Language Driven Image Editing via Transformers

Rodrigo Santos, António Branco and João Silva
University of Lisbon, Faculty of Sciences
NLX—Natural Language and Speech Group
Departamento de Informática, Faculdade de Ciências de Lisboa
Campo Grande, 1749-016 Lisboa, Portugal
{rsdsantos, ambranco, jrsilva}@fc.ul.pt

***Abstract*—With the emergence of specifically tailored neural architectures that cope with both modalities, cross-modal language and image processing has attracted increasing attention. A major motivation has been the search for a quantum leap in language understanding supported by visual grounding, which has been oriented mostly to solve tasks where language descriptions of images are to be provided, and vice-versa, where images are to be generated on the basis of keywords.**

Adopting a distinct angle of inquiry, this paper addresses rather the cross-modal challenge of language driven image design, focusing on the task of editing an image on the basis of language instructions to modify it. And adopting as well a distinct research path, which dispenses with specifically tailored architectures, the approach proposed here resorts rather to a general purpose, suitably instantiated neural architecture of the Transformer class.

Experimentation with this approach delivered very encouraging results, empirically demonstrating that this is an effective methodology for language driven image design and the basis for further advances in cross-modal processing and its applications with affordable compute and data.

***Index Terms*—Language Driven Image Design, Conditional Image Alteration, Computer Vision, Natural Language Processing**

I. Introduction

Image Generation has been a prominent driver for progress in Artificial Intelligence and the recently enhanced capabilities of this technology have been raising a lot of media attention and user enthusiasm. This has been concomitant with an explosion of interest for research on the subject from various research groups, and companies ([10], [18], [20], [22], [23], [25], [34]).

Models such as DALL-E [23], DALL-E 2 [22], PARTI [34], IMAGEN [25], among others, have shown the creative capabilities of these models, that appear to rapidly approximate the ones of the human minds. In those papers, the authors have trained their models on image-caption pairs, where the caption is used to guide the generation of further images that did not exist before.

In particular, the DALL-E model [23] raised a lot of interest since its release, and delivered promising results in such a task, by receiving a description in the form of a snippet of text (e.g. "a green clock in the form of an hexagon") and creating an image, out of many possible others, that humans recognize as one that could correspond to that input description. Its successors, namely DALL-E 2, PARTI, IMAGEN, further improved this performance offering higher resolution images, as well being able to undertake also more restricted tasks, where, for instance, a specified subarea of the image is to be completed on the basis of the input snippet of text.

A few factors have facilitated the emergence of this technology: (i) the gathering of massive datasets containing image-caption pairs, such as LAION [26], together with many other proprietary datasets that are not freely available; (ii) the advent of Conditional Generative neural models with the ability of crunching data faster, and with better results than previous approaches—including the Transformer architecture [28], which became one of the mainstream approaches for virtually any language processing task [3], [5], [21], given its ability to cope with the intrinsically compositional nature of language and the meaning conveyed by contextualized expressions; (iii) the enhanced computational power to put everything together—that is nevertheless hardly accessible to most research groups and organizations; and also (iv) the pursuit of combining major advances from Computer Vision and Natural Language Processing research areas.

Historically the image and language processing domains have progressed for decades quite independently of one another, with each focusing on the analysis and generation of its own modality, and the useful applications that can be built on that. Recently, though, there emerged very promising prospects for progress in cross-modal processing. A major motivation has been the realization that the so-called grounding is needed for a quantum leap in language understanding [2], and a major promising enabler has been the emergence of underlying technology that can be successfully applied to both modalities and their cross-modal processing [7], [20], [23], [29].

In the image to language direction, there has been considerable progress in the task of image captioning, that is of generating a language description for an input image [13], [20], [30], [31], and the subsidiary task of image retrieval from a language description [12], [16], [24], [33]. In the language to image direction, in turn, success has come mostly, as mentioned, from the advances on the Conditional Image Generation task. Notwithstanding the aforementioned progress, and many challenges and research opportunities are still waiting to be handled.

