Re-enacting video shots with fictional characters

Joanna Materzynska
MIT
jomat@mit.edu

David Bau
Harvard
davidbau@mit.edu

Antonio Torralba
MIT
torralba@mit.edu

Abstract

Re-enacting video shots with popular fictional characters can serve as a tool for creating fan-fiction or creative content. Previous methods require an image template of the target character to change an appearance of a person in the source image. In this work, we propose a simple method for manipulating a person's appearance in a video shot into one of a fictional character using only its name and a bounding box of the person in the original frame. Our method encodes source image using a VQGAN encoder and optimizes the latent code, such that the detected region of the person in the reconstructed image is similar to the target character name in a shared embedding space. To preserve the semantics of the original frame we apply the perceptual loss between the original frame and target frame. Our results indicate how much do large vision and language models know about pop-culture. We demonstrate our results on several video clips from the TV show Friends, generated characters from various movies and TV shows, and finally demonstrate that our simple method can be applied more broadly as an open vocabulary, text to image manipulation technique.

1. Introduction

Movies entertain us, provide comfort, shock us, educate us, and provide a perspective. As such a rich form of expression, it is not surprising that the characters of influential movies continue living in our imagination. Not only can we easily visualize the appearance of a characteristic fictional character, but we can also imagine his or her behavior in any situation that we find ourselves in. In fan fiction, people re-imagine their favorite characters, for example, creating a new episode of their favorite tv show or painting their favorite characters in a new setting.

In this work, we provide a creative tool for movie reenactment, a text-driven method that allows for changing the appearance of a person in a video to one of a famous movie character. Figure 1 shows an example of our method applied to the characters from the tv show Friends and characters from the movie Lord of the Rings. Inspired by the recent advances in vision and language research, we use a CLIP model [27] to measure the similarity between the textual embedding of movie character's names and images. Using a pre-trained image generation model, and person detector, we maximize the cosine similarity between the generated character and the text embedding. To preserve the semantics of the video, we use the perceptual similarity loss commonly used for style transfer applications.

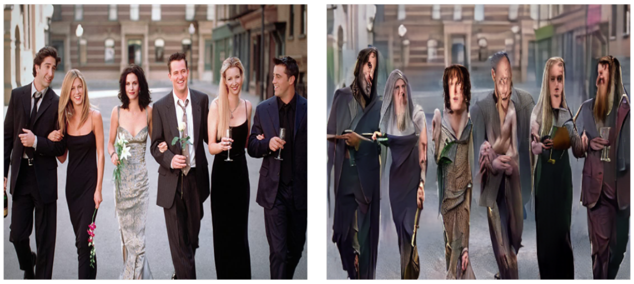

Figure 1. An example of movie character translation. Given the source image, person detection bounding boxes and a user defined mapping ('Ross': 'Aragorn', 'Rachel': 'Gandalf', 'Monica': 'Frodo', 'Chandler': 'Gollum', 'Phoebe': 'Legolas', 'Joey': 'Gimli'), we optimize a latent code to generate the output image.

Our method allows for open-vocabulary transfer of fictional characters in a video. More broadly, the approach can used for general image manipulation with text input.

2. Related Work

Image-to-image translation. Images can be transformed to a new target domain by training a conditional image translation model using either paired (Pix2Pix) training [13, 33], unpaired (CycleGAN) training [36, 16, 12]. A recent proposal trains atop an unconditional GAN [29]. Unlike these methods, we aim to allow transformation of video shots without training a new model for each manipulation, and without access to a large data set of target images.

Style transfer. With only a single target style image, classical texture transfer methods [4, 10] and recent neural image style transfer methods [6, 11, 19, 32, 21] can modify

