the components of recipes. Here we only show the augmented ingredients in Figure 3 for simplicity. Besides, some word typos can be seen in the visualizations, which are actually contained in the original vocabulary of the given dataset.

To be specific, from left to right in Figure 3, we present the original data and the augmented data generated from the replacement rates r from 0.1 to 1. We can observe that the StyleGAN2 generated images based on the output of the latent space alignment module, can generally produce images with decent quality and enough diversity. Although the augmented text with the random replacement transforms looks plausible and meaningless, we can still find some correspondence between the augmented text and the generated images. For example, in the first row, when we set r as 0.3, the augmented text contain the semantic information of *white belly and breast* and *black neck and crown*, we can see the text semantics has been reflected in the generated image. When we change r to 0.4, where the text has *bronzish belly*, the bird color of the augmented image can also change accordingly. In the second row, the augmented data has minor changes within $r \in [0.1, 0.3]$, as the length of each ingredient is short. When r increases, we can see the large diversity of the augmented images.

Moreover, we show the comparison between the original and projected images in Figure 4. The projected images are generated from the projected $\mathbf{w}$ of Equation 1. We observe most projected CUB images can preserve the original image semantic information and object textures. While in Recipe1M dataset, where the food images

replacement rate 0.1→ 1.0

this bird has a white belly and breast, black neck and crown, and short pointy bill.

this roosting has a white belly and breast, black neck and crown, and short pointy bill.

uniquely comically has a struts belly and breast, black neck and crown, and short pointy bill.

pupil slam has rapidly white belly and breast, black neck smnall crown, and short pointy bill.

name bird has a bronzish belly and armpit convert neck retains crown, and short pointy whited

dully fluffier toucan a white belly tank breast, sequence makings highlighter chartreuse and short pointy bill.

this bird large obsidian less who belly subtle black whitey movement crown, and littl fantail bill.

bue pointed has top white belly and hinds ite flecked and disgruntled juts trailing enchanting cast

this distinguishes plack crest adjacent misshapen festers thhis black extraordinary finch and cavorts pointy during

trace brasts butterball tuxedo fascinating outlines and splattering curbed th firm crown, dirt wit dar foraging

jet metalli bill european coppery peach blacked regions geen tipping dusted sunk sold dual coffee free

weight light fat free vanilla yogurt, fresh sliced strawberries, low-fat granola

weight light fat free currants yogurt, fresh sliced strawberries, low-fat granola

weight light fat free currants yogurt, squirrel cup fresh sliced strawberries, low-fat granola

light blackpepper free vanilla yogurt, oelek cup fresh sliced strawberries, low-fat granola

weight light thick free vanilla yogurt, fresh one-third diamond, clove low-fat granola

weight light fat free jigglers yogurt, shreded foster fresh sliced strawberries, family-size cups half-fill granola

wavy juices, fat free switches yogurt, fresh oil* valrhona, peaces cups granola

platters leaf, fat free basted peels, xanthan accumulates activia monde, fresh rickey choped, granola

squishy creamier drummies tostados distinctive chunks, bavarois gina, pan-fry scallopine sliced, poivre arbol

jellybeans, nikki sunomono, madeleine fat, cooking cans, strawberries, desert cups , breastbone

salt-cured kraut, noticeably stickiness conversation softer, gun padron browns, redhot scorching

Figure 3: The visualizations of the augmented images generated from the augmented text, where we use the random replacement strategy. From left to right, we show the original text and images, and then the augmented data produced with various replacement rates from 0.1 to 1. The words in red colour denote the correspondence with the generated images.

original

projected

Figure 4: The original and projected images.

contain more fine-grained ingredients, the projected images may fail to preserve all the original information, since the StyleGAN2 model is not capable of generating perfect images for such complex scenarios. Nevertheless, our proposed algorithm can produce large amounts of plausible images with diverse semantics and improve the retrieval performance.

5 DISCUSSION

Why to use StyleGAN instead of the text-conditioned GAN? When we train a generation model using the text-conditioned GAN structure, we only have limited text-image pairs, thus the model can hardly produce diverse images compared to the unconditioned GAN model. The recent proposed StyleGAN gives a new perspective to generate manipulate-able images, to be specific, it introduces to map the random noise $\mathbf{z}$ to another latent codes $\mathbf{w}$, and the images are generated from the mapped latent codes $\mathbf{w}$. In this way, the StyleGAN model can learn a more disentangled latent space $\mathcal{W}$. Some research works [46, 50] demonstrate this architecture outperforms the conditional GAN on the text-to-image generation task. Therefore, here we take the advantage of the disentanglablity of StyleGAN by uncovering the semantics of the latent space $\mathcal{W}$ and further manipulate the augmented images through the latent space $\mathcal{W}$. Our experiment results also depict its superiority over the text-conditioned GAN on the cross-modal data augmentation task, as shown in Table 3.

6 CONCLUSION

This paper proposed a novel framework for paired cross-modal data augmentation, which can generate an unlimited amount of paired data for training cross-modal retrieval models. Specifically, we use the strategy of random text replacement to produce the augmented text. To generate the corresponding augmented images from the augmented text, we first adopt the StyleGAN2 model to generate images with high quality and diversity. Then we proposed to bridge the gap between the text and image data by our latent space alignment module, which maps the text features into the latent space $\mathcal{W}$ of StyleGAN2. We use the output of the learned alignment module for StyleGAN2 to generate the augmented images, thus allow us to obtain the augmented text-image pairs. We further evaluate the quality of the augmented data, through the image-to-text retrieval task on two public datasets. The promising experimental results demonstrate our proposed method can effectively boost the performance of state-of-the-art models on several different benchmarks.