The research reported here was supported partially by PORTULAN CLARIN—Research Infrastructure for the Science and Technology of Language, funded by Lisboa 2020, Alentejo 2020 and FCT—Fundação para a Ciência e Tecnologia under the grant PINFRA/22117/2016.

2375-0197/22/$31.00 ©2022 IEEE
DOI 10.1109/ICTAI56018.2022.00139

In the present paper we propose and address a challenge of language driven image design, consisting of editing an image on the basis of language instructions that are explicitly meant to change that input image. Here the output image is conditioned not only on a text snippet but also on another image, such that the input image is appropriately altered taking into account the language input.

For example, given (an image of) a piece of furniture, the model is asked to change its color. And then possibly its height, its shape, the perspective, or the direction of the light, etc. This should allow one to iteratively and interactively modify the design of some object (e.g. a mug, a shoe, etc.) without any specific image manipulation software, and with no knowledge of how to work with it.

In this paper we present exploratory research results on affordable Language Driven Image Design (LDID), through the task of Conditional Image Alteration (CIA).

These contributions rely on a model driven by the research challenges addressed here, but can serve also as the basis of a new type of image editing that will allow one to interactively and on the fly modify the design of some object without dedicated image manipulation software. This can be exploited in a wide range of innovative applications, for instance in supporting a shopping assistant that progressively matches images altered by language instructions against current stock and suggests increasingly suitable products, among many others examples.

The remainder of this document is structured as follows: Section II describes the neural model used in this study; Section III explains the experiments performed and introduces the data sets used; Section IV presents the results obtained; Section V proceeds with error analysis; Section VI compares with previous work on Conditional Image Generation; Section VII discusses related work; and Section VIII closes the paper with concluding remarks.

II. Model

The Transformer [28] has become a staple architecture in Natural Language Processing in the past few years, with its variants now underlying the state-of-the-art of virtually any language processing task.

Models using only the decoder part of the Transformer, such as GPT-2 [21] or GPT-3 [3] are deemed to be the best for language generation tasks.

Recently, these models have shown promise for image processing tasks, namely in image generation [23], [29], showcasing their capacity to handle multi-modal input, and how general purpose the Transformer architecture can be, coping also with data rooted in signals that are not linguistic in nature.

Since such decoder models have been created for text, some adaptation is required in order to handle images. Interestingly, we found the changes done to the model architecture by previous work [23] can be dispensed with, and that the modifications can be restricted solely to the way the input data is pre-processed.

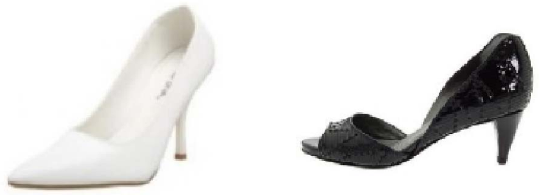


Fig. 1. The pair of images is associated with the following textual instruction: "are black with a thicker heel". Left: source image; right: target image

Accordingly, the input images have to be tokenized before being fed to the model. To do so we pass the images through a Vector-Quantized Variation Auto Encoder (VQ-VAE) that is both capable of describing an image with tokens according to an internal vocabulary of images and of constructing an image from those tokens [9]. By passing an image through a VQ-VAE, one gets a sequence of tokens that represents the image. This sequence can be used to train a decoder model like it is done with the sequence of tokens for language, given that the image tokens also have their own embedding in the embedding layer.

We resorted to a GPT-2 small model [21] as an affordable option for Language Driven Image Design, namely its current implementation from the transformers package of HuggingFace,[1] including their English pre-trained GPT-2 as well.[2]

As training parameters for the GPT-2, we use a batch size of 6 with gradient accumulation of 16, meaning that at each step our model back-propagates with 96 training instances. We evaluate on the development set every 250 steps, and stop training when the development set loss does not decrease from its lowest point after 5 evaluations.

