Synthesizing Drawings and Paintings with Conditional Adversarial Networks

Vadim Tschernezki
Heidelberg University

August 28, 2017

Abstract

Adversarial training has shown to be the state-of-the-art method for translating one possible representation of a scene into another, given sufficient training data. This work investigates a particular part of that research topic: synthesizing images from sketches of drawings and paintings. For this purpose, conditional adversarial networks with a U-net shaped generator are chosen. The sketches used for training the networks are generated with Holistically-Nested Edge Detection (HED). Finally, the trained models are tested on machine-computed sketches and human-drawn sketches. The main findings are: 1) the higher the abstraction level of the input images, the less likely the colors and details are captured by the model 2) testing the model on machine-computed sketches results in images having a rich amount of details and correctly chosen colors whereas these qualities are missing in the case of using human-drawn sketches.

1 Introduction

Formulating abstract tasks and letting someone or *something* execute it and fill in the details in this process is a desire that appears frequently from a human point of view, as this saves many resources. An interesting example of such task is: "Given these paintings, create a similar painting from a sketch." Having a machine that is able to execute it, would not only help non-artists turn sketches into something that is more art-like, but might also be useful from a professional point of view: In the process of designing animation, comics and games, redundant work could be saved.

In this work we will explore the possibilities of using conditional adversarial

networks as proposed by Isola et al. [7] for that task, starting in section 2 with related work and going into more detail regarding the mechanics behind such networks in subsection 3.2. The process for generating paintings and drawings from sketches is fully automated: An edge detection algorithm produces edges from the original images. Then, the edges with the corresponding original are fed into the conditional adversarial network while training. Finally, the network is supposed to be able to synthesize images from human-drawn (machine-computed) sketches (edges). The quality of this ability is illustrated in the results section 4.

2 Related Work

2.1 Style transfer

Synthesizing images from edges might be interpreted as transfering the style of a sketch-like version to a colorful, detailed version. One example in the field of computer vision is represented by the work of Gatys et al. [3]. They successfully transfered the style from paintings created by artists such as van Gogh and Picasso to photos. Due to the problem of colors and other image details being typically semantic-based, this approach fails when images with different contexts are provided ("color ambiguity problem") [1]. Therefore images must be labeled such as proposed by Cheng et al. [1] or individual loss functions must be designed [7, p. 2], which makes it a tedious task.

2.2 Manga colorization

Mangas have drawn the attention of several researchers to the topic of colorizing them computationally [1]. For example, Hensman and Aizawa [5] trained a conditional adversarial network only with one training sample and then generate images for the same manga, where the one picture was taken from, with that model. Other experiments were conducted by Furusawa et al. [2]. They developed "Comicolorization", a semi-automatic colorization system for manga images which focusses on colorozing same characters with the same color across a manga.

[1] From a commercial point of view, coloring mangas might be especially interesting as mangas are typically drawn in black and white.

2.3 Image-to-Image translation with conditional GANs

Isola et al. [7] investigated conditional adversarial networks as a general-purpose solution to image-to-image translation problems. In particular, they tested cGANs on a variety of tasks and datasets, including both graphics tasks, like photo generation, and vision tasks, like semantic segmentation. They demonstrate that their approach is effective at synthesizing photos from label maps, reconstructing objects from edge maps, and colorizing images, among other tasks. The conducted experiments in this work are meant to extend the experiments of Isola et al. regarding the task of synthesizing drawings and paintings from sketch-like versions generated with HED as sophisticated edge detection system.

