Transformers, Dall-E Mini, and Next Level Text to Video

Preprint · March 2021

CITATIONS
0

3 authors, including:

Mingyue Zhang
Nanjing University
3 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Transformers, Dall-E Mini, and Next Level Text to Video

Mingyue Zhang, Zhiwen Li Xu Zhang

1. Introduction

In our past blog posts, we have introduced the Transformer model architecture and its application in language translation. The Transformer is a powerful neural network framework. When the Transformer is trained over a large amount of natural language data, it learns an incredible amount of insight into the language itself. One of the examples is the GPT 3 model developed by OpenAI.

In most of our previous blog posts, we have been focusing on the intricacies and details of specific deep learning models and architectures, including the algorithm's inner workings and data structure. Developing and practicing model building from scratch has its limit, mostly bounded by the computing power and data sources. As for GPL 3, we're taking a point of view at the application layer, and building something that is fun and demoable. In this blog post, we discuss how we made use of multiple off-the-shelf deep learning models as tools to create a piece of media.

Uncanny Hat Cat (What We Did) We used several readily available models or open deep-learning powered products to construct a derivative re-telling of Dr. Seuss's The Cat in the Hat. The original text was used as the sole input from the original work. Using a language model, we generated summarizations of paragraphs from the original text. Then, we passed those summarizations to another model to create illustrations. The text was also passed to a text-to-speech product to produce narration. We then assembled the derived text, illustrations, and narration into the final video product.

Text Summarization with GPT-3 GPT-3 has the ability to interpret text commands. Fundamentally, GPT-3 is a model that was trained to predict the next word given a huge corpus of data. If provided inputs have text a command, GPT-3 is able to comprehend the command, and "predict" the corresponding output with regard to the command. We asked GPT-3 to summarize and paraphrase one of Dr. Suess' poems, The Cat in the Hat. In this way, GPT-3 helps us to generate novel content that we can use in our final video.

Query Development

Interacting with GPT-3 requires the user to develop queries in order to get the desired response. For the purpose of paraphrasing the poem, we supplied 1 to 3 stanzas individually to GPT-3 and asked it to summarize/paraphrase the text. In this way, we generated a new poem that roughly follows the original poem in content and style.

Image Generation with Dall-E Mini To make illustrations, we made use of the Dall-E Mini model, a model that takes in text and produces images. We fed the text summarization from the previous step into Dall-E Mini, and the model generated several multiple illustrations based on the text prompt. In the spirit of human-computer collaboration, the authors selected the illustration among the candidates to be paired with the text in the final video.

Text-to-speech For this project, we use Google's text-to-speech to generate the voice used in the video. Compared to a couple years ago, text to speech technologies are now quite mature. It's easy to supply a piece of text to a system and generate speech that is easy for humans to understand. Google's system also allows you to select different types of voices. We were quite easily able to generate the voiceover used in our video with a few quick and easy steps.

Creating the Video Finally, we manually stitch together the generated images, text, and voice-over to create the final video. We use iMovie for editing our movie with the necessary images, text, and audio.

On Creating Art with Deep Learning We were able to create a piece of (mediocre) art using several off-the-shelf models and products. And we were able to do it without taxing too many resources. While our final output is interesting for its novelty, the ease of use shows that a more generally entertaining piece of art can be possible with better models and/or more mastery of the models as tools.

Conclusion In conclusion, we were able to create a video using 3 different machine learning models. The models that we use span across computer vision, natural language processing, and speech. Our project shows how much easier it has become to apply technologies from multiple different disciplines to create applications for users. As these technologies continue to advance, we should continue to explore newer and better ways of applying tech. The applications of the future are likely going to employ better and more numerous models.

2. Related Work

Controllable image synthesis has been a long term objective in computer vision and computer graphic. In the earlier

works [24, 46], researchers used many aligned image pairs (i.e., visual domain guidance) as the source domain and target domain to obtain the translation model that translates the source images to the desired target images.

