A Deep Learning Pipeline for the Synthesis of Graphic Novels

Thomas Melistas,[1]* Yannis Siglidis,[2]* Fivos Kalogiannis,[1]* Ilan Manouach[3]*
[1]School of Electrical and Computer Engineering, National Technical University of Athens, Greece
[2]Department of Mathematics, ENS Paris Saclay, France
[3]Department of Media, University of Aalto, Finland
melistas.th@gmail.com, ioannis.siglidis@ens-paris-saclay.fr, phoekalogiannis@gmail.com, ilan.manouach@aalto.fi

Abstract

In this paper, we present what is to the best of our knowledge, the first deep learning pipeline to produce a synthetic graphic novel. Our method can synthesize from scratch engaging sequences of graphic novel pages, focusing on the Manga genre. To achieve this, we extract images and text from around 670 thousand Manga pages, which we use separately in order to train state-of-the-art generative architectures, such as GPT-2 for text generation and StyleGAN2 for image synthesis. Using these as sources of synthetic content, we develop a set of algorithmic aesthetic rules in order to bring together complete and continuous Manga pages.

Introduction

There is little consensus among comics scholars on whether comics is a language, but it's relatively agreed that comics is a sequential system of communication, consisting of both linguistic and non-linguistic signs (Groensteen et al. 2007). Comics create, most commonly, a narration, the contents of which are images and text, while its form is the panel layout, the placement and shape of text bubbles and the succession of panels and pages. More generally, we can consider comics synthesis as the generation of images and text, as well as their common arrangement, in a way which suggests a sense of narration and/or dialogue. While comics traditionally unfold a structured storyline and contain text and images that are directly related, types of comics that explore more unconstrained and creative directions have emerged. Our approach is motivated by such works, since we produce and experiment with more abstract and unstructured narrations.

The task of artificial comics synthesis has never before been considered in its full spectrum. Previous attempts have focused on automating specific parts of the pipeline, but as far as we know, no previous work has ever attempted to automate the full procedure. Progress made on the last few years in generative modeling, especially the development of the Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) and the Transformer architecture (Vaswani et al. 2017), has implied their potential use in creative and commercial applications that rely on content synthesis. Our work is mainly positioned inside this context, aspiring to question and challenge creativity in the multi-modal and complex setting of graphic novels.

We focus on a specific graphic novel form which originated in Japan, called *Manga*. Manga comics come in a huge variety and quantity, being classified into many genres on the basis of their targeted audience, the main subject of their plot and their artistic style. They were originally distributed in black-and-white prints. Nowadays, vast web-communities of Manga enthusiasts, known as *scanlators*, share scanned Manga comics in low quality through the web, usually poorly translated in English. Both because of their world-wide popularity and their abundance, but also because of their automated production process[1], they are a form of art for which we could speculate their, at least partial, automation through Artificial Intelligence in the coming years. To support even more this claim, Manga follow a certain consistency in drawing style, which differentiates them a lot from other types of comics, and are found in abundance, with many successful Manga series consisting of thousands of pages.

We present our contribution, a deep learning pipeline for synthesizing complete and continuous pages in the form of a graphic novel. Our work consists of two main parts: (a) the necessary pre-processing or content extraction step, namely the extraction of images and text from raw Manga comics and (b) the synthesis of the content and its assembly into the form of a graphic novel. We thoroughly describe all steps of the above process and underline the challenges we encountered, as well as the techniques we adopted to surpass them.

Firstly, we describe the pre-processing procedure. The first step is the extraction of Regions of Interest from Manga pages, namely image panels and text bubbles, using a region proposal based convolutional network. Next, we train a U-NET segmentation network (Ronneberger, Fischer, and Brox 2015) to precisely segment and isolate text bubbles, which are then inpainted in order to obtain clean images. Finally, we get text transcriptions of the detected areas, using

*Equal Contribution

[1] A typical production line of manga comics for example involves dozens of people handling specialized roles in a quasi-taylorist production belt, often in ways that have been criticized for resembling a sweatshop, while distribution has been increasingly involving massively digitized operations of logistics and global supply chains. Comics is, an *industrial* form of artistic expression.

an Optical Character Recognition model, which we domain-adapt by fine-tuning it to commonly used Manga fonts, increasing its recognition performance. For all these steps, we have used all resources that where available to us, namely existing datasets, manual data annotation and ad-hoc synthetic datasets.

