ManiTrans: Entity-Level Text-Guided Image Manipulation via Token-wise Semantic Alignment and Generation

Jianan Wang[1] Guansong Lu[2] Hang Xu[2] Zhenguo Li[2] Chunjing Xu[2] Yanwei Fu[1]
[1]School of Data Science, Fudan University [2]Huawei Noah's Ark Lab
{jawang19, yanweifu}@fudan.edu.cn {luguansong, xu.hang, li.zhenguo, xuchunjing}@huawei.com

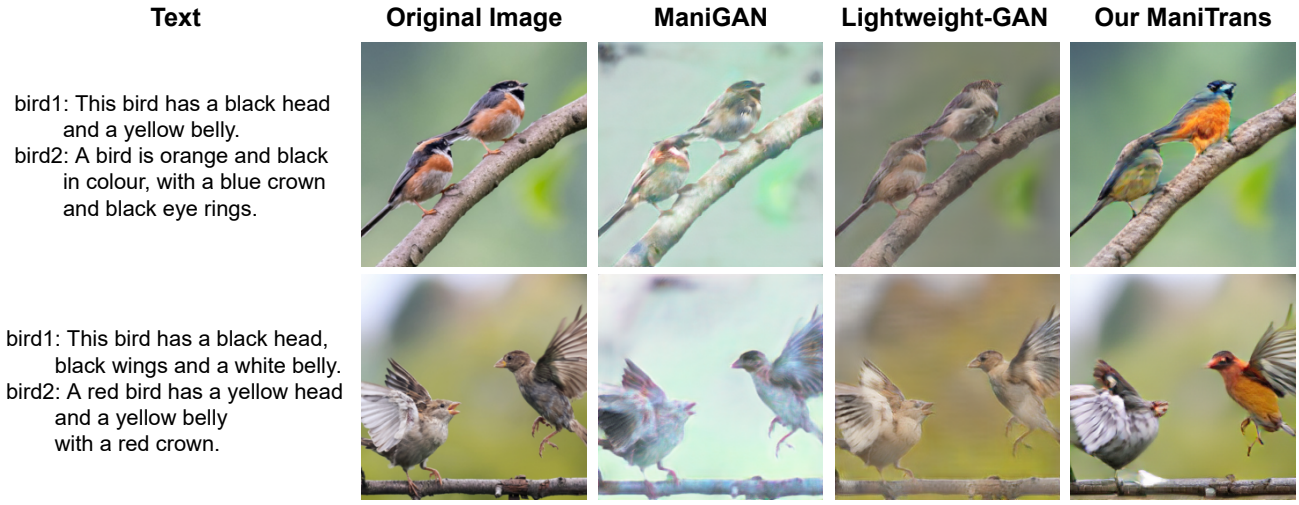

Figure 1. Results of manipulating multiple entities according to different texts with ManiGAN [26], Lightweight-GAN [28] and our ManiTrans. Our ManiTrans can manipulate different entities accordingly, while two baseline methods fail.

Abstract

*Existing text-guided image manipulation methods aim to modify the appearance of the image or to edit a few objects in a virtual or simple scenario, which is far from practical application. In this work, we study a novel task on text-guided image manipulation on the entity level in the real world. The task imposes three basic requirements, (1) to edit the entity consistent with the text descriptions, (2) to preserve the text-irrelevant regions, and (3) to merge the manipulated entity into the image naturally. To this end, we propose a new transformer-based framework based on the two-stage image synthesis method, namely **ManiTrans**, which can not only edit the appearance of entities but also generate new entities corresponding to the text guidance. Our framework incorporates a semantic alignment module to locate the image regions to be manipulated, and a semantic loss to help align the relationship between the vision and language. We conduct extensive experiments on the real datasets, CUB, Oxford, and COCO datasets to verify that our method can distinguish the relevant and irrelevant regions and achieve more precise and flexible manipulation compared with baseline methods.*

1. Introduction

There are various active branches of image manipulation, such as style transfer [16], image translation [21, 62], and Text-Guided Image Manipulation (TGIM), by taking advantage of recent deep generative architectures such as GANs [17], VAE [25] and auto-regressive models [50]. Particularly, the previous TGIM methods either operate some objects by text instructions [12, 15, 60], such as "adding" and "removing" in a simple toy scene, or manipulating the appearance of objects [4] or the style of the image [23, 52]. In this work, we are interested in a novel challenging task of entity-Level Text-Guided Image Manipulation (eL-TGIM), which is to manipulate the entities on a natural image given the text descriptions, as shown in Fig. . Critically, our eL-TGIM is much more difficult than the vanilla TGIM task, as it demands much manipulation ability in the fine-grained entity level. Thus, it is nontrivial to directly extend previ-

ous methods to this eL-TGIM task, as they can not effectively identify and edit the properties of entities as empirically shown in Fig. .

Generally, the major obstacle of the TGIM task lies in distinguishing which parts of the image to change or not change. To tackle this problem, existing TGIM methods [10, 26, 28, 34] propose many different manipulation mechanisms, such as word-level discriminator [28, 34] and text-image affine combination module [26], to differentiate the candidate editing regions from the other image parts. These methods unfortunately are still very limited to be applied to manipulate the entities in nature images. For example, Fig. shows that previous methods can only manipulate the texture/color of an object, while they fail to generate reasonable entity-level manipulation results from the text description.

To this end, we propose a novel framework of Manipulating Transformers (ManiTrans) with the token-wise semantic alignment and generation for the eL-TGIM. Thus, to tackle this task, we propose the two key ideas of *Transformer based image synthesizer (Trans)*, and *entity-level semantic manipulator (Mani)*. Specifically, the recent transformer-based architecture [14, 37, 43] has been proposed for image synthesis and has shown great expressive power. We thus present a novel component of Trans, by first learning an autoencoder to downsample and quantize an image as a sequence of discrete image tokens and then fit the joint distribution of this sequence with a transformer-based auto-regressive model.