3 Method

3.1 Dataset

The data consists of 137 drawings excerpted from Mittler et al. [8], 574 paintings from Pablo Picasso scraped from WikiArt [10] and further 591 hard-edge-style paintings from WikiArt [10]; each forming one dataset with the corresponding sketch-version [2]. In order to extract high-quality sketch-like images with Holistically-Nested Edge Detection (HED) (introduced in 3.3), only paintings and drawings that show (relatively) clear edges were chosen (such as paintings/drawings from those sources). The images were shuffled in order to avoid overfitting to any sequential pattern, which is given in particular in the book [3]. The datasets were used separately for training three different conditional GANs. Each dataset was split with a ratio of seven to three into training and testing sets. The images of the datasets can have different shapes, because they are cropped and resized automatically before they are fed into the network. New datasets can be created in three steps: Extract images with HED, create dataset from edges and colored images by using the provided tool for aligning them, feed in the dataset into the conditional GAN. The edges were captured accurately in many cases, except if edges from faces were to extract such as in figure 1.

[2] In the experiments of Isola et al. [7] the smallest training set mentioned consists of 400 images.

[3] The drawings represent a story, therefore similar subjects and objects appear sequentially.

Figure 1: Images from the dataset that was created by scraping drawings from Mittler et al. [8] and extracting their edges with HED. The detail-level of images from this dataset is higher compared to the other both.

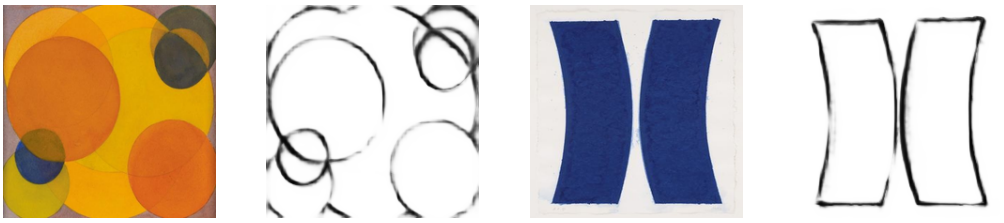

Figure 2: Images from the dataset that was created by scraping drawings from WikiArt [10] and extracting their edges with HED. It contains hard-edge-style paintings.

Figure 3: Images from the dataset that was created by scraping drawings from WikiArt [10] and extracting their edges with HED. It contains picasso paintings after 1920.

3.2 Conditional Generative Adversarial Networks

3.2.1 Objective

Recently, Generative Adversarial Networks (GAN) were introduced as a method for training a generative model [4] [4]. The basic principle of such networks consists in training two adversarial (competing) models simultaneously: A generative model G captures the distribution of the images, while a discriminative model D tries to distinguish between samples that are original and samples that were generated by G. Through this procedure, G learns a mapping from random noise vector z to output image y, i.e. $G : z \rightarrow y$.

In comparison to GANs, conditional GANs are additionally conditioned on

[4] The code used for the experiments borrows heavily from Isola et al. [7]

auxiliary information such as the original images. This conditioning is performed by feeding this information into the generator and discriminator as additional input layer. In case of G, the observed image x and random noise vector z are combined in joint hidden representation whereas in case of D, x and y are combined and used as inputs. Formally, the objective can be expressed as in (1). In terms of a two-player minimax game, where G tries to minimize that objective against an adversarial D, the resulting equation is (2) [7].

$$\begin{aligned}\mathcal{L}_{cGAN}(G,D) =& \mathbb{E}_{x,y\sim p_{data}(x,y)}[\log D(x,y)]+ \\ & \mathbb{E}_{x\sim p_{data}(x),z\sim p_z(z)}[\log(1-D(x,G(x,z))] \end{aligned} \tag{1}$$

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) \tag{2}$$

A method for improving the results of the generator is to mix the objective function with a more traditional loss function such as L1 or L2 distance [7]. The effect of this approach is that the generators predictions are forced to be near the ground truth regarding the norm. L1 distance has shown to reduce blurring [7]. Therefore we combine the objective of the generator with the L1-norm in (3). Formulating the final objective again in terms of a two-player minimax game with the loss defined in (1) and the regularization term from (3) we arrive at the final objective in (4).