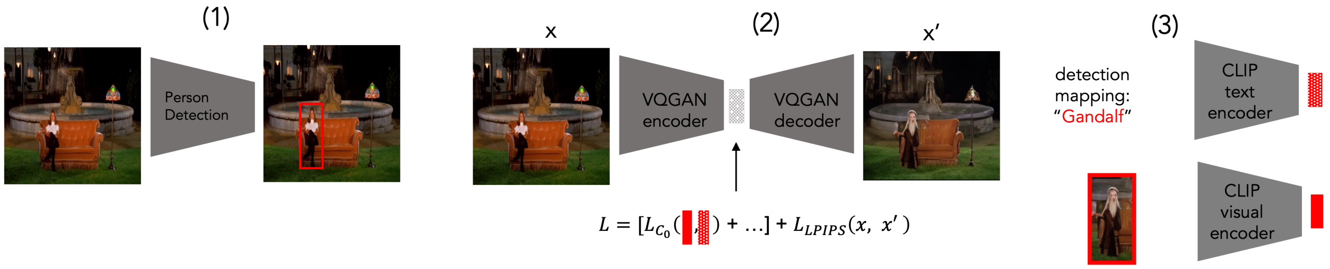

Figure 2. The overview of our method. A pre-trained person detector is used to detect people in each frame. In the training step, we encode the source image using a VQGAN encoder and optimize its latent code, such that the detected region of the person in the reconstructed image is similar to the target character name in a shared embedding space. The loss consist of perceptual metric between the original image and the reconstructed image and the similarity score between the target text embedding and the manipulated image region embedding.

a source image so that its style resembles that of a the target style image while retaining the content of the original source image. Our work differs because we modify high-level semantics such as the identity of a character rather than only low-level style, and we alter the image using text, without an example target image.

GAN-based semantic image manipulation. GAN-based image generation methods can modify an image by steering the latent representation of image semantics [7, 14, 30]. Or a GAN can be used to paint semantic modifications into local portions of an image [24, 2, 3]. Unlike these methods, which allow editing using a finite vocabulary of semantic concepts, our goal is to allow an open vocabulary of characters to be added to a shot.

Movie synthesis. Previous work to create movie scenes from text have retrieved and composited animated characters from a database [9], or trained models to compose simplified scenes from scratch from textual scene descriptions [20, 31]. The problem of creating a new movie is challenging and previous work has focused on animated scenes and synthetic worlds with objects such as blocks and balls. Unlike those approaches, our goal is not to create a new movie scene from scratch, but to use text guidance to modify the characters within an existing realistic movie shot.

Text-based image synthesis and manipulation. More similar to our method are text-to-image modeling methods which generate an image to based on a given text [18, 17, 35]; these have traditionally trained generative models on image and text pairs using a GAN discriminator and text encoders. Autoregressive models have also shown good performance at this task [28, 5]. Since our goal is to modify shots rather than synthesize new images, our method builds upon the more recent approach of modifying images by optimizing the output of a GAN to minimize losses defined by the powerful CLIP [27] image-text similarity network [23, 22, 1, 25]. Our method differs from previous approaches because our goal is to add recognizable human characters, composing multiple characters into an existing shot while preserving layout and pose.

3. Method

The outline of our method is shown in Figure 2. We apply our approach per image, on frames from a video shot. To detect and track the people in a video shot (1), we use an off-the-shelf detector [8] and SPT tracker [15]. We reconstruct each frame using a VQGAN [5] (2) and compute both the visual CLIP [27] embedding of the person detection region of the generated frame and the textual embedding of the target fictional character name (3). We apply image augmentations, including random affine transformation, random perspective transformation, color jitter, and random erasing of a small area of an image to the encoded image. Our objective function maximizes the cosine similarity between the textual embedding and the augmented image embedding. We also apply a perceptual loss between the original image and the generated image, to preserve the semantics of the original image. We could trivially extend this framework to enforce a global style of the image by adding another loss term analogous to (3) that would tie the embedding of the entire image to a style description.

4. Experiments

4.1. Video Shot Re-enactment

We optimize latent codes of each frame for 400 epochs; the generated images have a resolution of 600x600. To improve the quality of generation of people and specifically human faces, we fine-tune the VQGAN on the five first seasons of Friends for 10 epochs. To measure the improvement of the face quality, we compute the PSNR of reconstructions of frames from season 6 episode 1 using a fine-tuned model and a model trained on ImageNet, obtained PSNR scores are comparable with the later model marginally better (+0.13). Qualitative results of image generation using different VQGAN models are shown in Fig 5.