Furthermore, to successfully identify the entities for editing, we propose the Mani component which includes a semantic alignment module, and Contrastive Language-Image Pre-training (CLIP) module. The former module helps the generation model Trans to locate and modify the text-relevant image tokens given the textual guidance. Thus our ManiTrans generation model can manipulate the image locally and preserve the irrelevant contents to a greater extent, as in Fig. . On the other hand, we repurpose the recent CLIP module as one type of semantic loss to further boost the visual-semantic alignment between the input textual guidance and the manipulated image. Essentially, such semantic loss, proposed in our ManiTrans, is complementary to token-wise classification loss, and thus efficiently serves as a pixel-level supervision signal to train our model.

We evaluate our method on multiple datasets including: CUB [51], Oxford [36] and COCO [32]. Quantitatively and qualitatively comparison against previous methods shows that our method can better manipulate the entities of an image by text while keeping the background region unchanged. Besides manipulating the texture/color of one object, our method also shows superior capability for manipulating the structure of objects guided by various textual descriptions, as shown in Fig. and Fig. 4 respectively. This can not be done in previous methods.

In summary, our contributions are as follows:

• We propose a transformer-based entity-level text-guided image manipulation framework with token-wise semantic alignment and generation, named ManiTrans, which can not only manipulate the texture/color of a single object, but also manipulate the structure of an object and manipulate multiple objects.

• We propose a semantic alignment module to locate the text-relevant image tokens for flexible manipulation, and a semantic loss for better visual-semantic alignment and detailed training signal.

• We repurpose and utilize the transformer-based image synthesizer, and CLIP module as the semantic loss in our ManiTrans framework, which is nontrivial technically.

• We quantitatively and qualitatively evaluate our method on the CUB, Oxford and COCO datasets, achieving superior/competitive results against baseline methods.

2. Related work

Text-to-image Generation Text-to-image generation focuses on generating images to visualize what texts describe. There are many good GAN-based models [44, 55, 57, 58]. Li *et al*. [27] further introduce a word-level discriminator network to provide the generator network with fine-grained feedback. Besides GANs, recent works also explore applying transformer-based network for text-to-image generation [9, 13, 42]. In contrast, rather than generating images according to texts, we focus on entity-level manipulating input images according to texts.

Text-guided Image Manipulation Text-guided image manipulation has attracted extensive attention as it enables the users to flexibly edit an image with natural language [4, 10, 12, 15, 23, 24, 26, 28, 34, 52, 54, 60] . Particularly, Li *et al*. [26] introduces a multi-stage network with a novel text-image combination module to generate high-quality images. Li *et al*. [28] propose a new word-level discriminator along with explicit word-level supervisory labels to provide the generator with detailed training feedback related to each word, achieving a lightweight and efficient generator network. Recently, due to the good synthesizing capability of StyleGAN, researchers devote to image manipulation by pre-trained StyleGAN models [39, 54]. Patashnik *et al*. [39] adopt the CLIP model for semantic alignment between text and image, and propose mapping the text prompts to input-agnostic directions in StyleGAN's style space, achieving interactive text-driven image manipulation. On the contrary, our module for image synthesis is trained from the scratch, rather than built upon pre-trained StyleGAN models. Thus our framework should in principle be much more flexible to be deployed to real-world visual applications.

Semantic Image Synthesis The task of semantic image synthesis aims to generate a photo-realistic image from a

semantic label. Isola *et al*. [21] propose a unify framework based on conditional GANs [33] for various image-to-image translation tasks, including $Semantic\ labels \leftrightarrow photo$, $Edges \rightarrow Photo$, $Day \rightarrow Night$, and so on. Chen and Koltun [5] adopt a modified perceptual loss to synthesize high-resolution images to tackle the instability of adversarial training. Wang *et al*. [53] propose a novel adversarial loss and a new multi-scale generator and discriminator architectures for generating high-resolution images with fine details and realistic textures. Park *et al*. [38] propose a spatially-adaptive normalization layer to modulate the activations using input semantic layouts and effectively propagate the semantic information throughout the network. Such works enable users to synthesis images with a finite number of semantic concepts associated with the semantic labels, while our method focuses on manipulating the input images according to the input texts, which is more flexible and with an unlimited number of semantic concepts.

Vision and Language Representation Learning Numbers of vision-language pre-training models [6, 22, 29–31, 41, 48, 59, 63] learn cross-modal representations for various down-stream tasks, including image-text retrieval, image captioning, visual grounding, and so on. They adopt the network architecture of ResNets [18] and/or Transformers [11, 50], and mainly use two kinds of learning tasks for pre-training: cross-modal contrastive learning and masked language modeling. Specifically, the recently CLIP [41] model is trained on a large-scale dataset, and shows superior performance on zero-shot tasks. We repurpose the CLIP model as one supervision loss to help train our framework for eL-TGIM.

3. Method

Fig. 2 shows the architecture of our ManiTrans framework, which is composed of Mani and Trans. In this section, we first introduce the architecture of our model. Then we introduce the language guidance and vision guidance mechanism for model training. Finally, we introduce the semantic alignment module for flexible image manipulation during the inference phase.

3.1. Manipulating Transformers Model

Our transformer-based image manipulation model consists of an autoencoder model for downsampling and quantizing the input image as discrete tokens, and a transformer model for fitting the joint distribution of image tokens.