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y\sim p_{data}(x,y),z\sim p_z(z)}[\|y - G(x,z)\|_1]. \tag{3}$$

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda\mathcal{L}_{L1}(G). \tag{4}$$

3.2.2 Network Architecture

An important feature of the network is the U-Net shaped generator [5]. The receptive fields have a shape of 256×256 and are able to process 3 channels. The first convolutional layers have 64 filters. Batches are processed normalized. All ReLUs used are leaky with a slope of 0.2, except if used in the decoder part of the generator. All convolutions represent 4×4 spatial filters with a stride of 2.

[5]The network architecture is based on [7]. See their section about the network architecture for in-depth details about the network.

With *Ck* denoting a Convolution-BatchNorm-ReLU layer with k filters and *CDk* a Convolution-BatchNorm-Dropout-ReLU layer with a dropout rate of 50%, then the generator and discriminator layers can be summarized as follows:

- Generator
  - Encoder: C64-C128-C256-C512-C512-C512-C512-C512
  - U-Net Decoder: CD512-CD1024-CD1024-C1024-C1024-C512-C256-C128
- Discriminator: C64-C128-C256-C512-C512-C512

U-Net Generator One of the main tasks of image-to-image translation problems is to map a high resolution input image to a high resolution output image. Such tasks are usually approached with a encoder-decoder network. In order to reduce effects of the bottleneck problem, skip connections are added between the layers as in the general shape of a U-Net [9]. In particular, skip connections are added between each layer i and layer $n-i$, where n is the total number of layers. Each connection concatenates all channels at layer i with those at layer $n-i$.

3.3 Holistically-Nested Edge Detection

Holistically-Nested Edge Detection (HED) is a recently developed end-to-end edge detection system based on fully convolutional neural networks that automatically learns the type of rich hierarchical features that are crucial if approaching the human ability to resolve ambiguity in natural image edge detection settings [11]. Its main features are the holistic approach, i.e. train and predict in an image-to-image fashion despite not explicitly modeling structured output, and nested or inherited and progressively refined edge maps produced as side outputs. Xie and Tu [11] have shown significant improvements regarding quality of detected edges and prediction time compared to other state-of-the-art edge detection algorithms on the BSD500 dataset and the NYU Depth dataset. Therefore this edge detection algorithm seems to be well suited for creating sketches for the conditional GAN.

Figure 4: Each row shows edges, synthesized image and ground truth for the drawings from the book, hard-edge-style paintings and Picasso paintings, respectively. These images were synthesized from the test datasets.

4 Results

The adversarial networks were trained 1000 epochs with adaptive momentum estimation (Adam). The momentum term chosen for Adam equals 0.5 and the initial learning rate equals 0.0002. The decrease of the loss started to stagnate in the range of 700 - 1000 epochs (see appendix A, 7).

The results for the hard-edge-style paintings seem to be least similar to the ground truth with respect to color and shape. A reason for miss-colouring might be the abstract style of these paintings: Since many similar geometric forms are used, the adversarial network can not derive any relation between edges and the respective color in the ground truth (i.e., many circles have different colors, therefore it is nearly impossible to conclude from the shape of a circle to any concrete color). Concrete examples illustrating this problem can be found in figure 4 rows B, such as the squares in (B.41) that are predicted to have colors - probably due to other square like geometries such as in B.12 being colored.

The relationship between level of concreteness and exactness of coloring and level of blurring are also visible in the synthesized images of the paintings of Picasso and book drawings. The synthesized images for Picasso's paintings seem to be more promising compared to those generated from the hard-edge-style paintings. In particular, the colors are more similar to the ones in the ground truth. Also, the geometries and the colors are less blurry. The synthesized book drawings show even more exact coloring. In figure 4 rows A, nearly all clothes of the characters are colored correctly. Also other elements such as the ground show frequently correct coloring. The adversarial network seems to show least problems when predicting images for drawings from the book or more general, when being trained with images showing a high amount of details.