To investigate the effects of using the perceptual loss, we compare three generated images obtained both when using the perceptual loss component during training and when using only the embedding similarity loss. The images ob-

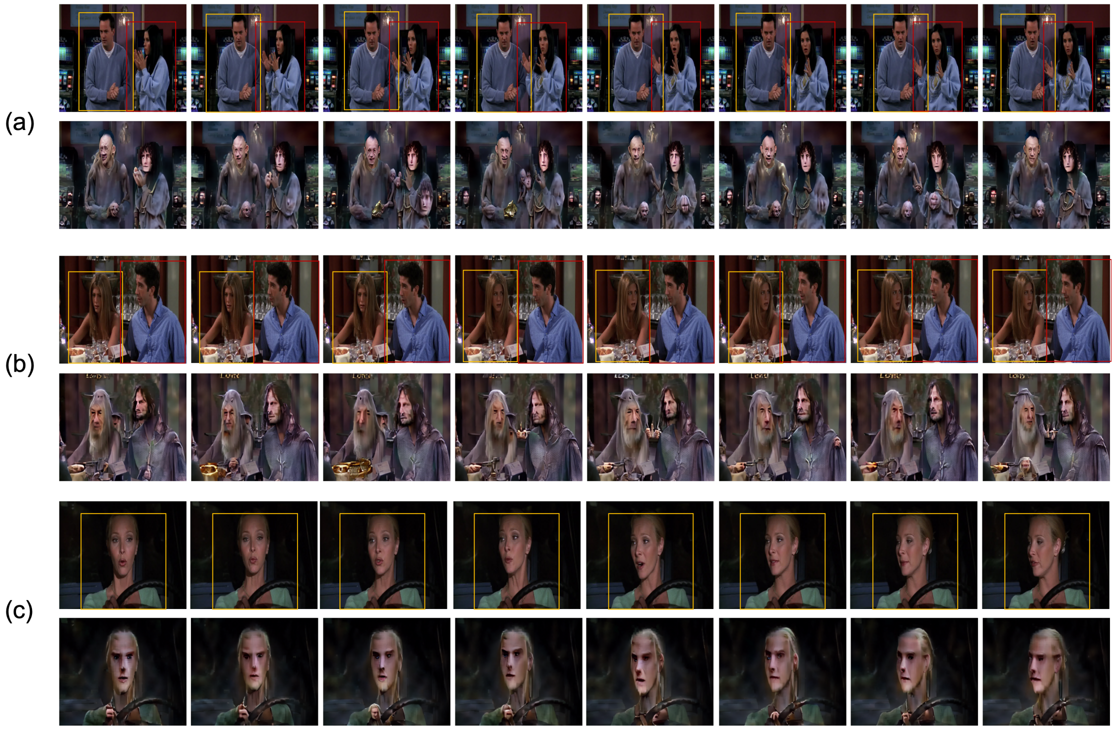

Figure 3. Qualitative results on three shots from Friends. The top row shows original frames, and the one below the generated shots. In the (a) image target characters are ”Frodo” and ”Golum”, the (b) image; ”Gandalf” and ”Aragron”, and the (c) image: ”Legolas”. Despite the coarse conditioning using only bounding boxes, the generated images preserve the pose and facial expression of the source characters.

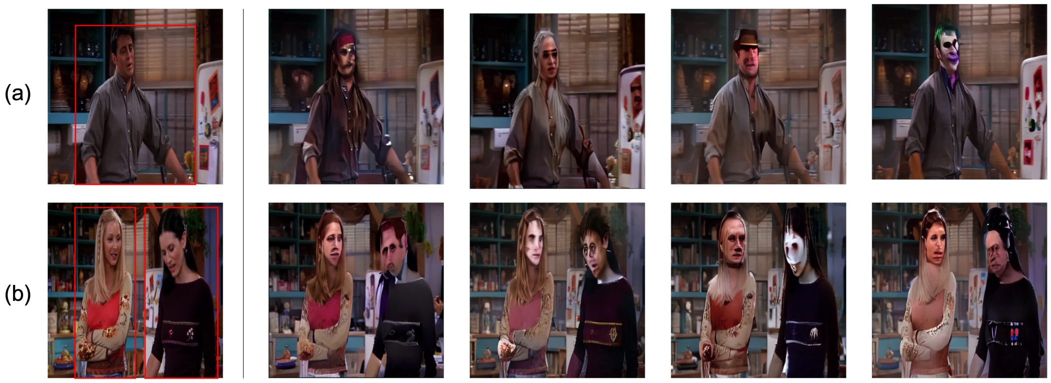

Figure 4. Movie Characters discovery, using our method, we transfer the original frame’s character to one of a fictional character. Target characters in the top image (a) are: ”Captain Jack Sparrow”, ”Daenerys Targaryen, the mother of Dragons”, ”Indiana Jones”, ”The Joker”, the target character pairs in image (b) are (”Pam Beesly”, ”Michael Scott”), (”Hermione Granger”, ”Harry Potter”), (”Hannibal Lecter”, ”the girl from the Ring”), (”Princess Leia”, ”Darth Vader”).

tained when training without the perceptual loss lose the semantic structure of an image and the target characters are generated appear in multiple places in the image. This experiment is shown in Figure 6.