The autoencoder based Trans model consists of three components, a convolutional encoder E, a convolutional decoder G and a codebook $\mathbf{Z} \in \mathbb{R}^{K \times n_z}$, containing K n_z-dimensional latent variables. All of them are learnable. Given an image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$, E encodes the image into a two-dimensional latent feature map $\mathbf{Q} \in \mathbb{R}^{h \times w \times n_z}$. The codebook is utilized to quantize the latent feature map by replacing each pixel embedding with its closest latent variables within the codebook element-wisely as follows:

$$\hat{\mathbf{Q}}_{ij} = \arg\min_{\mathbf{z}_k} \parallel \mathbf{Q}_{ij} - \mathbf{z}_k \parallel^2 . \tag{1}$$

For reconstruction, the decoder G takes the quantized latent feature map $\hat{\mathbf{Q}}$ as input and returns an generated image $\hat{\mathbf{X}}$ close to the original image, i.e., $\hat{\mathbf{X}} \approx \mathbf{X}$.

For image generation, the quantized feature map $\hat{\mathbf{Q}}$ can be modeled as a sequence of discrete tokens, denoted as a sequence of discrete token indices $\mathbf{I} \in \{0, \ldots, K-1\}^{h \times w}$. Each token roughly corresponds to an image patch of the size $\frac{H}{h} \times \frac{W}{w}$. Thus, the prediction of a token sequence is equivalent to synthesizing an image. In practice, we refer to uni-directional Transformer [50] to predict the image sequence autoregressively as follows:

$$P(\mathbf{I}_{\leq i}|\mathbf{T}) = \prod_{j}^{i} P(\mathbf{I}_j|\mathbf{I}_{<j}, \mathbf{T}), \tag{2}$$

where $\mathbf{T}$ is the sequence of text tokens of the caption paired with image $\mathbf{X}$.

To introduce positional information of the two modalities in Transformer, we learn two sets of positional embeddings. One is axial positional embeddings [20] for the visual sequence from a spatial grid. The other is sequence embeddings as BERT [8] for text sequence.

The autoregressive task minimizes cross entropy losses applied to the reconstruction of text tokens and image tokens, respectively [42],

$$\mathcal{L}_{txt} = -\mathbb{E}_{\mathbf{T_i}} \log P(\mathbf{T}_i|\mathbf{T}_{<i}), \tag{3}$$

$$\mathcal{L}_{img} = -\mathbb{E}_{\mathbf{I_i}} \log P(\mathbf{I}_i|\mathbf{I}_{<i}, \mathbf{T}). \tag{4}$$

Remark. *One consequent training idea is the masked sequence modeling by optimizing the loss for the paired text and image tokens. However, unlike most existing vision-and-language models [35, 49, 61] taking detected regions as an image sequence, our model accepts patch sequence, which will be an inexact alignment with text. Moreover, fine-grained correspondences of image patches and attribute tokens are difficult to be aligned. For example, aligning "a red crown" and "a red belly" within the detected bird needs to precisely recognize not only the color but also the body parts. To avoid noisy training signals, we do not adopt masked sequence modeling for training.*

3.2. Training with Language and Vision Guidance

Language Guidance. The transformer model determines the basic image tokens at the top level, and the autoencoder model holds the convolutional decoder complementing the texture in detail at the bottom level. Training these two

