MUSE: Textual Attributes Guided Portrait Painting Generation

Xiaodan Hu¹, Pengfei Yu¹, Kevin Knight², Heng Ji¹, Bo Li¹, Honghui Shi¹
¹University of Illinois at Urbana-Champaign
²DiDi Labs
{xiaodan8,hengji}@illinois.edu, kevinknight@didiglobal.com

Abstract—We propose a novel approach, MUSE, to automatically generate portrait paintings guided by textual attributes. MUSE takes a set of attributes written in text, in addition to facial features extracted from a photo of the subject as input. We propose 11 attribute types to represent inspirations from a subject's profile, emotion, story, and environment. Then we design a novel stacked neural network architecture by extending an image-to-image generative model to accept textual attributes. Experiments show that our approach significantly outperforms several state-of-the-art methods without using textual attributes, with Inception Score score increased by 6% and Fréchet Inception Distance (FID) score decreased by 11%, respectively. We also propose a new attribute reconstruction metric to evaluate whether the generated portraits preserve the subject's attributes. Experiments show that our approach can accurately illustrate 78% textual attributes, which also help MUSE capture the subject in a more creative and expressive way.¹

I. INTRODUCTION

We aim to teach computer to automatically generate portrait paintings, guided by textual attributes. Portrait is a special genre in painting, where the goal is to present not only the outward appearance of a specific human subject, but also their inner significance inspired by admiration or affection for the subject. A good portrait needs to be realistic, and thus it's important to take a photo of the subject as input. Some recent attempts [1], [2], [3], [4] try to automate the conversion from photo to portrait. Most of these methods are only based on visual style transfer [1], [2], [5]. However, art works reflect not only the artist’s hard work and dexterous technique but also often carry their personal emotions and memories on the subjects, due to the intimate relationship between the artist and the subject, either bound together before or during the painting. As Aristotle stated, “The aim of Art is to present not the outward appearance of things, but their inner significance; for this, not the external manner and detail, constitutes true reality.” [6]. From a good portrait, we can often reveal the story of the subject’s life, such as hobby, personality, mood, or a special occasion which may even involve the artist, from a certain facial expression, hairstyle, or the artist’s clever use of colors and lines.

¹We have made all of the data sets, resources and programs related to this new benchmark available at https://github.com/xiaodanhu/MUSE.

We represent a story (i.e. the inspirations) of the subject with 11 text attributes, as shown in Table I. We design a new portrait generation framework called MUSE, which takes these inspirational textual attributes in addition to face regions as input for portrait generation. We first feed the extracted face from the input photo into the encoder part of a UNet [7], a convolutional network architecture that has show promising results on image generation [8], [9]. Then we directly feed the textual attributes into an attribute encoder. Finally the textual attribute embeddings are integrated with the hidden representation of the input photo as an input to the attribute-based decoder to generate the portrait. Figure 1 illustrates an example where the facial expression and hair style are automatically changed based on a set of textual attributes.

Moreover, existing evaluation metrics such as Inception Score (IS) [10] only check the overlapped visual content between the machine-generated and human-generated portraits. But this is not how human assessors approach and appreciate a painting. We design a novel metric to evaluate how many text attributes human assessors can reconstruct from the system generated portraits. Experiments show that our method outperforms state-of-the-art on all measures, and portrait generation is an effective way to acquire and illustrate textual attributes. In summary, the main contributions of this paper are as follows:

- We propose the first inspiration-to-portrait generation framework MUSE that takes text description of attributes into account to generate portraits aligned with its background story such as underlying emotions.
- We develop a novel neural network architecture incorporating textual attributes for portrait generation. Rather than using a binary sequence of facial attributes as input, we apply attribute embeddings, which are initialized from portrait data and optimized during training. Instead of preserving the attributes of input photos, we design a novel discriminator to encourage diversity and realism.
- We create a large portrait generation data set containing 3,928 photo-portrait pairs with manually annotated attributes as a new benchmark, along with 51,939 portraits without annotations, and will share the resources with the community for further exploration.
| Text Attribute | Value Examples                                                                 | #Values |
|----------------|---------------------------------------------------------------------------------|---------|
| Age            | Child; Young adults; Middle-aged adults; Older adults                            | 4       |
| Clothing       | Blazer; Coat and Jacket; Choir & Religious Robe; Dressing Gown; Dress; Shirt; Sweater | 14      |
| Facial Expression | Smile; Smirk; Sneeze; Glance; Wink; Wrinkle the nose; Long face; Blank expression | 14      |
| Gender         | Male; Female; Other                                                             | 3       |
| Hair           | Straight hair; Wavy hair; Black hair; Blond hair; Short hair                    | 5       |
| Mood           | Calm; Excited; Happy; Angry; Apathetic; Sad                                    | 6       |
| Pose & Gesture | Lying; Sitting; Squatting or crouching; Standing; Riding; Shooting; Sleeping; Bowing | 9       |
| Setting        | In the room; In the hallway; On the street; At the river; In the courtyard or a garden | 9       |
| Style          | Impressionism; Realism; Classical Art; Modernism; Chinese paintings; Japanese paintings | 7       |
| Time           | Before 1970; After 1970                                                        | 2       |
| Weather        | Rainy; Stormy; Sunny; Cloudy; Hot; Cold; Windy; Foggy; Snow                    | 9       |