For the second step, we start with training a StyleGAN-2 architecture (Karras et al. 2020) on the inpainted Manga images that we previously extracted. To alleviate the low quality and diversity of the generated samples, we explore two different approaches: (1) we train a conditional model, providing labels that we acquire from a tag estimator trained on Anime art and (2) we perform transfer learning using a model pre-trained on Anime faces. To fit the industry standards for image quality, we chain the generation procedure with a super-resolution up-scaling network trained on Manga content. Next, regarding text generation, we fine-tune a GPT-2 language model (Radford et al. 2019) on text extracted from our Manga dataset, as well as on a diverse set of monolingual corpora from different genres of literature. We finally generate both image and text content in a sequential manner and place them inside randomized and standard panel layouts, bringing them together in the form of graphic novel pages.

Related Work

In this section, we discuss relevant research, regarding the comics medium and the procedure of content synthesis. We can make a first distinction between: (a) research that focuses on the analysis of comics and extracts information that is crucial for specific tasks and (b) work that deals with the synthesis of Manga related content, such as animated characters.

The first approaches on the field of comics analysis, focus on using traditional computer vision techniques for the extraction of basic comics features. One of the earliest tasks, providing motivation in this research field, was panel and text extraction, mostly oriented towards automating the process of formatting comics for reading in mobile devices (Yamada et al. 2004; Ho, Burie, and Ogier 2012; Li et al. 2014) or copyright protection (Sun and Kise 2013). With the rise of deep learning, research steered, with great success, towards the use of neural networks for similar or other, previously unexplored tasks. The creation of appropriate datasets, such as eBDtheque (Guérin et al. 2013) and Manga109 (Fujimoto et al. 2016) (both of which we also use in our work) helped to build more robust tools for detecting and extracting comics features, such as panels (Ogawa et al. 2018; Zhou et al. 2020), text bubbles (Dubray and Laubrock 2019) or characters (Qin et al. 2017), using deep convolutional architectures. Their extracted contents can be used in many ways, for example to create an indexing system with content analysis (Nguyen, Rigaud, and Burie 2018), sketch-based Manga retrieval (Matsui et al. 2016), semi-automatic comic colorization (Furusawa et al. 2017) or making comics more accessible to the visually impaired (Rayar, Oriola, and Jouffrais 2020).

Regarding the synthesis of comics, little work has been done towards the form of narration itself, with a few notable examples, such as the synthesis of Manga-resembling layouts (Cao, Chan, and Lau 2012). Previous work has mainly focused on synthesizing Manga related artwork. The evolution of the GAN architecture during the past years, has increased the interest for generation non-photographic animated characters, typically found in Manga, although it is most commonly used for generating photographic images. In (Su et al. 2020) a GAN architecture is trained to create Manga faces from photographs, while preserving the original face features. A notable example of image generation is the work of Gwern, *"this waifu does not exist"* (Gwern 2019b), a website hosting a StyleGAN-2 generator for female Manga character portraits with some text, independently generated by GPT-3 (Brown et al. 2020), accompanying each. The basic image generation component of this work was recently updated, introducing a revised StyleGAN-2 architecture, and was featured in the *"this anime does not exist"* project (Aydao 2021). Important changes of Aydao's approach are the doubling of the feed-forward embedding layer's width (consequently doubling the dimension of the latent vector) and decreasing the amount of regularization, leading to slower but more stable training.

Content Extraction

Throughout our work, we use a custom private dataset, consisting of 667,181 black-and-white Manga pages with English text in 72 dpi. Unfortunately, we are not in the position of making this dataset public or publishing the extracted content, as it is part of a private collection assembled from diverse sources, subject to copyright law. Furthermore, we do not perform any human annotation on the above dataset, so the evaluation of all the techniques we present next is mainly done visually from random samples, since our data is unlabeled.

Panel and Text Bubble Detection

As a first step we detect panels (areas in which images are located) and text bubbles (areas in which text is located). As a primary annotated resource for this task we use the Manga109 dataset, which contains handcrafted annotations for panels, characters and text bubbles in the form of rectangular bounding boxes. We found, that training a Faster R-CNN model (Ren et al. 2015) on this small dataset is effective for extracting bounding boxes on our larger dataset. Faster R-CNN incorporates a Regional Proposal Network that shares features with a detection network and is widely used in relevant tasks, even for extracting comics features. Our implementation is largely based on the `MMDetection` framework (Chen et al. 2019).