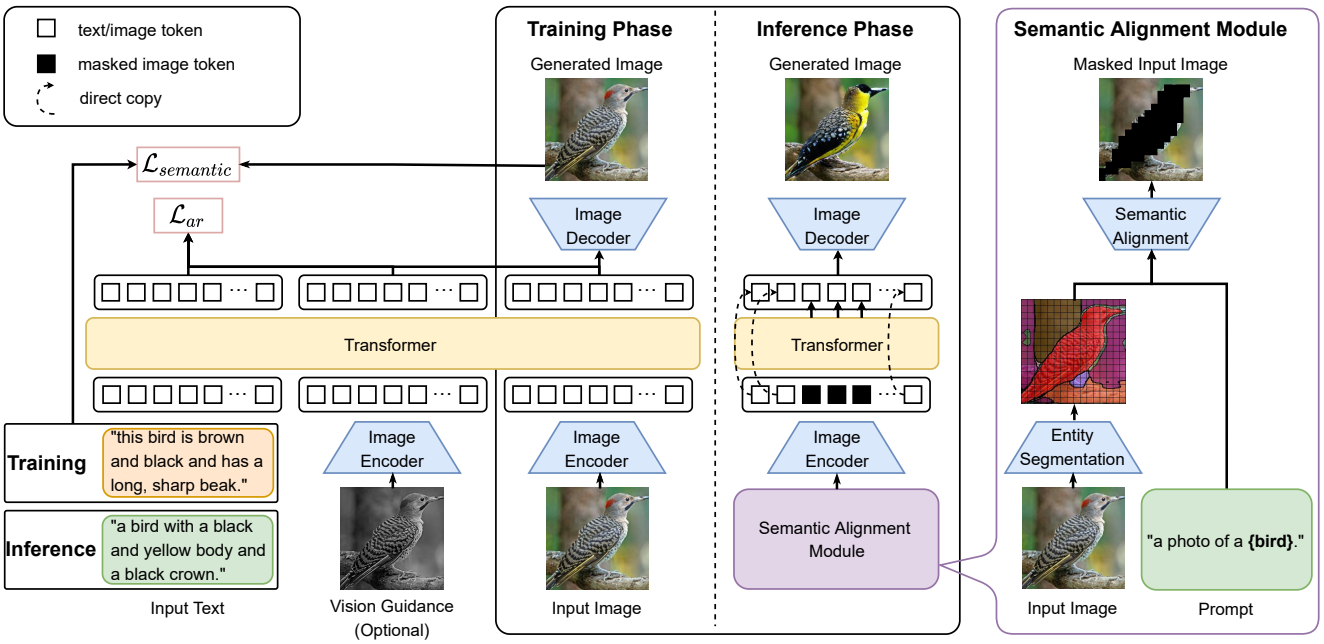

Figure 2. The architecture of our ManiTrans framework for entity-level text-guided image manipulation. The basic idea of ManiTrans is to manipulate only the image tokens which correspond to the text. ManiTrans adopts Transformer as the generation model, which takes the input text tokens and vision guidance tokens (gray/sketch image generated from the input image) as input, and generates image tokens auto-regressively. Besides the classification loss on each token, ManiTrans adopts a semantic loss to help the model capture the visual-semantic alignment between the input text and the manipulated image in the training phase. In the inference phase, ManiTrans adopts a semantic alignment module to locate the text-relevant tokens to be manipulated and the generation model manipulates only these image tokens. The vision guidance is optional and is necessary in the case of editing only the appearance of an entity.

models separately implies splitting the generation stream stiffly. To this end, in our Mani component, we propose a semantic loss for the token prediction not only considering the downstream decoding but also improving the ability to capture the relation between text and image.

The CLIP [39] is a vision-and-language representation learning model, trained with 400 million image-text pairs and has shown excellent visual-semantic alignment capability by achieving superb performance on the task of zero-shot image classification. It is optimized by a symmetric cross entropy loss over the cosine similarities of a batch of image and text embeddings. In this work, we leverage the strong model CLIP to guide our token prediction, through

$$\mathcal{L}_{semantic} = 1 - D(G(\hat{\mathbf{I}}), \mathbf{T}), \tag{5}$$

where D is the cosine similarity between the CLIP embeddings of its two arguments, as shown in Fig. 2. Note that the gradient back-propagation is implemented by straight-through estimator [2].

Vision Guidance. With the text descriptions, our model can replace an entity with other specific entities. For only editing the appearance of an entity, we need to provide the model with the prior information of the original entities' shape. Specifically, we convert the image to grayscale and append the quantized grayscale image tokens $\mathbf{V} \in \{0, \dots, K-1\}^{h \times w}$ to the text sequence as another condition for the tokens to be manipulated. The grayscale image token sequence $\mathbf{V}$ shares the positional embeddings with the image token sequence $\mathbf{I}$, for the same modality and spatial positions. For the identities of vision guidance and input text token sequence, we append two special separation tokens [BOV] and [BOT] to the beginning of them respectively. We apply the cross entropy loss on the vision guidance tokens as well,

$$\mathcal{L}_{gray} = -\mathbb{E}_{\mathbf{V_i}} \log P(\mathbf{V}_i | \mathbf{V}_{<i}). \tag{6}$$

We randomly select 50% samples to train with vision guidance. The total loss to train the transformer model is a combination of the four losses, which can be divided into two parts,

$$\mathcal{L}_{ar} = \lambda_1 \mathcal{L}_{img} + \lambda_2 \mathcal{L}_{gray} + \lambda_3 \mathcal{L}_{txt}, \tag{7}$$

$$\mathcal{L}_{total} = \mathcal{L}_{ar} + \lambda_4 \mathcal{L}_{semantic}, \tag{8}$$

where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are the balancing coefficients.

3.3. Inference with Entity Guidance

We design a semantic alignment module to locate the image patches to be manipulated by input text automatically

in the inference phase. The semantic alignment module is a two-step module, (1) to find the tokens of every entity and (2) to select the text-related entities to be manipulated, where each step bases on a strong existing model.

In the first step, we refer to entity segmentation [40] to recognize each entity on the original image $\mathbf{X}$, as Fig. 2 shows. The segmentation is implemented on the original image size, and we use the bilinear interpolation to resize the binary mask map of each entity to the same size of latent feature map $\mathbf{Q}$. The pixels whose values are larger than 0 represent that the tokens at the same position belong to the entity. In our preliminary experiments, we compare the bilinear interpolation with max-pooling for finding the entity tokens. The max-pooling dilates the tokens for the bilinear interpolation, however, due to the stack of convolutions in the first stage, the receptive field of the tokens by max-pooling is beyond the entity area and overlaps with other entities'. Thus, we use the bilinear interpolation to map the segment mask and token mask for a more precise alignment.

In the second step, we set a text prompt word to select the relevant entities. We leverage the FILIP [56], a CLIP-style model optimized by token level similarity, to calculate the similarities between image token and text token. For example, as Fig. 2 shows, we set "bird" as a prompt word to search the bird entities in the image, and then we average the similarities between tokens of each entity and the prompt word "bird". The entities whose similarities are higher than θ are the text-related entities.

Remark. *There is another prevalent way, word-patch alignment, to align a pair of text and image tokens, especially in many multi-modality transformer methods [41, 56]. The word-patch alignment begins from the word tokens to sort the patch tokens, which takes the image patches separately and neglects the information of the entity tokens as a whole during the alignment. Thus the selected image tokens may well scatter within or around an entity area. Manipulating these scattered tokens gets a messy image, where the foreground stays while the background changes. A comparison between word-patch alignment and our semantic alignment method is included in Section 4.4.*

4. Experiments

We compare ManiTrans against ManiGAN [26] and Lightweight-GAN [28]. Results of competitors are reproduced using the code/model released by the authors.

Datasets. Following common practice, we benchmark ManiTrans on three public datasets, including CUB [51], the Oxford [36] and the more complicated COCO [32] datasets. The statistics of these datasets are in the Appendix. We preprocess these datasets as in [55, 57].