Collecting paired data is usually of high cost in practical application. It is even impossible to acquire plausible paired data in many applications, e.g., translating real images to cartoon images. Thus, unsupervised methods [77, 27, 62] attracts lots of attention as it can be trained under unpaired setting. To achieve reliable generation performance, certain labeling or expert guidance are also expected. e.g., old movie restoration [43] or genomics [53]. Thus, some semi-supervised learning methods [28, 49, 5] are introduced into image synthesis to further promote the quality of generated images. Semi-supervised approaches leverage only source images with a few source-target aligned image pairs for training but can achieve more compelling generation results compared with unsupervised setting. On the other hand, humans can learn from only one or limited exemplars to achieve meaningful results. As described in meta-learning and few-shot learning [74, 54], humans can effectively use prior experiences and knowledge when learning new tasks, while neural network usually overfit to the limited data without generalization capability. Thus, few-shot or one-shot learning models are also explored in many works [38, 34, 35, 36]. The dataset settings can be different, most of these image generaito techniques tend to learn a one-to-one mapping and only generate single-modal output. However, in practice, the translation between domain is inherently ambiguous, as one input image may correspond to multiple possible outputs. Multimodal generation translates the input image from one domain to a distribution of potential outputs in the target domain while remaining faithful to the input. These diverse outputs represent different samples but preserve the similar characteristic as the source image.

Most of computer visions problems can be seen as an image-to-image translation problem, mapping an image from one domain to another image in different domain. As an illustration, super-resolution can be viewed as a concern of mapping a low-resolution image to a similar high-resolution one; image colorization is a problem of mapping a gray-scale image to a corresponding color one. The problem can be investigated in supervised and unsupervised learning methods. In the supervised approaches, paired of images in various domains are available [24]. In the unsupervised models, only two separated sets of images are available in which one composed of images in one domain and the other composed of different domain images—there is no paired samples representing how an image can possibly translated to a corresponding image in different domain. For lack of corresponding images, the unsupervised image-to-image translation problem is considered more difficult, but it is more feasible because training data collection is easier.

When assessing the image translation problem from a likelihood viewpoint, the main challenge is to learn a mutual distribution of images in different domains. In the unsupervised setting, the two sets composed of images from two minor distributions of different domains, and the task is to gather the cooperative distribution by utilizing these images. However, driving the joint distribution from the minor distributions is extremely ill-posed problem. In this section, we discuss the image-to-image translation methods. Image-to-image translation is similar to style transfer, which as the input receives a style image and a content image. The model output is an image that has the content of the content image and the style of the style image. It is not only transferring the images' styles, but also manipulates features of objects. This section lists several models that are proposed for image-to-image translation from supervised methods to unsupervised ones.

2.1. Supervised Translation

Isola et al. [24] proposed to merge the different network losses of Adversarial Network with L_1 regularization loss, therefore the particular generator not only trained to pass the discriminator filtering but also to produce images that contain realistic objects and similar to the ground-truth images. L_1 generates less blurry images as compared to L_2, it was the reason for using L_1. The conditional GAN loss is formulated as:

$$\begin{aligned}\ell_{cGAN}(G,D) = E_{(x,y)\sim p_{data}(x,y)}[\log D(x,y)]+ \\ E_{x\sim p_{data}(x),z\sim p_z(z)}[\log(1-D(x,G(x,z))].\end{aligned} \quad (1)$$

in which $x, y \sim p(x, y)$ denotes to the images that have different styles but belong to the same scene, similar to the standard GAN [18], $z \sim p(z)$ represents random noise, thereby L_1 loss for pressuring self-similarity is defined as:

$$\ell_{L_1}(G) = E_{x,y\sim p_{data}(x,y)}, z \sim p_z(z), [||y - G(x,z)||_1], \quad (2)$$

the general objective is specified by:

$$G^*, D^* = arg^{\min_G \max_D} \ell_{cGAN}(G,D) + \lambda \ell_{L_1}(G) \quad (3)$$

in which the hyperparameter of λ is used to balance the two loss functions. Moreover, in [24], the authors pointed out that, the noise z does not have noticeable influence on the result, therefore, they proposed to use the noise in the form of dropout during training and test in place of samples that belongs to random distribution. In this model, the structure of the G is based on the new structure of U-Net that has multi-scale connections to join each encoder layer to the same layer decoder for sharing low-level information like edges of objects. In [24] the authors proposed `PatchGAN`. The proposed model rather than classifying the whole image attempts to classify the `N x N` path of each image and

seek the average scores of patches for obtaining the final score of the image. From the experiments it has been observed, for obtaining the high frequency details, it is sufficient to limit the discriminator to focus on the local patches.