An important detail for a human observer are faces such as the ones of the characters from the book. The conditional GAN was partially able to capture their structures. The drawings and their generated edges have shown that learning faces is a hard task for the network, since HED did not capture the contours of the elements of faces in most cases. The face contours in figure 4 (A.1) show already best-case results; in many cases, no edges of nose, eyes or mouth were extracted such as shown already in figure 1. Consequentially, the adversarial network generates faces with unnatural characteristics.

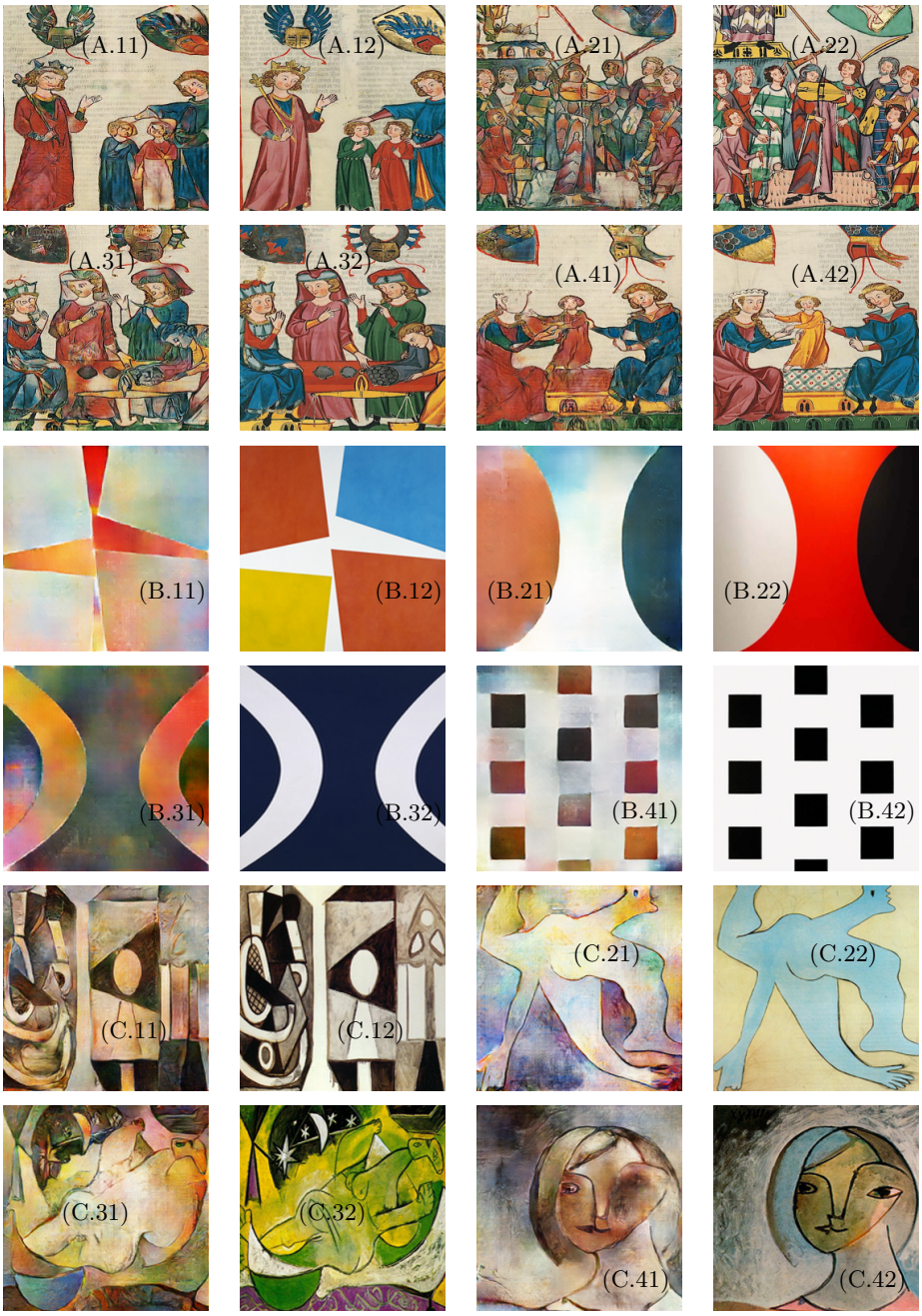

Figure 5: Synthesized and original images (11 and 12, 21 and 22, ...) respectively for book drawings in first two rows (A), hard-edge-style paintings in rows three to four (B) and Picasso paintings in the last two rows (C). The more concrete the paintings are, the more the colors are chosen correctly and the less blury artifacts occurs, i.e. book drawings show best results, then Picasso paintings and at last hard-edge-style paintings. Images were chosen arbitrary from the test dataset.

Figure 6: Rows B and C contain sketches that I drew myself. I tried to capture the style of abstract paintings in case of Row B and Picassos painting in row C. Row A contains sketches that were extracted from [6], i.e. the book that corresponds to the colorized version [8]. All sketches were binarized with Canny edge detection. The images were synthesized with the models that were trained with edges generated with HED. A.11 and A.21 were chosen in order to make the results comparable with A.31 and A41 from figure 4.

The drawings and paintings that were generated with natural sketches (not sketch-like images generated with HED), show the least quality regarding color matching and details as can be seen in figure 4. Considering the results for hard-edge-style paintings and paintings from Picasso as shown in figure 4, these results do not seem unexpected. Still, the images generated with sketches from the book seem to be miss-colorized heavily when compared to the images generated from HED edges.

5 Conclusion

The results of the experiments show that adversarial networks can be trained on paintings and drawings with HED as preprocessing step for edge extraction. Problems occur mostly, if the input images show a high level of abstraction such as in the case of hard-edge-style paintings. In comparison, the adversarial network is even able to colorize the clothes of the characters from the book in most cases correctly.