4.1. Quantitative Metrics

To evaluate the quality of manipulated images, we use the Inception Score (IS) [45] as the quantitative evaluation metric. To evaluate the visual-semantic alignment between the text descriptions and manipulated images, we calculate the cosine similarity between their embeddings extracted with CLIP text/image encoders, called CLIP-score. Besides, we conduct an image-to-text retrieval experiment and report Recall@N for quantitative comparison. In the image-to-text retrieval, for each manipulated image, the text candidates consist of the input text, which serves as the positive sample, and 99 randomly sampled descriptions as negative samples. Such 100 text candidates are sorted in descending order according to their cosine similarity with the manipulated image. Recall@N calculates the percentage of images, whose positive sample occurs within the top-N candidates. As we use the ViT-B/32 CLIP model during training, for a fair comparison, we refer to the ResNet50 CLIP model to compute the CLIP-score. Additionally, following [34], to compare the quality of the content preservation, we compute the L2 reconstruction error by forwarding images with positive texts.

The higher the IS, the higher quality of the manipulated images. Higher CLIP score and R@N indicate better visual-semantic alignment between the input texts and the manipulated images. The lower the L2 error, the higher content preservation quality.

4.2. Experimental Setup

The model at the first stage inherits from the VQGAN [14] pretrained on ImageNet, where the codebook size is 1024, the image size is 256×256, and the latent feature map size is 16×16. At the second stage, our transformer has 24 layers, 8 heads with 64 dimensionalities for each head. We replace the traditional Feed-Forward Network (FFN) with a GEGLU [47] variant, which adds a Gated Linear Units (GLU) [7] with GELU [19] activation to the first hidden layer of FFN. We use Byte-Pair Encoding [46] to tokenize the text, with vocabulary size 49408. We limit the text length to 128 and learn a padding token for each position as DALL·E. Our transformer has 152M parameters, a little larger than BERT-Base 110M. The hyper-parameters of autoregressive loss $\lambda_1, \lambda_2, \lambda_3$ are set to $7/9, 1/9, 1/9$ and λ_4 of language guidance loss is 5 for all the datasets. The CLIP model for the semantic loss is ViT-B/32 [1] . For the semantic alignment module, we use the entity segmentation model based on Swin-L-W7 [2] and the FILIP-large [56] model for similarity computation. The similarity threshold θ is 0.163.

For a good initialization of the transformer, we pretrain

[1] https://github.com/openai/CLIP
[2] https://github.com/dvlab-research/Entity

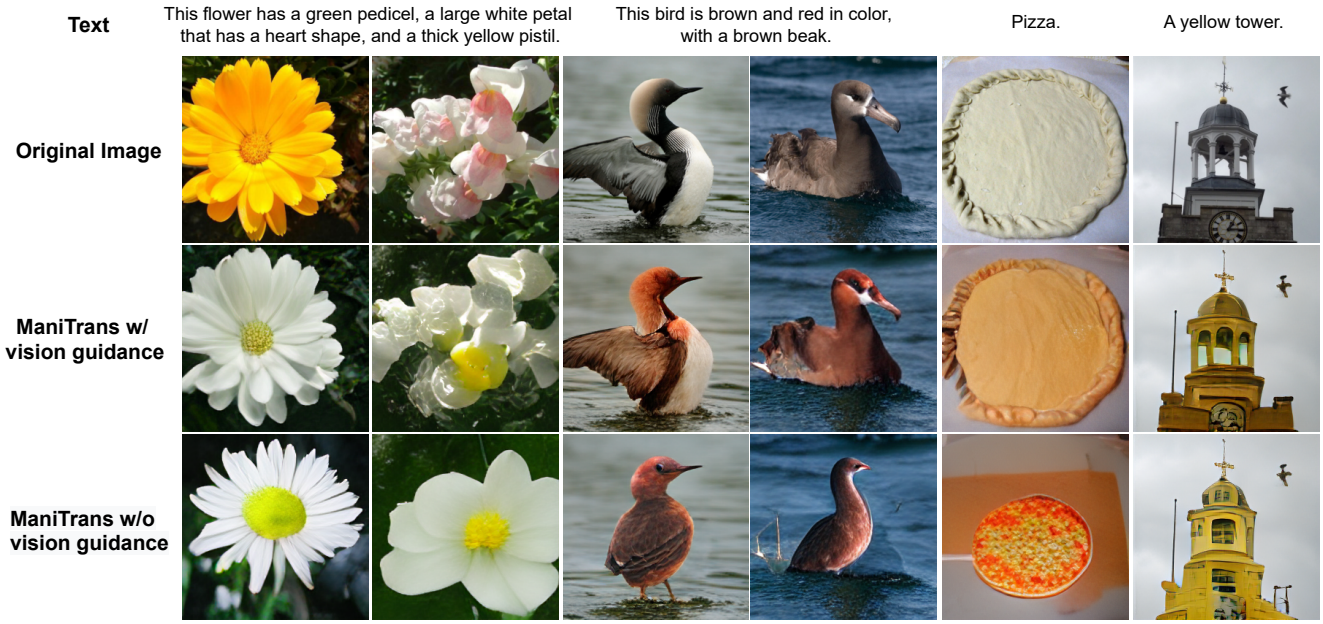

Figure 3. Our manipulation results w/ and w/o vision guidance. With vision guidance, ManiTrans can well keep the structure of an object while without vision guidance, ManiTrans can flexibly generate an entity with a different structure according to the text. The prompt word for the text "Pizza." is "dough".