Yoo et al. proposed an algorithm for supervised image-to-image translation, while having a secondary discriminator D_{pair} that evaluates whether or not a pair of images from multiple domains is related with each other. The loss of D_{pair} is calculated as follows:

$$\begin{aligned} \ell_{pair} &= -t \log[D_{pair}(X_s, X)] \\ &+(t-1)\log[1 - D_{pair}(X_s, X)], \\ s.t.t &= \begin{cases} 0 & if X = X_t \\ 0 & if X = \hat{X}_t \\ 0 & if X = X_{\bar{t}} \end{cases} \end{aligned} \tag{4}$$

where the input image from the source domain is represented by X_s and its groundtruth image is denoted by X_t in the target domain, an irrelevant image in the target domain is represented by $X_{\bar{t}}$. The generator in the proposed model transfers X_s into a single image $\hat{X}_t$ in the associated domain. The authors proposed an efficient pyramid adversarial networks to generating synthetic labels based on target domains for road segmentation in remote sensing images. Zareapoor et al. proposed a semi-supervised adversarial networks for dataset balancing in mechanical devices. The authors integrate multi-instance learning into adversarial networks for human pose estimation. As the results show, the proposed model has high accuracy and fast performance. Shamsolmoali et al. to handle the imbalanced class problems, proposed a capsule adversarial networks based on minority class augmentation. Some authors proposed a general learning framework assign the generated samples to a distribution over a set of labels instead of a single label. The effectiveness of their proposed model is proved through a set of experiments. Zhang et al. proposed DRCW-ASEG method in order to generate synthetic examples for multi-class imbalanced problem. The authors shown that their proposed strategy is able to improve the classification accuracy.

There is no noise input in the generator of pix2pix. A novelty of pix2pix is that the generator of pix2pix learns a mapping from an observed image y to output image $G(y)$, for example, from a grayscale image to a color image. As a follow-up to pix2pix, pix2pixHD [60] used cGANs and feature matching loss for high-resolution image synthesis and semantic manipulation. With the discriminators, the learning problem is a multi-task learning problem. Chrysos et al. [8] proposed robust cGANs. Thekumparampil et al. [59] discussed the robustness of conditional GANs to noisy labels. Conditional CycleGAN [39] uses cGANs with cyclic consistency. Mode seeking GANs (MSGANs) [40] proposes a simple yet effective regularization term to address the mode collapse issue for cGANs. GANs are also utilized to achieve image composition [33, 3, 70, 64], Based on cGANs, we can generate samples conditioning on class labels [45, 44], text [50, 22, 72]. In [72, 71], text to photo-realistic image synthesis is conducted with stacked generative adversarial networks (SGAN) [23]. cGANs have been used for convolutional face generation [15], face aging [1], multi-modal image translation [58, 67], panoramic image generation [14, 55], exemplar-based image synthesis [76, 73, 69], synthesizing outdoor images having specific scenery attributes [25], natural image description [9], and scene manipulation [61]. Most cGANs based methods [11, 48, 52, 13, 56] feed conditional information y into the discriminator by simply concatenating (embedded) y to the input or to the feature vector at some middle layer. cGANs with projection discriminator [41] adopts an inner product between the condition vector y and the feature vector. Two-domain I2I can solve many problems in computer vision, computer graphics and image processing, such as image style transfer [77, 31], bounding box and keypoints [51, 68] which can be used in photo editor apps to promote user experience and semantic segmentation (c.) [47, 79], which benefits the autonomous driving and image colorization (d.) [57, 32], and domain adaptation [42, 6, 37, 65, 66]. If low-resolution images are taken as the source domain and high-resolution images are taken as the target domain, we can naturally achieve image super-resolution [63, 75].

2.1.1 Multimodal Outputs

Multimodal image translates the input image from one domain to a distribution of potential outputs in the target domain while remaining faithful to the input.