**TABLE I**

**TEXT ATTRIBUTES TO REPRESENT PORTRAIT INSPIRATIONS.**

---

**Fig. 1. Overview of the model architecture.** Given an input photo \( x \) and attribute embeddings \( v \), the model can generate the portrait \( \hat{y} \). The generator \( G \) is an attribute-aware UNet that incorporates attribute embeddings in hidden representation. The discriminator \( D \) is a stack of convolutional layers followed by fully connected layers. By using the adversarial loss \( L_a \) and the attribute classification loss \( L_c \), the discriminator \( D \) can both recognize the realness of the generated portrait \( \hat{y} \) and classify \( \hat{y} \) into different classes of each attribute.

- We propose a novel evaluation metric based on human attribute reconstruction to better assess the quality of generated portraits.

### II. MUSE: PORTRAIT GENERATION APPROACH

**MUSE** takes two sources of input to generate portraits: (1) a photo \( x \) of the subject; (2) a set of textual attributes in the form of type-value pairs \( \{(t_i, c_i)\}_{i=1}^n \). **MUSE** contains a generator \( G \) and a corresponding discriminator \( D \) for adversarial training. To demonstrate the importance and effectiveness of textual attributes, we modify a state-of-the-art image generator, UNet ([7], [9]), to incorporate attribute embeddings \( v \) into generation. The discriminator \( D \) takes the system generated portrait \( \hat{y} \) together with its corresponding human generated portrait \( y \) to evaluate the generator \( G \). The overall architecture of **MUSE** is depicted in Figure 1.

#### A. Attribute Embedding

We propose 11 textual attributes to represent the inspirations, as shown in Table 1. We select these attributes from various online resources including the LitCharts Library, ClarkandMiller.com, and ManyThings.org.

We assign each type-value pair \( (t_i, c_i) \) a unique embedding \( v_i \). Given a set of \( n \) attribute type-value pairs \( \{(t_i, c_i)\}_{i=1}^n \) as input, we concatenate embeddings for \( n \) pairs to get the input attribute embedding \( v = [v_1; v_2; \ldots; v_n] \in \mathbb{R}^{nd_w} \).

The attribute values in Table I are correlated. For instance, the clothing under rainy weather is more likely to be blazer or coat. To capture such inter-dependency between multiple attribute values, we initialize \( v_i \) by using domain-specific attribute embeddings trained from portrait data, where we use skip-gram methods in Word2Vec [11] and consider attribute values corresponding to the same portrait as a bag of words in one context window.

#### B. Attribute-aware UNet

UNet ([7], [9]) is a high-quality image-to-image generative model. Given an input photo \( x \), UNet first encodes \( x \) into a hidden representation \( h \in \mathbb{R}^{dh} = G_{enc}(x) \) using multilayer Convolutional Neural Networks (CNNs), which is further decoded into an output image \( \hat{y} = G_{dec}(h) \) using a stack of transposed convolutional layers.