In Figure 3 we show the re-enactment of shots from the TV show Friends, (the shots used for testing are from season 6 episode 5) generated using the proposed approach. In 3 (a) used target characters are (’Gollum’, ’Frodo’), in the 3 (b) (’Gandalf’, ’Aragorn’), and the 3 (c) ’Legolas’. While our method only provides coarse conditioning of the target pose, the generator can transfer head position, facial expressions, and body pose. The characters in each frame are distinguishable despite the fact, that we are only using the a textual description of the character’s name.

4.2. Movie Characters Discovery

To visualize various fictional characters, we test our method on two different frames and four sets of character mapping. In Figure 4 (a) the queried characters are ”Captain Jack Sparrow”, ”Daenerys Targaryen, the mother

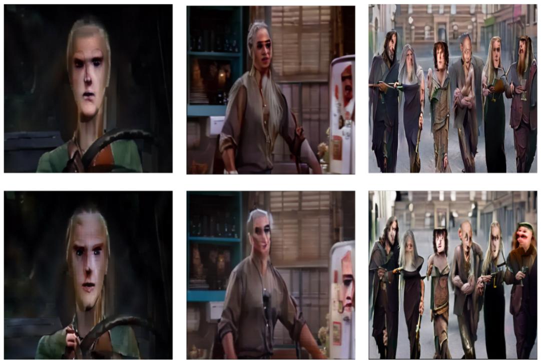

Figure 5. The image generation using a VQGAN model trained on five seasons of the TV show Friends (bottom row) and a VQGAN model trained on ImageNet (top row).

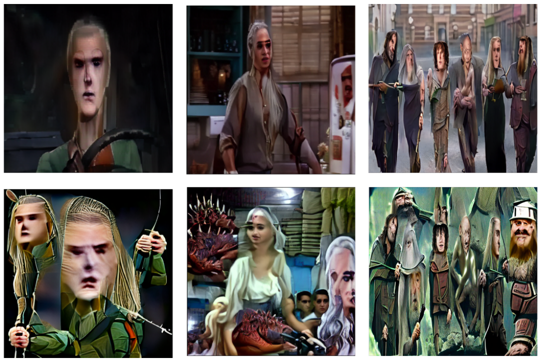

Figure 6. The image generation with the perceptual loss (bottom row) and without using the perceptual loss (top row).