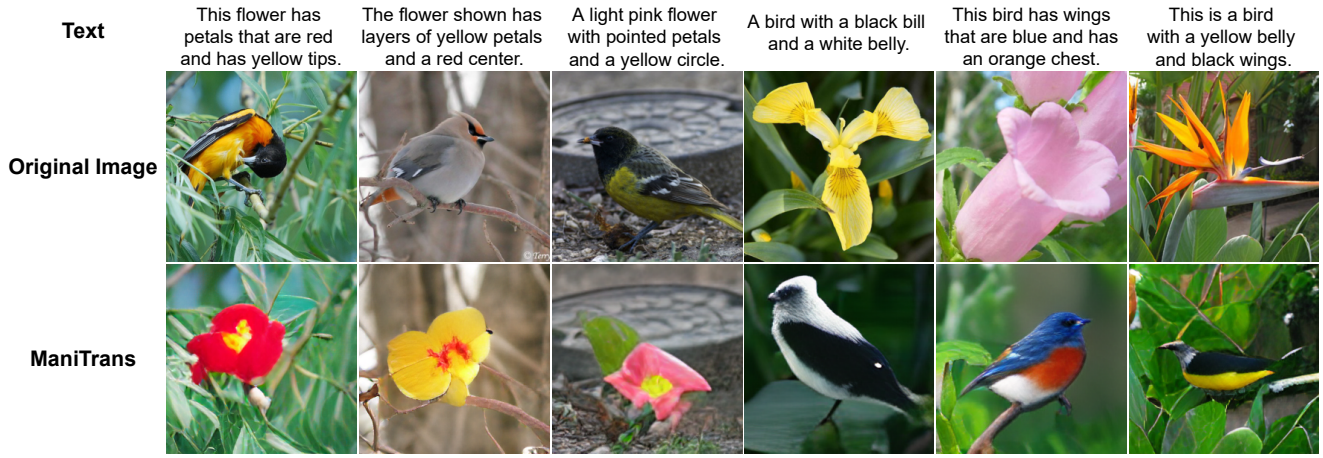

Figure 4. Manipulation results from bird to flower and flower to bird with our proposed ManiTrans.

our transformer on CC12M [3] without language and vision guidance. We use AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.96$ to train 12 epochs with batch size 112. The learning rate linearly ramps up to $6 \times e^{-4}$ for the first 5k iterations and is halved whenever training loss does not decrease for 50000 iterations. With the same optimizer, we fine-tune our model on the three datasets with the same two steps. The first step fine-tunes the model without vision guidance. The second step adds the vision guidance into training with 50% samples. Each step lasts 500 epochs with batch size 96 and the learning rate linearly ramps up to $5 \times e^{-4}$ for the first 1k iterations and is halved when training loss does not improve for 10 epochs. We implemented the proposed ManiTrans with Huawei MindSpore [1] and Pytorch; both implementations show comparable performance and efficiency.

As discussed in Section 3.3, we set a prompt for the entity to be manipulated in the inference phase. Particularly, CUB and Oxford have specific category images, where we set "bird" and "flower" as the prompt word respectively. COCO contains various category entities, and we set a prompt word for each text guide of each image. Almost all the prompt words of COCO are the nouns of their text in the following experiments and we will state the prompt words for special examples.

Model	CUB				Oxford				COCO			
	IS	CLIP-score	R@10	L2-error	IS	CLIP-score	R@10	L2-error	IS	CLIP-score	R@10	L2-error
ManiGAN	4.19 ± 0.04	21.30	10.49	0.05	4.37 ± 0.11	21.59	14.21	**0.02**	22.6 ± 0.40	11.91	14.50	0.03
Lightweight-GAN	4.66 ± 0.06	18.88	10.00	0.13	4.35 ± 0.09	17.55	11.58	0.12	**24.80 ± 0.94**	**13.65**	14.49	0.03
Ours	**5.02 ± 0.11**	**23.56**	**34.82**	**0.01**	**4.50 ± 0.06**	**23.34**	**36.49**	0.03	21.45 ± 0.41	13.10	**21.32**	**0.02**

Table 1. Quantitative comparison between ManiGAN [26], Lightweight-GAN [28] and our ManiTrans. IS - Inception Score, CLIP-score - averaged cosine similarity with CLIP embeddings, R@10 - recall within top 10 candidates, L2-error - L2 reconstruction error. IS, CLIP-score and R@10 are the higher the better, L2-error is the lower the better.