It should be noted, that while it is common for Manga panels to have boundaries that are not parallel to the borders or are more complex than quadrilateral, there is no large annotated dataset which contains non-rectangle polygons or masks for object detection. Moreover, modeling rectangular (or even square) images of a fixed resolution is the predominant approach, used by most of the existing computer vision and image generation architectures. Adding white margins to the rectangular images is a possible solution, but while it can increase expressiveness and data variability, it introduces white areas that dominate the generated images and complicates the final page assembly. As we mention below, we arrive to the solution of cropping and resizing accordingly.

Finally, we develop an algorithm that sorts the detected panels and text bubbles according to the Manga reading order, while it associates each bubble with the panel it belongs to. Manga comics are read from right to left and from top to bottom. This applies to both panels and text bubbles. We consider a panel preceding another if it is located higher on the page or if it is on the same level but on the right-hand side. Panels are considered to be on the same level if the horizontal border lines of one are contained on those of the other or if the difference between their respective upper or lower borders is smaller than half the height of the shortest panel. Each bubble is then associated with one panel, based on the distance of their centers. Bubbles that belong to the same panel are sorted following the same procedure. The above algorithm is robust to unconventional panel layouts, common in Manga comics and enables us to extract and encode structural information, such as the sequences of images and the text-image correspondence, as well as getting a complete and ordered story from the extracted text.

Text Bubble Segmentation

In this step, we remove all the text bubbles from the extracted images, as such a visual feature would insert noise and dominate the generated samples. To achieve this, we first detect the exact region the bubbles occupy and then use proprietary software to do content-aware inpainting of the area underneath. The Faster R-CNN model, that we have used above, is not sufficient for this task, as it only can provide us with a rectangle bounding box. An effective solution to this problem would require pixel-level masks which cover the exact area that needs to be inpainted. This is rather an image segmentation task, which we approach using a U-Net architecture (Ronneberger, Fischer, and Brox 2015), following the implementation of the `fastai` library (Howard and Gugger 2020).

We use the labeled *eBDtheque* dataset (Guérin et al. 2013) to train our model. It contains pixel level masks for text bubbles of comics that span various styles and traditions. Since it is a small dataset, consisting of only 100 annotated comic pages, we augment it using custom-made synthetic data. To produce them, we place handcrafted text bubbles, with fill-in real text, on pages whose bubbles we have removed successfully during a previous iteration. To achieve a first rough estimation of the bubble area, we have implemented a flood-filling algorithm to find connected components based on pixel intensity around letter markers, extracted by the Faster R-CNN model that we have previously trained. To make our model more robust, we train it on augmented versions of both the original and our synthetic dataset. Images are randomly cropped and rescaled, followed by a random affine transform, to which random brightness is applied subsequently.

Text Transcription

The last step of this process is the extraction of text from the detected text bubbles. To achieve this, we use the `Tesseract` Optical Character Recognition (OCR) Engine (Smith 2007). Using a model that was pre-trained on English text led to character recognition of poor performance. This is not unexpected, since comics and especially Manga contain specific and uncommon fonts, as well as not casual monochromatic backgrounds and disrupted or skewed text, which the initial trained model is naturally unaware of. Unfortunately, we could not use the Manga109 dataset for this purpose, as it contains text transcriptions only in Japanese, and as far as we know, no other annotated dataset was available.