Another finding is that adversarial networks are barely able to synthesize images with high quality from human-drawn sketches after being trained with machine-computed edges. The reason becomes clear when comparing the edges from figure 4 and figure 4: both versions differ strongly with respect to edge thickness and edge detail. The extent of this problem might be reduced by preprocessing the edges generated with HED such that the edges share the same thickness. Another option might be to train the network with human-generated sketches, although this approach reduces the applicability of this method strongly and makes it less general.

References

[1] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. *CoRR*, abs/1605.00075, 2016. URL `http://arxiv.org/abs/1605.00075`.

[2] Chie Furusawa, Kazuyuki Hiroshiba, Keisuke Ogaki, and Yuri Odagiri. Comicolorization : Semi-automatic manga colorization. *CoRR*, abs/1706.06759, 2017. URL `http://arxiv.org/abs/1706.06759`.

[3] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015. URL `http://arxiv.org/abs/1508.06576`.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. URL `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

[5] Paulina Hensman and Kiyoharu Aizawa. cgan-based manga colorization using a single training image. *CoRR*, abs/1706.06918, 2017. URL `http://arxiv.org/abs/1706.06918`.

[6] Heidelberg historische Bestände digital. *Outlines from the Illuminations in the Manuscript Collection of the Songs of the Minnesingers: Commonly Called the Manesse Manuscript*. Heidelberger Bibliotheksschriften. Heidelberger Bibliotheksschriften, 1876. URL `http://digi.ub.uni-heidelberg.de/diglit/heidhs4036/`

0009.

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arxiv*, 2016.

[8] E. Mittler, W. Werner, H. Drös, and Universitätsbibliothek Heidelberg. *Codex Manesse*. Heidelberger Bibliotheksschriften. Braus, 1988. ISBN 9783925835209. URL https://books.google.de/books?id=c-cfAQAAIAAJ.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

[10] WikiArt. Wikiart — wikipedia, the free encyclopedia, 2017. URL https://en.wikipedia.org/w/index.php?title=WikiArt&oldid=791185313. [Online; accessed 26-August-2017].