Our generator \( G \) employs the UNet architecture. The input photo \( x \) is first encoded into the hidden representation \( h \). Then we aggregate the hidden representation \( h \) and attribute embeddings \( v \) of the expected portrait as

\[
h^a = \sigma(W_h h + W_v v + b),
\]

where \( \sigma \) is the sigmoid function.
C. Loss Functions

We apply adversarial training simultaneously for the generator and discriminator \((G, D)\) to learn the mapping from the input photos \(X\) and input attribute embeddings \(V\) to the output portraits \(Y\). Given training samples \(\{(x^{(i)}, v^{(i)}, y^{(i)})\}_{i=1}^N\) where the input photo \(x^{(i)} \in X\), input attribute embeddings \(v^{(i)} \in V\) and the human generated portrait \(y^{(i)} \in Y\), we denote the data distribution as \(x \sim p_{\text{data}(x)}, v \sim p_{\text{attr}}\) and \(y \sim p_{\text{data}(y)}\). While \(G\) tries to generate realistic portraits \(\hat{y}\) similar to the portraits \(y\) in \(Y\) domain, \(D\) tries to distinguish between \(\hat{y}\) and \(y\).

We compute the adversarial loss following GAN [13], with input photo \(x\), attribute embeddings \(v\), and \(y\) as corresponding human generated portraits rescaled to the same size as the outputs of \(G\). Specifically, the adversarial losses of the generator \(G_i\) and discriminator \(D\) are as follows:

\[
L_a^G = -\mathbb{E}_{x \sim p_{\text{data}(x)}, v \sim p_{\text{attr}}} \log D(G_{\text{dec}}(G_{\text{enc}}(x, v)))
\]

\[
L_a^D = \mathbb{E}_{x \sim p_{\text{data}(x)}, v \sim p_{\text{attr}}} \log D(G_{\text{dec}}(G_{\text{enc}}(x, v)))
- \mathbb{E}_{y \sim p_{\text{data}(y)}} \log D(y).
\]

In addition to the adversarial loss \(L_a^G\), we further use \(L_1\) distance to force the generator not only to generate realistic portraits to fool the discriminator but also get close to the human generated portrait. The \(L_1\) loss can be obtained as:

\[
L_{L1} = \mathbb{E}_{x, v, y} \|y - G_{\text{dec}}(G_{\text{enc}}(x, v))\|_1,
\]

While the adversarial learning is employed on the system generated portrait \(\hat{y}\) to ensure its visual reality, \(\hat{y}\) is also expected to correctly contain the desired attributes \(v\). Hence, an attribute classifier \(F\) is used to constrain the system generated portrait \(\hat{y}\) with \(v\). Let \(\tilde{v}\) denote the one-hot vectors of \(v\), the attribute classification loss \(L_c\) can be obtained as:

\[
L_c = \mathbb{E}_{x \sim p_{\text{data}(x)}, v \sim p_{\text{attr}}} [\rho(F(G_{\text{dec}}(G_{\text{enc}}(x, v))), \tilde{v})],
\]

where \(\rho\) is the summation of binary cross-entropy losses of all attributes as follows:

\[
\rho(\tilde{v}, \hat{v}) = \sum_{i=1}^n \frac{-\tilde{v}_i \log \hat{v}_i - (1 - \tilde{v}_i) \log (1 - \hat{v}_i)}.
\]

Overall, by combining the adversarial loss and the attribute classification loss, the final objective functions of the generator \(G\) and the discriminator \(D\) are as follows:

\[
L^G = L_a^G + \lambda_1 L_c + \lambda_2 L_{L1},
\]

\[
L^D = L_a^D + \lambda_3 L_c,
\]

where \(\lambda_1, \lambda_2, \lambda_3\) are hyper-parameters that control the relative importance of the losses.
distribution of Class distribution is then marginalized by assuming uniform on ImageNet [24] predefined classes. Suppose $v \in V$ using the standard metric Inception Score (IS) [10] based given an attribute of face expression.

$$\text{IS}(V_g) = \frac{1}{|V_g|} \sum_{v \in V_g} \sum_{v \in V_g} \text{KL}(p_v || q)$$

A higher IS indicates the generated images are more realistic in the sense that their conditional distributions concentrate on a small subset of classes. Although IS is widely used, it does not compare the generated results with real samples [26]. Fréchet Inception Distance (FID) [26] is another popular metric for conditioned image generation, which measures the Fréchet distance between the generated and real (gold standard) image distribution. Lower FID is better, indicating the generated results and target samples are more similar.

Table II shows the proposed MUSE outperforms the baseline with the IS score increased by 6% and the FID score decreased by 11%. By taking attribute embeddings as additional input, the model can build the alignment between the input photo and target portrait more easily, making the hidden representation more meaningful to produce more reasonable composition. E.g., the attribute value blond hair can provide clearer guidance of generating a particular hair color rather than some unpredictable behavior.

### C. Illustrating Textual Attributes

We propose a new metric to check if the generated portraits reflect the input attribute values, and further explore how attribute semantics are learned and grounded into generated portraits.