REFERENCES

[1] Hangbo Bao, Li Dong, and Furu Wei. 2021. BEiT: BERT Pre-Training of Image Transformers. *arXiv preprint arXiv:2106.08254* (2021).

[2] Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173* (2017).

[3] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. 2018. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *ACM SIGIR*.

[4] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15789–15798.

[5] Jingjing Chen, Lei Pang, and Chong-Wah Ngo. 2017. Cross-modal recipe retrieval: How to cook this dish?. In *International Conference on Multimedia Modeling*. Springer, 588–600.

[6] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic Embeddings for Cross-Modal Retrieval. *arXiv preprint arXiv:2101.05068* (2021).

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 702–703.

[8] Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552* (2017).

[9] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity Reasoning and Filtration for Image-Text Matching. In *AAAI*.

[10] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).

[11] Mikhail Fain, Andrey Ponikar, Ryan Fox, and Danushka Bollegala. 2019. Dividing and Conquering Cross-Modal Recipe Retrieval: from Nearest Neighbours Baselines to SoTA. *arXiv preprint arXiv:1911.12763* (2019).

[12] Han Fu, Rui Wu, Chenghao Liu, and Jianling Sun. 2020. MCEN: Bridging Cross-Modal Gap between Cooking Recipes and Dish Images with Latent Variable Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14570–14580.

[13] Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941* (2019).

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[15] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).

[16] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303

[17] Gao Huang, Zhuang Liu, Geoff Pleiss, Laurens Van Der Maaten, and Kilian Weinberger. 2019. Convolutional networks with dense connectivity. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[18] Mengmeng Jing, Jingjing Li, Lei Zhu, Ke Lu, Yang Yang, and Zi Huang. 2020. Incomplete Cross-modal Retrieval with Dual-Aligned Variational Autoencoders. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3283–3291.

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.

[20] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676* (2020).

[21] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4401–4410.

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. CVPR*.

[23] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334* (2021).

[24] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[25] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[26] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11336–11344.

[27] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4654–4662.

[28] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*. Springer, 121–137.

[29] Zechao Li, Jinhui Tang, Liyan Zhang, and Jian Yang. 2020. Weakly-supervised semantic guided hashing for social image retrieval. *International Journal of Computer Vision* 128 (2020), 2265–2278.

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[31] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. 2020. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10921–10930.

[32] Shweta Mahajan and Stefan Roth. 2020. Diverse image captioning with context-object split latent spaces. *arXiv preprint arXiv:2011.00966* (2020).

[33] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *European Conference on Computer Vision*. Springer, 681–699.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[35] Amaia Salvador, Erhan Gundogdu, Loris Bazzani, and Michael Donoser. 2021. Revamping Cross-Modal Recipe Retrieval with Hierarchical Transformers and Self-supervised Learning. *arXiv preprint arXiv:2103.13061* (2021).

[36] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3020–3028.

[37] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[38] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9243–9252.

[39] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[40] Yale Song and Mohammad Soleymani. 2019. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1979–1988.

[41] Zhiqiang Tang, Yunhe Gao, Leonid Karlinsky, Prasanna Sattigeri, Rogerio Feris, and Dimitris Metaxas. 2020. OnlineAugment: Online data augmentation with less domain knowledge. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 313–329.

[42] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*. PMLR, 10347–10357.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[44] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).

[45] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. 2021. PointAugmenting: Cross-Modal Augmentation for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11794–11803.

[46] Hao Wang, Guosheng Lin, Steven CH Hoi, and Chunyan Miao. 2021. Cycle-consistent inverse gan for text-to-image synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*. 630–638.

[47] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven CH Hoi. 2019. Learning Cross-Modal Embeddings With Adversarial Networks for Cooking Recipes and Food Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11572–11581.

[48] Hao Wang, Doyen Sahoo, Chenghao Liu, Ke Shu, Palakorn Achananuparp, Ee-peng Lim, and CH Steven Hoi. 2021. Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. *IEEE Transactions on Multimedia* (2021).

[49] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2018. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 394–407.

[50] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. TediGAN: Text-Guided Diverse Face Image Generation and Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2256–2265.

[51] Shuo Yang, Lu Liu, and Min Xu. 2021. Free lunch for few-shot learning: Distribution calibration. *arXiv preprint arXiv:2101.06395* (2021).

[52] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).

[53] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 833–842.

[54] Ying Zhang and Huchuan Lu. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 686–701.

[55] Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2017. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342* (2017).

[56] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. 2021. Improving Calibration for Long-Tailed Recognition. *arXiv preprint arXiv:2104.00466* (2021).

[57] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13001–13008.

[58] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. 2019. R2GAN: Cross-modal recipe retrieval with generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11477–11486.

[59] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764* (2019).

[60] Peiye Zhuang, Alexander G Schwing, and Oluwasanmi Koyejo. 2019. Fmri data augmentation via synthesis. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 1783–1787.