Actually, this multimodal translation benefits from the solutions of *mode collapse problem* [17, 2, 19], in which the generator tends to learn to map different input samples to the same output. Thus, many multimodal image translation methods [78, 4] focus on solving the mode collapse problem to lead to diverse outputs naturally. BicycleGAN [78] became the first supervised multimodal image translation work by combining cVAE-GAN [21, 29, 30] and cLR-GAN [7, 12, 13] to systematically study a family of solutions to the mode collapse problem and generate diverse and realistic outputs. Similarly, Bansal et al. [4] proposed PixelNN to achieve multimodal and controllable translated results in image translation. They proposed a nearest-neighbor (NN) approach combining pixelwise matching to translate the incomplete, conditioned input to multiple outputs and allow a user to control the translation through on-the-fly editing of the exemplar set.

Another solution for producing diverse outputs is to use *disentangled representation* [7, 20, 26, 10] which aims to break down, or disentangle, each feature into narrowly de-

fined variables and encodes them as separate dimensions. When combining it with image translation, researchers disentangle the representation of the source and target domains into two parts: domain-invariant features *content*, which are preserved during the translation, and domain-specific features *style*, which are changed during the translation. In other words, image translation aims to transfer images from the source domain to the target domain by preserving *content* while replacing *style*. Therefore, one can achieve multimodal outputs by randomly choosing the *style* features that are often regularized to be drawn from a prior Gaussian distribution $N(0, 1)$. Gonzalez-Garcia et al. [16] disentangled the representation of two domains into three parts: the *shared* part containing common information of both domains, and two *exclusive* parts that only represent those factors of variation that are particular to each domain. In addition to the bi-directional multimodal translation and retrieval of similar images across domains, they can also transfer a domain-specific transfer and interpolation across two domains.

3. Methods & Results

Text to image generation has been all over the internet and news. Some think the media attention is distracting to the AI community, while others start to anthropomorphize, assigning traits like creativity to these models. There is no denying however, that this technological milestone will transform visual arts forever. A natural progression from image generation is video generation and many people are excited for it. In this article, we'll discuss some AI models that generate videos from text prompts. Imagine putting the entire text of the Harry Potter series into this model, or maybe a Game of Thrones fan fiction, obviously rewriting the last two seasons since we know how that ended. But before we get ahead of ourselves, let's discuss the current state of text to video generation.

The current models are very limited in video generation mostly because of two important constraints. Unlike images, videos are represented as multiple frames per time unit, calculated as frames per second. This requires very high computational resources for training video data which is a very big issue. Another issue is the lack of accurate datasets. VATEX, the largest multilingual video description dataset, contains only about 41,250 videos and 825,000 captions in both English and Chinese. The datasets available for video synthesis are either tailored to very specific domains or the captions available do not accurately represent the frames at specific times. Now that we know the challenges facing video generation using text prompts, let's dive into the current AI models that have been released so far.

We'll begin with CogVideo. With 9.4 billion parameters, CogVideo is the largest pre trained model for text-to-video generation across multiple domains. It uses what they call a multi-frame-rate hierarchical training technique, with a resolution quality of 480 by 480, at 8 frames per second, for a total of 4 seconds. I hope I didn't lose you there with too many technical detail. If you have any questions, please leave them in the comments section below and I'll do well to answer. You can tell the project is in its infancy and there are several directions it can go to improve the videos generated. For now, the model takes in Chinese as input, so English texts have to be translated first. I'll share the interactive links in the description so you can play around with it.

Next, we have NUWA-Infinity, a collaborative effort between Microsoft Asia and Peking University. The project's goal was to produce high resolution images and longer form videos. The model is able to perform several visual synthesis tasks with very high quality. Let's discuss a few of them. The model takes an image as input and can generate a similar image with a very high resolution image of 38,912 by 2,048! That's about 19 times the size of the original input created by the AI model. The content of the produced image, shows new additions that still fit the context of the original version. The model can also produce high-resolution animation from just a a simple input image. Finally it can also produce Text-to-Video tasks with amazing quality. Look at this incredible Peppa Pig video generated by the model. Perhaps you can start writing different Peppa Pig scripts for your children or younger siblings so they have an endless supply of episodes. One issue with this model however, is that the datasets used to train this model are narrow in scope and therefore can't generalize well to other types of videos. This will surely change soon, and we're definitely keeping an eye on it.