**Subject Attribute Reconstruction Accuracy.** Since most of ImageNet classes are objects and the number of classes that appear in portraits is quite limited, the IS metric cannot evaluate the subtle details of human subjects in portraits. We use our attribute classifier $F$ trained on the portrait data

| Evaluation Method | Age | Clothing | Face | Gender | Hair | Mood | Pose | Setting | Style | Time | Weather | Average |
|-------------------|-----|----------|------|--------|------|------|------|---------|-------|------|---------|---------|
| Computer          | 0.88| 0.78     | 0.82 | 0.99   | 0.86 | 0.72 | 0.59 | 0.54    | 0.95  | 0.56 | 0.83    | 0.78    |
| Random            | 0.20| 0.07     | 0.07 | 0.33   | 0.17 | 0.14 | 0.10 | 0.10    | 0.13  | 0.33 | 0.10    | 0.16    |
outputs of the proposed MUSE (e.g., attribute embeddings) separate model for each attribute, for which we constrain the scenario where we change only one attribute. We train a Single Attribute Coherence approach successfully illustrates 78% attributes. The attribute reconstruction accuracy. We can see that our extracted from the human generated portrait. Table III shows the attribute value can be correctly grounded in the generated portraits. We compute the F-score of the estimated attributes against the textual attributes extracted from the human generated portrait. Table III shows the attribute reconstruction accuracy. We can see that our approach successfully illustrates 78% attributes.

Single Attribute Coherence We first consider a simpler scenario where we change only one attribute. We train a separate model for each attribute, for which we constrain the attribute embeddings \( v \) to contain only the specific attribute (e.g., hair) during training. Figures 2 to 5 show example outputs of the proposed MUSE compared with StarGAN [23] and AttGAN [8] by modifying hair, face expression, mood and gender respectively. We can see our models have successfully changed these attributes and learned to generate according to attribute values rather than performing some random behavior. In contrast, the StarGAN mistakenly interprets the attributes and AttGAN does not learn from the attributes at all, due to the small amount of training data and lacking of the mechanisms for cross-domain (photo to portrait) image transformation.

Multiple Attributes Here we consider a more complex scenario that we change all attributes for portrait generation. Figure 6 shows examples using a combination of 11 attributes listed in Table I. MUSE can capture the explicit attribute values such as blond and black hair, and capture the abstract concepts such as stormy vs sunny and cold vs hot by adjusting the background darkness levels.

Grounding of Attributes We further examine how the model ground attributes into generated portraits, by taking an demonstrative example of the hair attribute as shown in Figure 7. From these results we observe that MUSE can rely on either face outlines or facial features such as eyes or mouths to estimate the relative position of components including hair and facial area. In the fourth column we design a blank face with hair outline. Although MUSE cannot perfectly color the hair area, we do observe the model tends to color along the lines. We also show in the last column that MUSE works in a similar way for faces in profile.

Inter-dependency of Attributes As discussed in section II-A, the attribute values in Table I are correlated. Here we show some examples of two related attributes, weather and mood, in Figure 8. For example, the happy portrait in the same row is associated with a brighter background implying a sunny weather. We also use the pre-trained classifier \( F \) to quantita-

| Input Photo | Human Generated | Baseline | MUSE w. Female | MUSE w. Male | StarGAN w. Female | StarGAN w. Male | AttGAN w. Female | AttGAN w. Male |
|-------------|----------------|----------|----------------|------------|------------------|----------------|----------------|--------------|
| Calm; Classical Art; Sitting; Rainy; Cold; Male; Black hair; Short hair | Young adults; Calm; Classical Art; Before 1970; Stormy; Frown; Glance; Male; Coat; Blond hair | Calm; Classical Art; Sitting; Cold; Blank Expression; Female; Dress; Wavy hair; Black hair | Calm; Impressionism; Sitting; Stormy; Cloudy; Female; Dress; Black hair |

Fig. 7. Given drawn sketches and an attribute of hair color or weather, the attribute value can be correctly grounded in the generated portraits.

Fig. 6. Synthetic portraits of the proposed MUSE given combined attributes including age, mood, style, time, pose, setting, weather, face expression, gender, clothing and hair.