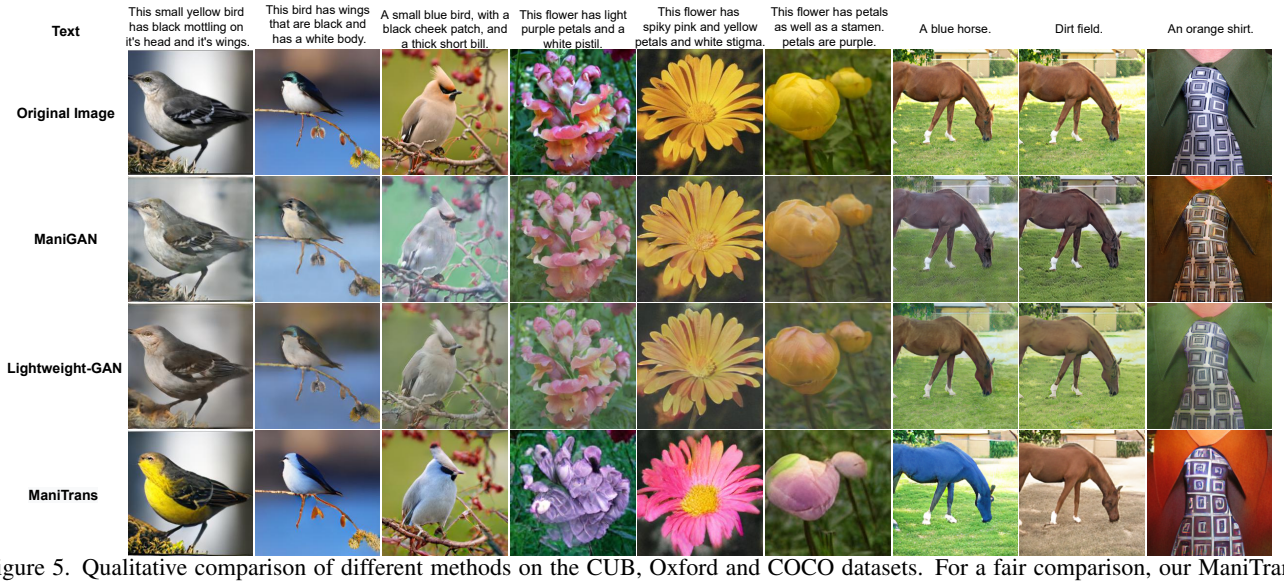

Figure 5. Qualitative comparison of different methods on the CUB, Oxford and COCO datasets. For a fair comparison, our ManiTrans uses the vision guidance to manipulate the images.

4.3. Main Results

In this section, we first qualitatively verify the manipulation ability of our model to edit or change the entity on the CUB, Oxford and COCO datasets. As Fig. 3 shows, our model can manipulate the images with the same object structure providing the vision guidance, i.e. the grayscale image, as prior shape information. Without the constraint of the vision guidance, our model generates an apparently different entity corresponding to the text description in the place of the original entity. Our framework merges the generation ability to the manipulation without any user manual mask but only the guidance of input text, where most existing models fail.

We also conduct an experiment trained on the mixture of datasets CUB and Oxford to verify a wider manipulation than on the same category. As shown in Fig. 4, ManiTrans generates reasonable manipulated entities which are corresponding to the text and fit to the background, in both bird-to-flower and flower-to-bird settings. For example, in the third column from the left, ManiTrans not only generates a flower consistent with the description but also complements the upper left corner of the manipulated flower with a leaf, which shows that ManiTrans also learns a combination of the object information and the background. Please refer to the supplementary material for more results.

4.4. Comparison with the State of the Art

Table 1 shows the quantitative comparison of our method against previous methods, including ManiGAN [26] and Lightweight-GAN [28]. On CUB and Oxford datasets, our ManiTrans achieves better results than other models on almost all metrics, except for the L2-error on Oxford datset, where ManiTrans is competitive with ManiGAN. It demonstrates that our method can generate high-quality manipulated images (IS), which are consistent with the text descriptions (CLIP-score and R@10) , and preserve the content of original images (L2-error).

For the more complicated dataset, COCO, ManiTrans outperforms the ManiGAN and Lightweight-GAN on the R@10 and L2-error and achieves competitive CLIP-score. The IS of our method are competitive with ManiGAN and Lightweight-GAN. However, as Fig. 5 shows, within many text-guided manipulation cases, ManiGAN and Lightweight-GAN both change the images slightly, more like applying a filter, while ManiTrans conducts manipulation according to the text. Typically, the former one is easier to generate high quality images than the latter and this is why their IS are a bit higher than our method.

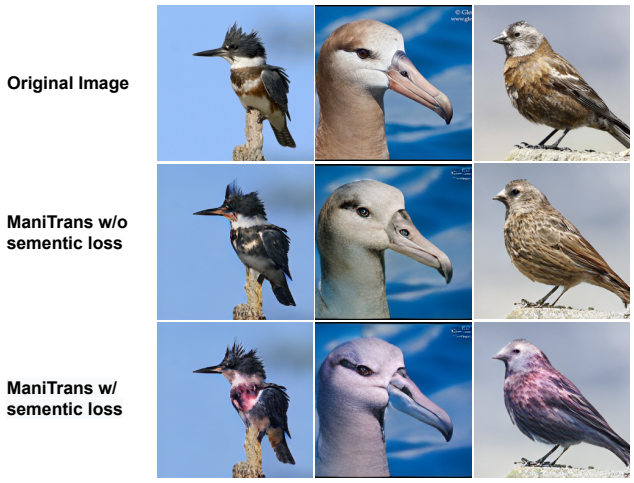

Figure 6. Qualitative comparison of our methods w/ and w/o semantic loss on CUB dataset. The input text is "This particular bird has a belly that is purple and gray.".

As Fig. 5 shows, compared with the original images, ManiGAN directs the images toward the semantic of text closer than Lightweight-GAN but changes the background style further from the original as well. Lightweight-GAN preserves the irrelevant contents better than ManiGAN while failing in transforming the text-relevant regions according to the descriptions. Our method outperforms them both on background preservation and foreground manipulation. As the second and third columns from the right of the Fig. 5 show, our ManiTrans can manipulate the horse and the field respectively on one image, while the baseline methods only change the whole image style with the text.

Effects of Semantic Loss. Fig. 6 compares the qualitative results of our model with the semantic loss or not on CUB dataset. The model trained with the semantic loss manipulates the bird as gray and purple, while the model trained without the semantic loss neglects the purple. It implies the semantic loss helps the model capture the relation between image and text.

Effects of Semantic Alignment Module. We compare our semantic alignment against the word-patch alignment qualitatively in Fig. 7. The two methods share the same similarity threshold θ for sorting the image tokens. As Fig. 7 shows, our semantic alignment selects the image patches corresponding to the bird precisely, while the word-patch alignment misses some patches corresponding to the bird and selects a few patches which belong to the background. With the inaccurate patches selected by word-patch alignment, only the right-wing turns to blue and the yellow leaks out. Although the color of two manipulated images both match the description, the qualitative result by the semantic alignment is better, resulting from more precise edited locations.