There are other video synthesis projects that are not necessarily text to video, but still deserve an honorary mention. The first is Transframer from Deepmind which is able to generate 30 seconds of video from just a single image provided. It's great for video prediction and changing views. The resolution quality is however not the highest, but may be improved upon in later iterations. Another state of the art model is from Nvidia Labs. They worked on a video generating model that accurately replicates object motion, camera angle adjustments, and evolving content. The model creates new video content and addresses the issue of long-term inconsistency where scenes may change unrealistically between time frames, for example clouds moving back and forth in an unnatural manner. Moving on, Runway is a video editing platform that announced plans to extend its editing capabilities by using text prompts to change scenes. For example, the link below is a video that changes the background based on the description entered. You can see the background doesn't show any dynamic objects, nor does the text generate a completely new video from scratch. Maybe future updates may incorporate more

complex editing. But for now, this is still a very compelling task. Finally, many creatives are generating videos from interpolating several text to image frames. By adding multiple frames and switching between images, the outcome shows mesmerizing animation results.

You can see we are on the cusp of innovative AI video generation tools. How many months or years do you think we have until we see high quality long form videos from text or image ? Let me know your predictions in the comments section below. Although this is all exciting news for the future, we can't forget to mention the possible negative societal impacts these sort of video generation tools may have. Like other types of generation tasks, the possibility of misinformation and disinformation is still a huge hurdle in the artificial intelligence community. Copyright issues will be brought to the forefront of this and the laws will have to be updated to reflect the changing times. I can't imagine Larry and Jerry particularly happy about seeing AI generated videos of their beloved Jerry Seinfeild sitcom. We might still be several years from this scenario but it doesn't hurt to start thinking about these possibilities. On the other hand, it opens up an endless opportunity for creativity, where anyone can generate videos based on scripts for educational, marketing or entertainment purposes. How would ad agencies or streaming services like Disney adapt to this ? I guess we'll have to wait and see.

References

[1] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2089–2093. IEEE, 2017. 3

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017. 3

[3] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. Compositional gan: Learning image-conditional binary composition. *International Journal of Computer Vision*, 128(10):2570–2585, 2020. 3

[4] Aayush Bansal, Yaser Sheikh, and Deva Ramanan. Pixelnn: Example-based image synthesis. In *International Conference on Learning Representations*, 2018. 3

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059, 2019. 2

[6] Jinming Cao, Oren Katzir, Peng Jiang, Dani Lischinski, Danny Cohen-Or, Changhe Tu, and Yangyan Li. Dida: Disentangled synthesis for domain adaptation, 2018. 3

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Neural Information Processing Systems*, pages 2172–2180, 2016. 3

[8] Grigorios G Chrysos, Jean Kossaifi, and Stefanos Zafeiriou. Robust conditional generative adversarial networks. *arXiv preprint arXiv:1805.08657*, 2018. 3

[9] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *IEEE International Conference on Computer Vision*, pages 2970–2979, 2017. 3

[10] Emily L Denton and vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4414–4423. Curran Associates, Inc., 2017. 3

[11] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using laplacian pyramid of adversarial networks. In *Neural Information Processing Systems*, pages 1486–1494, 2015. 3

[12] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 3

[13] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016. 3

[14] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017. 3

[15] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2, 2014. 3

[16] Abel Gonzalez-Garcia, Joost Van De Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *Advances in neural information processing systems*, pages 1287–1298, 2018. 4

[17] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks, 2017. 3

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Neural Information Processing Systems*, pages 2672–2680, 2014. 2

[19] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Neural Information Processing Systems*, pages 5767–5777, 2017. 3

[20] I. Higgins, Loïc Matthey, A. Pal, Christopher P. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 3

[21] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 3

[22] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7986–7994, 2018. 3

[23] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5077–5086, 2017. 3

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 2

[25] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*, 2016. 3

[26] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018. 3

[27] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865, 2017. 2

[28] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014. 2

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:1, 2014. 3

[30] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016. 3

[31] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, pages 1–16, 2020. 3

[32] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[33] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9455–9464, 2018. 3

[34] Jianxin Lin, Yingxue Pang, Yingce Xia, Zhibo Chen, and Jiebo Luo. Tuigan: Learning versatile image-to-image translation with two unpaired images. In *European Conference on Computer Vision*, pages 18–35. Springer, 2020. 2

[35] Jianxin Lin, Yijun Wang, Tianyu He, and Zhibo Chen. Learning to transfer: Unsupervised meta domain translation. *arXiv preprint arXiv:1906.00181*, 2019. 2