Fig. 5. Synthetic portraits of the proposed MUSE, StarGAN [23] and AttGAN [8] given an attribute of gender.
Sunny mood respectively evaluate the inter-dependency between weather and mood. Given the happy portraits, the predicted probabilities to be sunny and stormy are 38.28% and 0.22%, respectively. Similarly, given the sunny portraits, the predicted probabilities to be happy and sad are 39.20% and 0.31%, respectively.

**Affordances of Attributes** Not all the compositions of attributes are semantically valid. For example, it is not possible to have "smile" face expression and "sad" mood in the same portrait. Although it is challenging to inform a model with such affordances, we can use visual signal to improve the generalization capabilities of the model since invalid combinations will not appear in the visual domain. Table IV shows the attribute reconstruction accuracy, IS score and the FID score of the generated "smile sad" and "smile happy" portraits. Using a valid attributes combination, the model can generate a portrait with better quality and better reconstruction accuracy.

| Attributes   | Recons. Acc. ↑ Mood | Recons. Acc. ↑ Face | IS ↑ | FID ↓ |
|--------------|----------------------|---------------------|------|------|
| Smile+Sad    | 0.960                | 0.934               | 1.76 ± 0.193 | 0.085 |
| Smile+Happy  | 1.000                | 0.969               | 1.84 ± 0.212 | 0.073 |

**Table IV** Attributes affordances of the proposed MUSE evaluated on various combination of mood and face expression.

**D. Remaining Challenges**

Good art works should reflect not only explicit information of the subject but also implicit attributes such as personality, occasion, occupation, nationality, hobby. However, the portrait attribute types in the proposed data set are limited due to the lack of knowledge of the subject and the artist. Automatically extracting the implicit information from professionally written text descriptions of the portraits can enrich the attribute types and further enrich the generated portraits. In addition, instead of generating only face regions, artists often include more complete portrait paintings containing background landscape, pose & gestures and other objects. However, with the small amount of training samples, it is difficult to simultaneously handle all styles and content with large variance.

**IV. RELATED WORK**

Generative Adversarial Networks (GAN) [13] achieve great success in generating realistic images without much control [27]. Image-to-image generation [28], [2] takes an image as input condition, and transfers it into another image in a different domain. Another line of work takes natural language as input condition and generates images accordingly. [29] first introduces a conditional DC-GAN architecture which achieves positive results for generating low-resolution images (64 x 64), but it is not equally successful in higher resolution image generation. To address this problem, Zhang et al.[30] propose Stacked Generative Adversarial Networks (StackGAN) that first generate low resolution images and make refinements thereafter. [31] further proposes hierarchically-nested losses that refines images in multiple steps. [32] uses dual inference over conditional and unconditional latent variables for disentanglement of content and style. Rather than starting from low-resolution image generation, [33] first constructs semantic layouts from text and generates images based on them. Despite promising results, the above methods only incorporate sentence-level features without considering fine-grained attributes and thus yield unsatisfactory results when the input sentences are complex.

Attentional GAN (AttnGAN) [34] enables the generative networks to be trained on words of higher relevance, and develops an inter-modal attention mechanism to compute the similarity between the generated image and the relevant text description. Recent methods [35], [36] incorporate attention mechanism to improve semantic consistency. They rely on general caption-type instructions to generate images of flowers [37], birds [38] or common objects [39]. However, such instructions usually lack identity information, which makes these models impossible to generate images of a specific person, flower or bird. In contrast, we propose to take both photo and textual attributes as input. It is worth noting that although AttGAN also leverages facial attributes when generating images, our model is different from the following aspects: (1) our goal is to generate creative and abstract portraits instead of generating photo-realistic images. (2) we apply more complex and abstract attributes to guide the portrait generation. We include more values for each attribute type and some of them have shared semantics; our attributes are carefully designed for portraits and are more abstract (e.g., mood); we use trainable attribute embeddings to better represent the inter-dependency between multiple attribute values.

[40] describes an interesting user study to evaluate the results of generating animations from screenplays where users are asked to evaluate, on a five-point Likert scale [41], if the video shown was a reasonable animation for the text, how much of the text information was depicted in the video and how much of the information in the video was present in the text. Our attribute reconstruction metric aims at a similar goal, but we compute the scores based on pre-defined attribute categories and thus our metric is more objective.

**V. CONCLUSIONS AND FUTURE WORK**

We have developed a novel method, MUSE, which can generate portrait paintings guided by textual attributes. In the future we plan to extend MUSE to unstructured text descriptions and apply open-domain attribute extraction techniques to extract as input, and extend it to cover a wider range of entity types and attribute types.