As the VQ-VAE, we use the one from [9],[3] with a "vocabulary" for images of size 1024, which is added to the GPT-2 embedding map, and by means of which every image is represented.

Optionally, after the training of the GPT-2 model, we rank its outputs using CLIP[4] over the various images from the same input. After using two separate encoders, for image and for text, CLIP maps their encoding vectors into a common embedding so that a caption and its respective image end up with the same representation [20]. CLIP can thus support the ranking of images generated from a caption given that the encoded image that is closer (in vector space) to the encoded caption is the one more closely described by the caption.

III. Experiments

The central experiment of interest here is aimed at assessing how well the model is able to perform Conditional Image Alteration (CIA), i.e. generating an image both from another image and from a text snippet describing how the later should

[1] https://huggingface.co/docs/transformers/index

[2] https://huggingface.co/gpt2

[3] https://github.com/CompVis/taming-transformers

[4] https://github.com/openai/CLIP

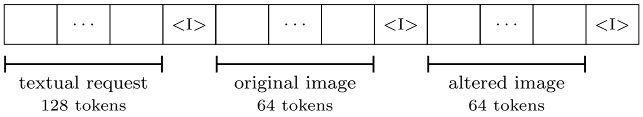


Fig. 2. Input representation

be altered. In addition, we also perform a comparison between a model with and without language pre-training.

A. Data sets

We resorted to the dataset from [12], which was developed for research on image retrieval,[5] and which we re-purposed for the task of interest here, different from that original image retrieval task.

Each instance in the dataset contains a source image of a shoe, a target image of another shoe, and a short textual description of how the source image can be manipulated to resemble the target image. An example from this dataset is displayed in Figure 1.

The data set has 10 750 examples, which were shuffled and split with a 80/10/10 proportion, resulting in 8600 examples for training, 1075 for development and 1075 for testing.

Additionally, we introduce additional variation to the images by flipping images horizontally (50% chance); rotating (between 0º and 20º clockwise or anticlockwise); distorting in order to simulate different perspectives (50% chance); increasing sharpness by a factor of 2 (50% chance); and finally maximizing their contrast (50% chance).

B. Input representation

Each instance in the data set is represented by 259 tokens, as in the schema of Figure 2: the first 128 are text tokens corresponding to the alteration request; followed by a token (<I>) marking the beginning of the source image; 64 image tokens from the source image; another <I> token marking both the end of the source image and the beginning of the target image; another 64 tokens from the target image; and finally, a last <I> token marking the end of the target image.

In an initial run, we provided to the model the source image followed by the textual alteration. However, the resulting model had worse performance than the one with the text in the first (leftmost) place, as described above. A possible explanation is that, by having the textual tokens first, the model can more easily learn the point from which no more textual tokens can occur—after the first <I>—and after that point can attribute low probabilities to textual tokens and focus solely on generating image tokens.

C. Prompt engineering

Models like GPT-2 [21] or GPT-3 [3], and also DALL-E [23] or CLIP [20] have been aligned with the emergence of so called prompt engineering. This concerns how the textual input is given to the model and how the user can craft it for the desired result to be eventually delivered.

Along this line and upon experimentation, we noted that including the designation of the type of the object image in the alteration text instead of this text stating only the alteration to be performed increases the performance of our model. For instance, the instruction “high heels are a darker tone” leads to better performance that just the instruction “are a darker tone”. This can be partly attributed to the fact that the model gets a confirmation of what image to generate (e.g. “high heels” vs. “rain boots”).

We use this approach to help CLIP rank the generated images, by prefixing the textual input with the expression denoting the type of object of the source image. Despite this improvement, the type of object of the source image may not always be the same as that of the target image, but in general a prompt prepared this way improves the performance when CLIP is used for ranking.

IV. Results

Typically there can be multiple outputs that are acceptable during the evaluation of generative tasks (e.g. summarization, machine translation, etc.), therefore such evaluation tends to be a problematic endeavour. While one could try to perform an automatic evaluation against a gold standard, small mismatches (of equally acceptable outputs) in comparison to the gold example inevitably make most such metrics, like accuracy, very brittle, leaving only some kind of distance metric to be resorted to.