[36] Jianxin Lin, Yingce Xia, Sen Liu, Tao Qin, and Zhibo Chen. Zstgan: An adversarial approach for unsupervised zero-shot image-to-image translation. *arXiv preprint arXiv:1906.00184*, 2019. 2

[37] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. In *Advances in neural information processing systems*, pages 2590–2599, 2018. 3

[38] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[39] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Conditional cyclegan for attribute guided face image generation. *arXiv preprint arXiv:1705.09966*, 2017. 3

[40] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1429–1437, 2019. 3

[41] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018. 3

[42] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[43] Aamir Mustafa and Rafał K. Mantiuk. Transformation consistency regularization – a semi-supervised paradigm for image-to-image translation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 599–615, Cham, 2020. Springer International Publishing. 2

[44] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4467–4477, 2017. 3

[45] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning*, pages 2642–2651, 2017. 3

[46] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 2

[47] 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 (CVPR)*, June 2019. 3

[48] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016. 3

[49] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with

ladder networks. In *Advances in neural information processing systems*, pages 3546–3554, 2015. 2

[50] 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 1–10, 2016. 3

[51] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *Neural Information Processing Systems*, pages 217–225, 2016. 3

[52] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *IEEE International Conference on Computer Vision*, pages 2830–2839, 2017. 3

[53] Mingguang Shi and Bing Zhang. Semi-supervised learning improves gene expression-based prediction of cancer recurrence. *Bioinformatics*, 27(21):3017–3023, 2011. 2

[54] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 2

[55] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6926, 2019. 3

[56] Kumar Sricharan, Raja Bala, Matthew Shreve, Hui Ding, Kumar Saketh, and Jin Sun. Semi-supervised conditional gans. *arXiv preprint arXiv:1708.05789*, 2017. 3

[57] Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Infrared image colorization based on a triplet dcgan architecture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–23, 2017. 3

[58] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2417–2426, 2019. 3

[59] Kiran K Thekumparampil, Ashish Khetan, Zinan Lin, and Sewoong Oh. Robustness of conditional gans to noisy labels. In *Neural Information Processing Systems*, pages 10271–10282, 2018. 3

[60] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 3

[61] Shunyu Yao, Tzu Ming Hsu, Jun-Yan Zhu, Jiajun Wu, Antonio Torralba, Bill Freeman, and Josh Tenenbaum. 3d-aware scene manipulation via inverse graphics. In *Neural Information Processing Systems*, pages 1887–1898, 2018. 3

[62] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 2

[63] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 3

[64] Fangneng Zhan and Shijian Lu. Esir: End-to-end scene text recognition via iterative image rectification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2059–2068, 2019. 3

[65] Fangneng Zhan, Shijian Lu, and Chuhui Xue. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–266, 2018. 3

[66] Fangneng Zhan, Chuhui Xue, and Shijian Lu. Ga-dan: Geometry-aware domain adaptation network for scene text detection and recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9105–9115, 2019. 3

[67] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, and Shijian Lu. Multimodal image synthesis and editing: A survey. *arXiv preprint arXiv:2112.13592*, 2021. 3

[68] Fangneng Zhan, Changgong Zhang, Wenbo Hu, Shijian Lu, Feiying Ma, Xuansong Xie, and Ling Shao. Sparse needlets for lighting estimation with spherical transport loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12830–12839, 2021. 3

[69] Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Rongliang Wu, and Shijian Lu. Modulated contrast for versatile image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18280–18290, 2022. 3

[70] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3653–3662, 2019. 3

[71] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1947–1962, 2019. 3

[72] 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 *IEEE International Conference on Computer Vision*, pages 5907–5915, 2017. 3

[73] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 3

[74] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *Advances in Neural Information Processing Systems*, pages 2365–2374, 2018. 2

[75] Yongbing Zhang, Siyuan Liu, Chao Dong, Xinfeng Zhang, and Yuan Yuan. Multiple cycle-in-cycle generative adversarial networks for unsupervised image super-resolution. *IEEE transactions on Image Processing*, 29:1101–1112, 2019. 3

[76] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet

v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11465–11475, 2021. 3

[77] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*, pages 2223–2232, 2017. 2, 3

[78] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Neural Information Processing Systems*, pages 465–476, 2017. 3

[79] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3