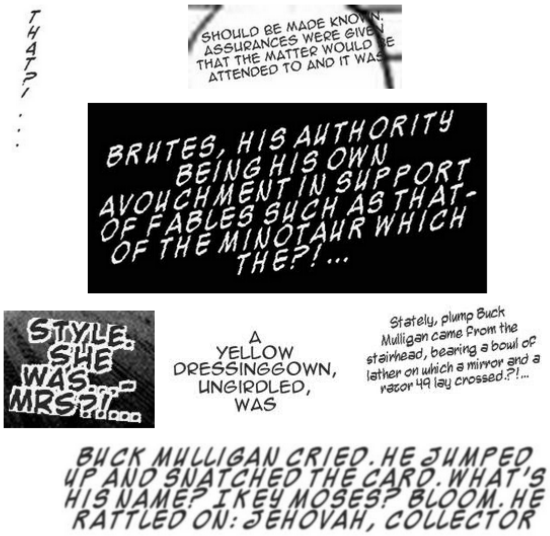
SHOULD BE MADE KNOWN. ASSURANCES WERE GIVEN THAT THE MATTER WOULD BE ATTENDED TO AND IT WAS
THAT?/...
BRUTES, HIS AUTHORITY BEING HIS OWN AVOUCHMENT IN SUPPORT OF FABLES SUCH AS THAT- OF THE MINOTAUR WHICH THE?!...
Stately, plump Buck Mulligan came from the stairhead, bearing a bowl of lather on which a mirror and a razor 49 lay crossed.?!...
A YELLOW DRESSINGGOWN, UNGIRDLED, WAS
STYLE. SHE WAS...- MRS?!...
BUCK MULLIGAN CRIED. HE JUMPED UP AND SNATCHED THE CARD. WHAT'S HIS NAME? IKEY MOSES? BLOOM. HE RATTLED ON: JEHOVAH, COLLECTOR

Figure 1: Examples of synthetic text images that were used to improve the OCR accuracy.

To improve the accuracy of OCR, we fine-tune the English model on synthetic pairs of text and generated text bubble images, with fonts resembling those found in Manga comics. As a text corpus we have used *"Ulysses"* by James Joyce, because of its casual tone and plethora of neologisms and onomatopoeia. The text was partitioned into chunks of varying sizes and then split into lines of varying lengths, in order to better represent text bubbles. Specifically, the number of words contained in each bubble was chosen uniformly in the range of 1 to 20 and line breaks were added to split each bubble to approximately 4 lines, with some added randomness. Also, some punctuation commonly used in Manga, such as triple dots and exclamation marks, were manually added. Using 40 fonts, common in Manga, we generate a total of 20,000 text images with varying word and line count, different backgrounds, font sizes, orientation, blurring and skewing. Some examples can be seen in Figure 1. We find that the fine-tuned model significantly improves the quality of text transcription.

Synthesizing a Graphic Novel

In this section we discuss the method we adopt in order to generate synthetic content, as well as the algorithmic procedure that we follow in order to assemble a graphic novel.

Synthesizing Images

We consider the development of the GAN architecture to be a pivotal moment in synthetic content creation, which at first proved its remarkable ability to generate mono-categorical photo-realistic images, such as human faces. One of the most influential architectures, widely used in a wide variety of creative applications is StyleGAN2 (Karras et al. 2020), yielding state-of-the-art results in generative image modeling. Its advantages among alternatives include its ability to be trained on images of higher resolution and the feasible computing resources needed compared to other methods, which has led to its adoption by a community of *StyleGAN artists*.

After the aforementioned extraction and the separation of images with aspect ratios close to 1x1, we end up with 1.7 million monochromatic images, which are resized and cropped to fit the 512x512 resolution. Training a StyleGAN2 architecture from scratch to this data proved insufficient, resulting in non-convergence. This was either reflected on a high Frechet Inception Distance (Heusel et al. 2018) measure (more than 30) or on complete divergence. Just tuning the hyper-parameters (for example decreasing the regularization or the learning rate) did not solve this issue. We attribute this poor training performance both to the absence of a center in our data, as well as to their poly-categorical nature and to their complex textures, something which is not the case in most traditional datasets which have been used for evaluation purposes, such as FFHQ (Karras, Laine, and Aila 2019). To overcome this, we experiment with two standard approaches: (a) boosting the learning process with label conditioning and (b) transfer learning.

The first approach requires a meaningful categorization or labeling of our images. The idea that organizing the diversity of our dataset could result in improved performance was explored in (Oeldorf and Spanakis 2019), which showed that meaningful conditions enable the model to learn a larger number of modes and produce more detailed, diverse and controllable outputs. Additionally, it seems that although simple label conditioning is supported in the standard implementation of StyleGAN2, it is largely unexplored by the (art) community.