This problem tends to be further aggravated for images since metrics such as BLEU [19] or METEOR [1], that are used to evaluate textual generative tasks, work by being able to refer to some text parts that are well defined substructures in an expression (e.g. spans of words), but for images there are no clear substructures that can be resorted to, and in most cases these distance metrics work only at the pixel level.

As a consequence, we resort to Mean Square Error (MSE) for evaluation. Despite being a rather rudimentary metric, particularly when compared with the previously mentioned ones for text, its evaluations come as a straightforward comparison between models.

A. Conditional Image Alteration

Figure 3 displays a couple of examples generated by our system from the respective source images and alteration instructions.

Table I presents the evaluation results for our Conditional Image Alteration (CIA) task.[6] All scores were obtained as the mean score of the top four ranked images, with the exception of the last line (as only one image was available).

The best results concentrate in the lower half of the table when CLIP is fed with the least amount of images. This seems to indicate that for the CIA task using CLIP hinders performance.

[5] https://github.com/XiaoxiaoGuo/fashion-retrieval

[6] Models were trained for 17 and 7 GPU hours, running on an NVIDIA Titan RTX 24G, with and without language pre-training respectively. Model inference (image generation) took less than a second.

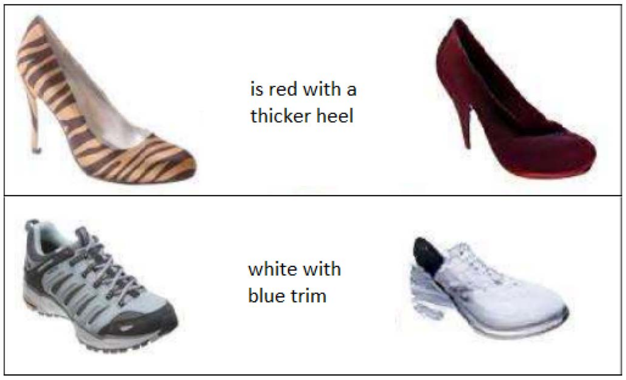


Fig. 3. Right column: source image; middle: instruction; right: image changed.

TABLE I
CIA AVERAGED SCORES OF TOP-4 IMAGES, WITH AND WITHOUT TEXTUAL PRE-TRAINING, USING MEAN SQUARE ERROR (LOWER IS BETTER). THE FIRST COLUMN INDICATES THE NUMBER OF GENERATED IMAGES GIVEN TO CLIP.

N. Images	Without pre-training	With pre-training
32	0.1103	0.1109
16	0.1076	0.1100
8	0.1074	0.1074
4	**0.1041**	0.1040
1	0.1049	**0.0967**

This is in contrast to what was found in previous work [23]. As mentioned above, the nature of the text used for CIA is probably the cause of such behaviour: In CIA the text snippet describes the *alteration* of the input image, while in previous work on Conditional Image Generation (CIG), the text describes the output image—and CLIP was trained to approximate the representations of images and their descriptions, i.e. a scenario favorable to CIG but not to CIA.

Additionally, considering the scores obtained, the model seems to get better results with textual pre-training than without language pre-training, with the best scores, for every line representing the number of examples fed to CLIP, coming for the model with textual pre-training.

V. ERROR ANALYSIS

One problem found concerns image clarity. Even though some images are correct, they have some fuzzy details. This is likely due to the reduced volume of the training data set.

Another problem arises when the target image is very different from the source image. In such cases, the model is basically asked to create a quite different object, for which the small size of the data set provided limited evidence.

Additional problems occur when the images to be generated are too similar to the source image, or the generated images are too similar to each other. While not necessarily a problem for the overall quality of the output, the first kind of cases becomes an issue for evaluation, as generated images may be more similar to the source image than to the target one. As for the second kind of cases, when the generated images are similar to one another, it may become a problem if object design is the intended use for the tool, and not just image alteration.