[11] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. *CoRR*, abs/1504.06375, 2015. URL http://arxiv.org/abs/1504.06375.

A Loss Curves

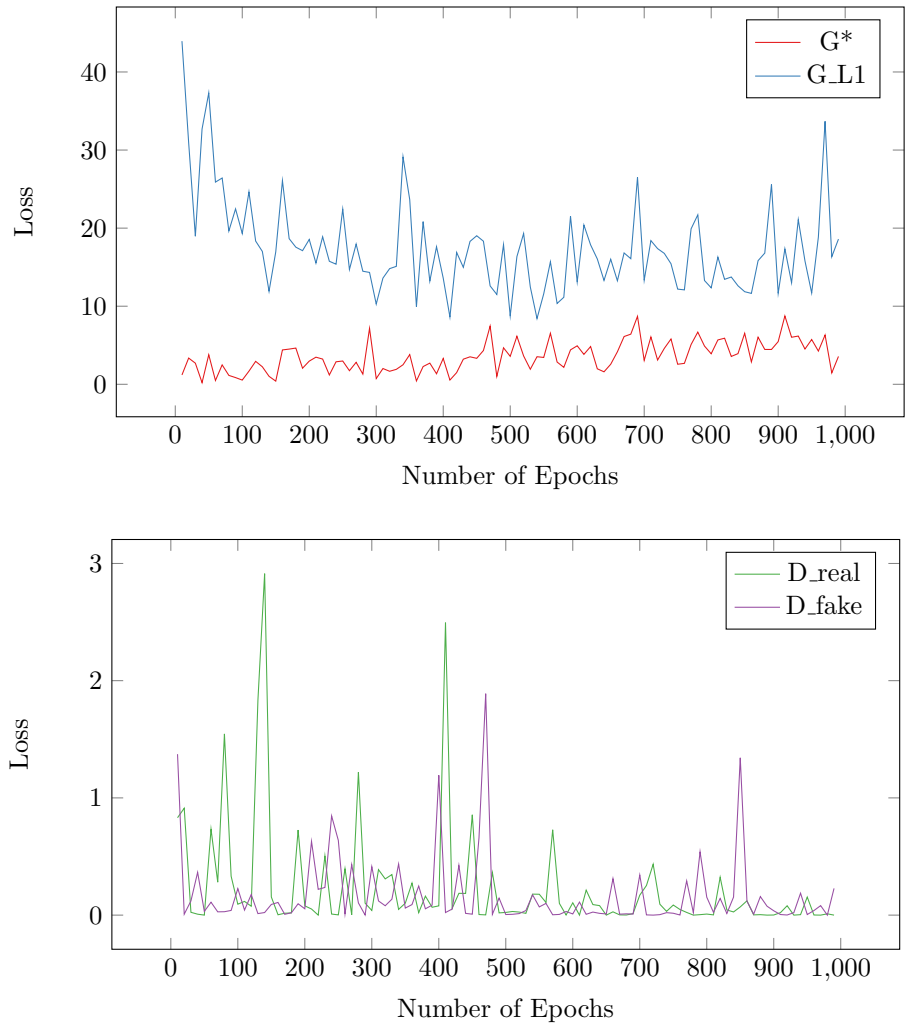

Figure 7: Both plots show the losses for training a conditionals adversarial network with paintings from Picasso (parameters are set as described in 3.2). The first plot shows the loss of the whole network G* and the generator G_L1. The second plot shows the loss of the discriminator for two cases: first, if calculated for real/original images and second, if calculated for images synthesized by the generator. The steadily decreasing loss of the discriminator and slowly stagnating loss of the generator might explain the increasing summed loss (G*): the discriminator seems to learn slightly better than the generator.