To label our images, we utilize a multi-tag ResNet classifier (He et al. 2015), pre-trained on the 5543 most popular descriptive tags of the Danbooru dataset (Anonymous, community, and Branwen 2021). Danbooru is currently the largest available public dataset of anime-style images, commonly found in Manga. Each image is accompanied by around 30 tags coming from a total of 434k predefined tags. Unfortunately, a single annotation by itself is rarely descriptive of a single image. Thus, we predict the tag scores for all our images and apply incremental PCA (Ross et al. 2008) to reduce their dimensionality to 20. Next, we perform clustering, using the incremental k-means algorithm (Pham, Dimov, and Nguyen 2004) in order to extract 20 categories in total. By adding label information, we obtain more diverse content with a lower FID, but which is much more figurative. A few examples can be seen in Figure 2. Although we did not try other variants of this procedure, we strongly suggest it as a subject of future work.

For the second approach, we fine-tune a publicly available model that has been trained on 512x512 female Manga faces (Gwern 2019a) on our dataset. This model is the core of the aforementioned *"this waifu does not exist"*. We observe that by following this approach, we manage to model face characteristics with a higher fidelity, which can in fact help to make our comic more engaging. On the other hand, this approach exhibits less diversity, and as expected is much more biased towards generating female faces. Some examples that showcase this can be found in Figure 3.

Sampling Procedure We sample the latent space of our model using a 513-dimensional vector (512 for the noise vector, plus one for truncation ψ) of an interpolated time-series of stock volumes. This results in a continuous sequence of transformations, a procedure that is generally used for the exploration of the latent space and which creates a sense of action for various characters that appear locally for certain latent vector values. After generating about 130K images, we manually classify them using the active learning technique of Pool-Based sampling (using as informativeness the maximum probability of any category), to four ad-hoc visual categories, plus a fifth which is used to discard images with ambiguous content. As the produced images have a 512x512 resolution, we bring them to the industry standard of 1024, using the domain-specific up-scaling model *"waifu2x"* (Nagadomi 2018).

Synthesizing Text

Language models assign probabilities to sequences of words and are commonly used to generate text, by iteratively choosing a word given the previous context. GPT-2 (Radford et al. 2019) was until recently the state-of-the-art model for language generation, superseded by its successor (Brown et al. 2020). It consists of billions of parameters and it has been trained on a huge corpus of texts scraped from the web. Even from early experiments, we have noticed its unique ability of domain adaptation when fine-tuning to a certain author or genre; its results could be seen in broad terms as the *impersonation* of an author or the imitation of the genre.

In our work, we use a distilled version of this model (distilGPT-2), which is contained in the `Hugging Face` library (Wolf et al. 2020). Knowledge distillation (Hinton, Vinyals, and Dean 2015) is a technique for reducing the size of a model, while decreasing its accuracy by a small factor. The distilled GPT-2 model is twice as fast and consists of 37% fewer parameters than the smallest full model available. We notice that by using the original versions of GPT-2, the sampled outputs can be a bit more poetic and fruitful, but much less reliable in coherence and syntactic precision. We train this language model on the Manga text that we have extracted above, where we use special tokens to separate text that belongs to different bubbles and panels.