Failed Case. Our method relies on a pretrained entity segmentation and FILIP model to locate the entities to be manipulated. Though the text prompt facilitates the model to locate different entities on one image, entities with similar appearances may be unexpectedly confused by the semantic alignment model. We give a failed example in Fig. 8, where the semantic alignment module takes the wildebeest as the relative entity to "giraffe". Consequently, we fail the manipulation with red patches in the wrong place.

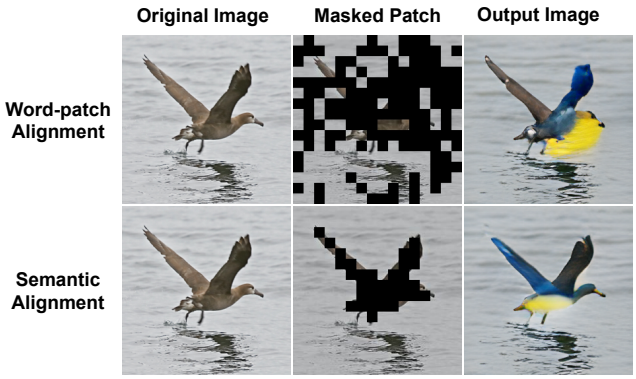

Figure 7. Qualitative comparison of our methods with our semantic alignment mechanism and word-to-patch alignment on CUB dataset. The input text is "This bird has wings that are blue and has a yellow belly.".

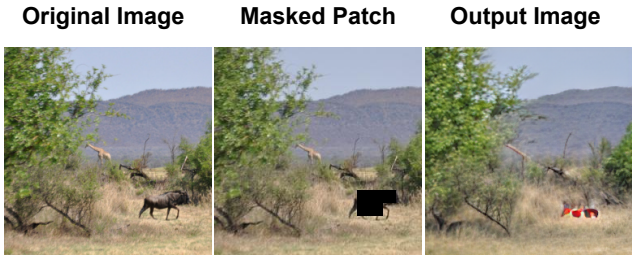

Figure 8. A failed manipulated example on COCO dataset. The input text is "A red giraffe.".

5. Conclusion

For the first time, this paper studies a new task – entity-level text-guided image manipulation. To tackle this task, we propose a novel framework – ManiTrans of Manipulating Transformers with the token-wise semantic alignment and generation. It is the two-stage framework with a semantic alignment module to manipulate the images on the entity level, which incorporates editing and generation.

Social Impact. Our ManiTrans offers a new tool to modify the images. Our framework can be used for many industrial applications on education, and potentially help the visually impaired people. We do not expect the negative social impact, as the synthesized images should be generated with the textual guidance.

Acknowledgement Partly funded by NSFC Project (62176061), and SMSTM Projects (2018SHZDZX01 and 2021SHZDZX0103) with corresponding authours: Yanwei Fu, and Hang Xu.

References

[1] Mindspore. In *https://www.mindspore.cn/*. 6

[2] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 4

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 6

[4] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8721–8729, 2018. 1, 2

[5] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017. 3

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3

[7] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017. 5

[8] 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. 3

[9] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *arXiv preprint arXiv:2105.13290*, 2021. 2

[10] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 3

[12] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10304–10312, 2019. 1, 2

[13] Patrick Esser, Robin Rombach, Andreas Blattmann, and Björn Ommer. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *arXiv preprint arXiv:2108.08827*, 2021. 2

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 2, 5

[15] Tsu-Jui Fu, Xin Wang, Scott Grafton, Miguel Eckstein, and William Yang Wang. Iterative language-based image editing via self-supervised counterfactual reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4413–4422, 2020. 1, 2

[16] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 1

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[19] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[20] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019. 3

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 3

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021. 3

[23] Wentao Jiang, Ning Xu, Jiayun Wang, Chen Gao, Jing Shi, Zhe Lin, and Si Liu. Language-guided global image editing via cross-modal cyclic mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2115–2124, 2021. 1, 2

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2

[25] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014. 1

[26] B. Li, X. Qi, T. Lukasiewicz, and Phs Torr. Manigan: Text-guided image manipulation. 2019. 1, 2, 5, 7

[27] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Controllable text-to-image generation. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 2065–2075, 2019. 2

[28] B. Li, X. Qi, Phs Torr, and T. Lukasiewicz. Lightweight generative adversarial networks for text-guided image manipulation. 2020. 1, 2, 5, 7

[29] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020. 3

[30] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 3

[31] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. *arXiv preprint arXiv:2103.00823*, 2021. 3

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 5

[33] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[34] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 42–51, 2018. 2, 5

[35] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3977–3986, 2021. 3

[36] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 2, 5

[37] Avd Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning. 2017. 2

[38] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 3

[39] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2, 4

[40] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Zhe Lin, Philip Torr, and Jiaya Jia. Open-world entity segmentation. *arXiv preprint arXiv:2107.14228*, 2021. 5

[41] 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. *arXiv preprint arXiv:2103.00020*, 2021. 3, 5

[42] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 2, 3

[43] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pages 14866–14876, 2019. 2

[44] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016. 2

[45] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016. 5

[46] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 5

[47] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 5

[48] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019. 3

[49] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1, 3

[51] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 2, 5

[52] Hai Wang, Jason D Williams, and SingBing Kang. Learning to globally edit images with textual description. *arXiv preprint arXiv:1810.05786*, 2018. 1, 2

[53] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 3

[54] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2256–2265, 2021. 2

[55] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pages 1316–1324, 2018. 2, 5

[56] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training, 2021. 5

[57] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2, 5

[58] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018. 2

[59] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 3

[60] Tianhao Zhang, Hung-Yu Tseng, Lu Jiang, Weilong Yang, Honglak Lee, and Irfan Essa. Text as neural operator: Image manipulation by text instruction. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1893–1902, 2021. 1, 2

[61] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165, 2021. 3

[62] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1

[63] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12647–12657, 2021. 3