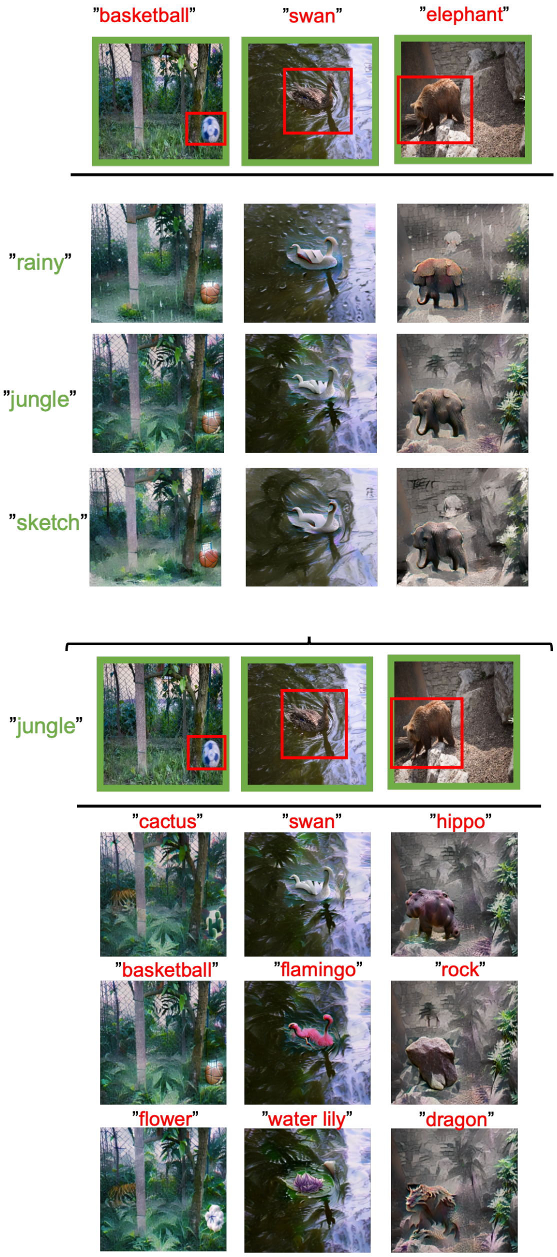

Figure 7. The proposed method can be applied to natural images as a text guided style transfer. At the top, in each column, the generated images share the same modified object and have different global style (green). At bottom, images in each column share the same style (”jungle”) and vary in changed object.

of Dragons”, ”Indiana Jones” and ”The Joker”, in Figure 4 (b) the character pairs are: (”Pam Beesly”, ”Michael Scott”), (”Hermione Granger”, ”Harry Potter”), (”Hannibal Lecter”,”the girl from the Ring”), (”Princess Leia”, ”Darth Vader”). It is interesting to observe that the generated people are recognizable for both very characteristic characters (eg. ”Captain Jack Sparrow”) as well as regular sitcom actors (”Pam Beesly”, ”Michael Scott”).

4.3. General application

Our method can also be applied to natural images. In Figure 7 we take frames from the DAVIS dataset [26], object detections obtain using [34] and use the proposed method to modify both the detected objects and the general style of an image. The top images demonstrate generated images with a fixed style; at bottom are different objects with the same style.

5. Conclusion

We present a simple method for re-enacting video shots with fictional characters using text. Our method allows for free-form text-to-image manipulation that can describe abstract fictional characters as well as specific objects. We use an off-the-shelf person detector and tracker, and apply our method per frame. We show that using the perceptual loss is crucial for preserving the semantics of the original image. We show qualitative results of our method on three video shots from the TV show Friends, seventeen different movie characters and natural images.

We believe our work can promote creative fan fiction generation and improvements in text guided style transfer.

References

[1] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021. 2

[2] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020. 2

[3] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5771–5780, 2020. 2

[4] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346, 2001. 1

[5] 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

[6] 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

[7] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *ICCV*, 2019. 2

[8] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 2

[9] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 598–613, 2018. 2

[10] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001. 1

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 1

[12] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 1

[13] 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

[14] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *ICLR*, 2020. 2

[15] Arne Hoffhues Jonathon Luiten. Trackeval. https://github.com/JonathonLuiten/TrackEval, 2020. 2

[16] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 1

[17] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020. 2

[18] Bowen Li, Xiaojuan Qi, Philip HS Torr, and Thomas Lukasiewicz. Image-to-image translation with text guidance. *arXiv preprint arXiv:2002.05235*, 2020. 2

[19] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *arXiv preprint arXiv:1705.08086*, 2017. 1

[20] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[21] Tianwei Lin, Zhuoqi Ma, Fu Li, Dongliang He, Xin Li, Errui Ding, Nannan Wang, Jie Li, and Xinbo Gao. Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5141–5150, 2021. 1

[22] Ryan Murdock. Aleph2Image. https://colab.research.google.com/drive/1Q-TbYvASMPRMXCOQjkxxf72CXYjR_8Vp, Feb. 2021. 2

[23] Ryan Murdock. The Big Sleep. https://colab.research.google.com/drive/1NCceX2mbiKOSlAd_o7IU7nA9UskKN5WR, Jan. 2021. 2

[24] 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. 2

[25] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021. 2

[26] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 4

[27] 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. 1, 2

[28] 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

[29] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021. 1

[30] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[31] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating compositional scenes from textual descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6710–6719, 2019. 2

[32] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, volume 1, page 4, 2016. 1

[33] 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. 1

[34] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 4

[35] Xiaoming Yu, Yuanqi Chen, Thomas Li, Shan Liu, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. *arXiv preprint arXiv:1909.07877*, 2019. 2

[36] 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