To address these issues, further techniques to enhance image diversity should be explored in future work, so that the model can suggest a more varied set of images to the user.

VI. CONDITIONAL IMAGE GENERATION

In order to be able to compare with previous work in Conditional Image Generation, we resorted to DALL-E mini,[7] a smaller version of DALL-E, since DALL-E is not available.

Considering that DALL-E performs a different task, we retrain our model on Conditional Image Generation using a subset of our dataset that contains images and their respective captions.

We randomly selected 25 captions in this data set, not used for training, and asked both models to generate their respective images (cf. Figure 4).

Following the same comparative evaluation approach used for CIG in DALL-E, in a best-of-five vote, the images generated by our model were always the most realistic the ones better matching the caption. The images generated by the other system happen to be scrambled pieces of disparate objects.

When compared to our model, DALL-E mini has 3.2 times more parameters (400 million parameters vs. 124 million parameters) and was trained on 5000 times more images (15 million images vs. 2880 images).

VII. RELATED WORK

Generative Adversarial Networks (GAN) [11], [27], [32], [37] showed promise in the field of Computer Vision for image generation tasks. A GAN is formed by two parts, a generator and a discriminator. The generator tries to create fake yet as realist as possible images, while the discriminator tries to distinguish the fake images produced by the generator from the real ones.

Despite this early success also being attributed to the use of Convolution Neural Networks (CNN) [17], the concept of GAN can be used with other deep learning approaches. Such is the case of the more recent work in [15], where two Transformer models [28] play the roles of generator and discriminator. Without using convolution, they nonetheless achieve competitive scores when compared to their CNN counterparts.

Transformers became known due to their success in language processing tasks of all kinds, and recently they have also been applied to other data modalities. Such is the case of DALL-E [23] for image generation from captions, and more recently DALL-E 2 [22] that improves upon its predecessor by incorporating the CLIP model for image and caption representation, and through the use of a diffusion model for image generation [6].

[7] https://huggingface.co/flax-community/dalle-mini

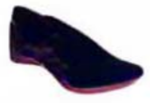
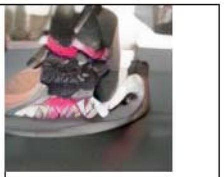
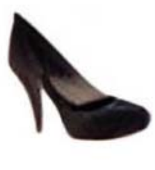
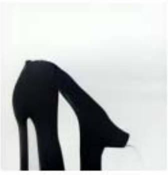
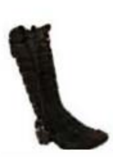
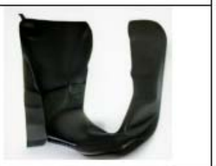
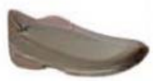
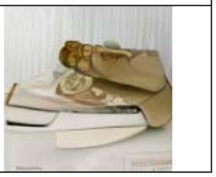
black sneakers with pink accents		
black suede pumps with a cone heel		
black low heel motorcycle boot		
beige sneakers		

Fig. 4. A few examples to compare our system with DALL-E mini in the task of Conditional Image Generation. Right column: caption; middle column: image generated by our model; left column: image generated by DALL-E mini.

Similarly, the work of [25] and [34] confirmed the success of such approaches through the study of other methods of encoding the caption and the impact of model size.

The approach proposed in [10] also achieves promising results in image generation with a pre-trained Transformer CLIP [20], only by training a genetic algorithm.

The architecture adopted in our model is similar to the backbone architecture on which DALL-E is based. Our model is different from DALL-E, however, in not having any specific optimization performed on the base Transformer, like it was done to set up DALL-E, and in being of a greatly reduced size (124 million vs. 12 billion parameters). Our system also differs in that it is geared for a task other than the Conditional Image Generation one, of DALL-E, namely the task of Conditional Image Alteration (CIA). It happens also that it was trained in a much smaller amount of data (10750 vs. 250 million examples).