**ACKNOWLEDGMENT**

This research is based upon work supported by U.S. DARPA GAILA Program HR00111990058. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.
REFERENCES

[1] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogies. ACM Transactions on Graphics, 36(4), 2017. 1

[2] 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, 2017. 1, 6

[3] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In Proceedings of the 2017 Advances in Neural Information Processing Systems. 2017. 1

[4] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In Proceedings of the 15th European Conference on Computer Vision, September 2018. 1

[5] T. Qiao, W. Zhang, M. Zhang, Z. Ma, and D. Xu. Ancient painting to natural image: A new solution for painting processing. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 521–530, 2019. 1

[6] Gordon C. Aymar. The art of portrait painting. Chilton Book Co., Philadelphia, p. 119, 1967. 1

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nissir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pages 234–241, Cham, 2015. Springer International Publishing. 1, 2

[8] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. AttGAN: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing, 28(11):5464–5478, 2019. 1, 3, 4, 5

[9] Xiaodan Hu, Mohamed A. Naiel, Alexander Wong, Mark Lamm, and Paul Fieguth. Runet: A robust unet architecture for image super-resolution. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPRW), pages 505–507, 2019. 1, 2

[10] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. CoRR, abs/1606.03498, 2016. 1, 4

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeff Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013. 2

[12] Vinod Naray and E. Geoffrey Hinton. Rectified linear units improve on convolutional networks for image super-resolution. In Proceedings of the 27th International Conference on International Conference on Machine Learning, 2010. 3

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Proceedings of the 28th Annual Conference on Neural Information Processing Systems. 2014. 3, 6

[14] Getty Museum. https://twitter.com/GettyMuseum. [Online]. 3

[15] Mett Winning. https://www.instagram.com/explore/tags/mettwinning/. [Online]. 3

[16] Pinchuk Art Centre. http://new.pinchukartcentre.org/. [Online]. 3

[17] Rijksmuseum. https://www.rijksmuseum.nl/en/. [Online]. 3

[18] Tussen Kunst & Quarantaine. https://www.instagram.com/tussenkunstenquarantaine/. [Online]. 3

[19] Booooooom. https://www.booooooom.com/2011/09/27/remake-a-project-by-booooooom-and-adobe/. [Online]. 3

[20] Pinterest. https://www.pinterest.com/. [Online]. 3

[21] Wikiart. https://www.wikiart.org/. [Online]. 3

[22] Wikidata. https://www.wikidata.org/. [Online]. 3

[23] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. CoRR, 2017. 3, 4, 5

[24] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE International Conference on Computer Vision, 2009. 4

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. CoRR, abs/1512.00567, 2015. 4

[26] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a nash equilibrium. CoRR, abs/1706.08500, 2017. 4

[27] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. CoRR, abs/1411.1784, 2014. 6

[28] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2017. 6

[29] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Proceedings of the 33rd International Conference on International Conference on Machine Learning, 2016. 6

[30] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xioaoliu Huang, Xiaoang Wang, and Dimitris N. Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, 2017. 6

[31] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 6199–6208, 2018. 6

[32] Qicheng Lao, Mohammad Havaei, Ahmed Pesaranhaghi, Francis Dutil, Lisa Di-Jorio, and Thomas Fevens. Dual adversarial inference for text-to-image synthesis. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 7566–7575. IEEE, 2019. 6

[33] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 7986–7994, 2018. 6

[34] 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. CoRR, abs/1711.10485, 2017. 6

[35] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgen: Learning text-to-image generation by redescription. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019. 6

[36] Hongchen Tan, Xiuping Liu, Xin Li, Yi Zhang, and Baoaii Yin. Semantics-enhanced adversarial nets for text-to-image synthesis. In Proceedings of the IEEE International Conference on Computer Vision, pages 10500–10509, 2019. 6

[37] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, October 7-14, 2011, pages 1500–1507. IEEE Computer Society, 2011. 6

[38] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[39] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tom ‘Tomas’ Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision – ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, volume 8693 of Lecture Notes in Computer Science, pages 740–755. Springer, 2014. 6

[40] YeYao Zhang, Eleftheria Tsipidi, Sasha Schriber, Mubbasir Kapadia, Markus Gross, and Ashutosh Modoi. Generating animations from screenplays. In Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics, 2019. 6

[41] Reinsis Likert. A technique for the measurement of attitudes. Archives of Psychology, 22(140):1–55, 1932. 6