Also, related to our research topic, [4] tackles the same task, though by means of a Generator/Discriminator architecture, with data that while similar to ours is not the same. To the best of our knowledge, that dataset is not publicly available, so no comparison was possible.

The authors of [14] also work with language guided image edition, but with different datasets that do not tackle the problem of object shape manipulation.

Work on image editing without language guidance can be found in the work of [36], [38], on different datasets.

The research presented here appears as a more streamlined approach for the tasks involved in Language Driven Image Design since most of the work is performed with a common decoder-only architecture, in the form of a GPT-2 small model. This is a generalist architecture that can be adapted for other tasks, as it was the case here with the CIG task, or any other task that can be represented by a sequence (text, audio, image, etc.).

VIII. Conclusion

The results in the present study were obtained by exploring Conditional Generative models of the Transformer class, by resorting to its GPT-2 instantiation, with 124M parameters only, for Language Driven Image Design. The task of interest here was Conditional Image Alteration, consisting of generating a new image given a source image and a textual instruction for its alteration, on which the proposed model showed an effective performance.

Additionally, another task of Conditional Image Generation, namely consisting of generating an image given a textual description, was also experimented here with the same data set. Highly encouraging results were obtained also in this respect, specially given that the data set used here was several orders of magnitude smaller than those that have been used in the literature for this task.

Empirical results also showed that extending the model with language pre-training helps to improve its performance for these image design tasks.

Very encouraging results were thus obtained and demonstrated that the approach proposed and experimented with here can support an effective solution to Language Driven Image Design and represents a promising research path whose potential is worth being further exploited.

Given the effective solutions proposed here, the present study contributes for AI-enhanced solutions that support interactive hands-free image editing via language instructions, which can be used to edit images without using dedicated image manipulation applications.

These research advances can support a wide range of innovative applications, such as enhancing a shopping assistant that progressively matches images refined by language instructions against current stock and suggests increasingly suitable products for the customer, among many others examples enticing our imagination.

As future work further evaluation of the models should be undertaken, primarily through the use of human evaluators in order to mitigate the lack of a suitable metric for the task at hands.

Additionally, considering the present study focus on the manipulation of a single object in the image, rather than of multiple objects in a scene, also as future work the task of scene manipulation [8], [35] should be investigated by exploiting the approach developed here with single object manipulation.

References

[1] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[2] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, 2020.

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[4] Yu Cheng, Zhe Gan, Yitong Li, Jingjing Liu, and Jianfeng Gao. Sequential attention gan for interactive image editing. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4383–4391, 2020.

[5] 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.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] 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. *arXiv preprint arXiv:2010.11929*, 2020.

[8] 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.

[9] 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.

[10] Federico A Galatolo, Mario GCA Cimino, and Gigliola Vaglini. Generating images from caption and vice versa via CLIP-guided generative latent space search. *arXiv preprint arXiv:2102.01645*, 2021.

[11] 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.

[12] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Schmidt Feris. Dialog-based interactive image retrieval. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 676–686, 2018.

[13] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6), feb 2019.

[14] 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.

[15] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong GAN. 2021.

[16] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2973–2980. IEEE, 2012.

[17] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

[18] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[20] 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.

[21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. 2019.

[22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[23] 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.

[24] 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.

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[26] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[27] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis, 2021.

[28] 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.

[29] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. NÜWA: Visual synthesis pre-training for neural visual world creation. *arXiv preprint arXiv:2111.12417*, 2021.

[30] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton Van Den Hengel. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1367–1381, 2017.

[31] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[32] 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. 11 2017.

[33] Aron Yu and Kristen Grauman. Fine-grained comparisons with attributes. In *Visual Attributes*, pages 119–154. Springer, 2017.

[34] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[35] 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.

[36] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pages 592–608. Springer, 2020.

[37] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5795–5803, 2019.

[38] Peiye Zhuang, Oluwasanmi Koyejo, and Alexander G Schwing. Enjoy your editing: Controllable gans for image editing via latent space navigation. *arXiv preprint arXiv:2102.01187*, 2021.