To introduce more interesting and diverse textual content,

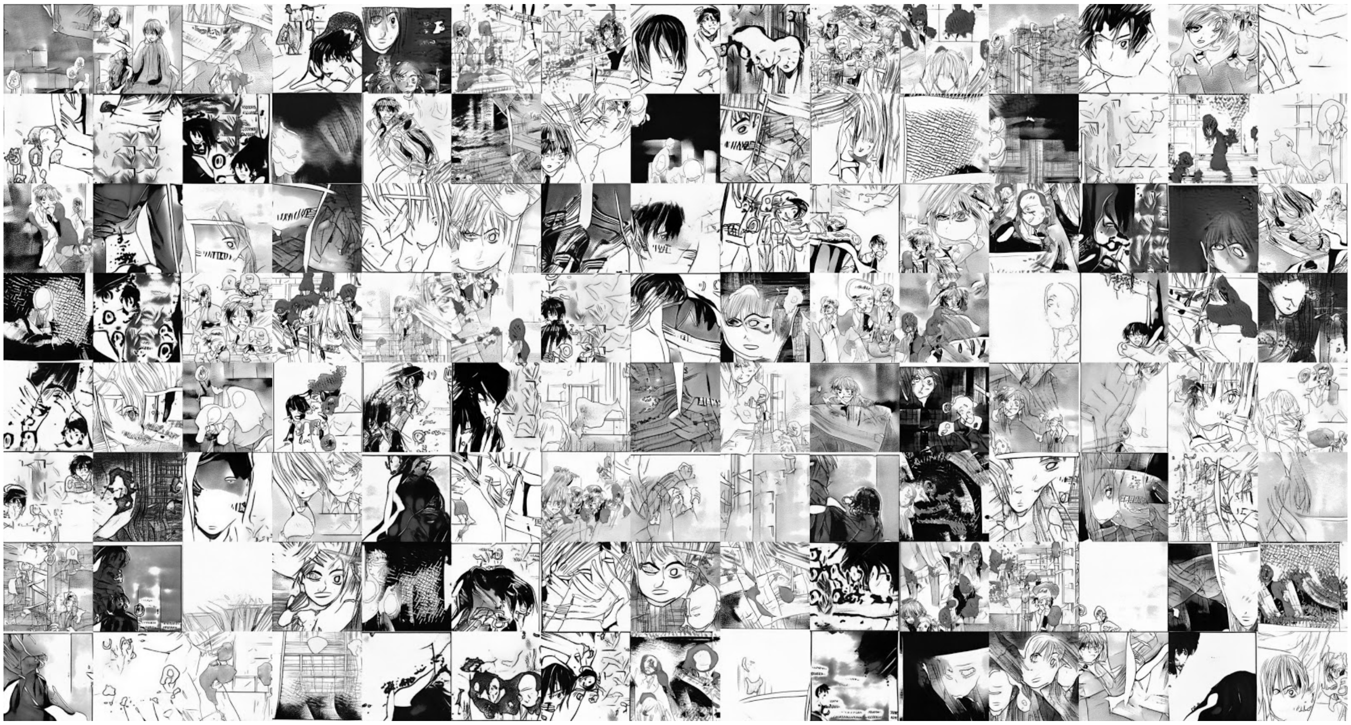

Figure 2: Progress of training on labeled 1x1 images @ kimg=1913

we train language models on a wide variety of publicly available monolingual corpora of literature. For this purpose, we compile separate English corpora from various works of Poetry, True-Crime, Science Fiction and Buddhist literature. Models which are fine-tuned on each of those different language models will be used later as distinct "voices" (Poetry will play the role of narration), to give the impression of a dialogue which unfolds between distinct characters, focusing on different subjects and emotions. To make the output more appealing and to exclude writing styles unrelated to graphic novels, we only keep sampled sequences that follow a set of simple semantic rules (for example ignoring chapter names or references), without a great loss of continuity, as texts of similar styles are generated in sequence. We observe that these models produce more rich and engaging samples than the one which was trained on the text extracted from our Manga dataset.

Assembling a Graphic Novel

As the final part of the described pipeline, we investigate the assembly of a comic from two independent and locally coherent streams of image and text content. We experiment with two processes to generate panel layouts for a fixed page size of `2480x3508px`: (a) a randomized approach of generating panels in multiple scales and rectangle shapes and (b) a standard 5 row by 4 column layout of square images.

We first develop a randomized layout synthesis, that follows several fixed constraints. We design a recursive function, which starts from a higher level and based on a biased coin-flip, decides whether it should split in half the current panel space, either horizontally or vertically. Each time, the level is decreased by one, starting from 4 and it stops when it reaches the number zero (see Figure 4). Additionally, while splitting a level in half, it also decides whether the split will be parallel to the borders of the page, or whether it should include a certain amount of randomly selected small inclination. Moreover, we randomly decide if the borders of each panel will have a gutter (the value of which is again selected randomly) or if they will be a so called *full-bleed*. All those distributions are parametrized at panel level and their detailed parameters are estimated through experimentation to fit certain aesthetic criteria. Figure 5 contains a few examples of different layouts that are synthesized using this procedure. The non-randomized approach (b) is not discussed because of its simplicity. It is mainly introduced, to balance out the fuzzy elements of the synthetic text and image content and to allow the reader to effectively engage with the current version of this work.

For each of the above procedures we have used different ways for selecting synthetic content. In the case of (a) we follow again a manually tuned random sampling procedure for each level, were we select images based on their categories. Moreover, to avoid having images with a lot of variation in small levels we use a very simple measure of Shannon-Entropy. To fit an image inside each panel, we resize each image to the smallest dimension of the panel, while keeping its aspect ratio constant and we crop it with equal margins from the borders of its largest dimension. For the more coherent version of the squared layout we ignore the categorical labels and use 4 local images which come from

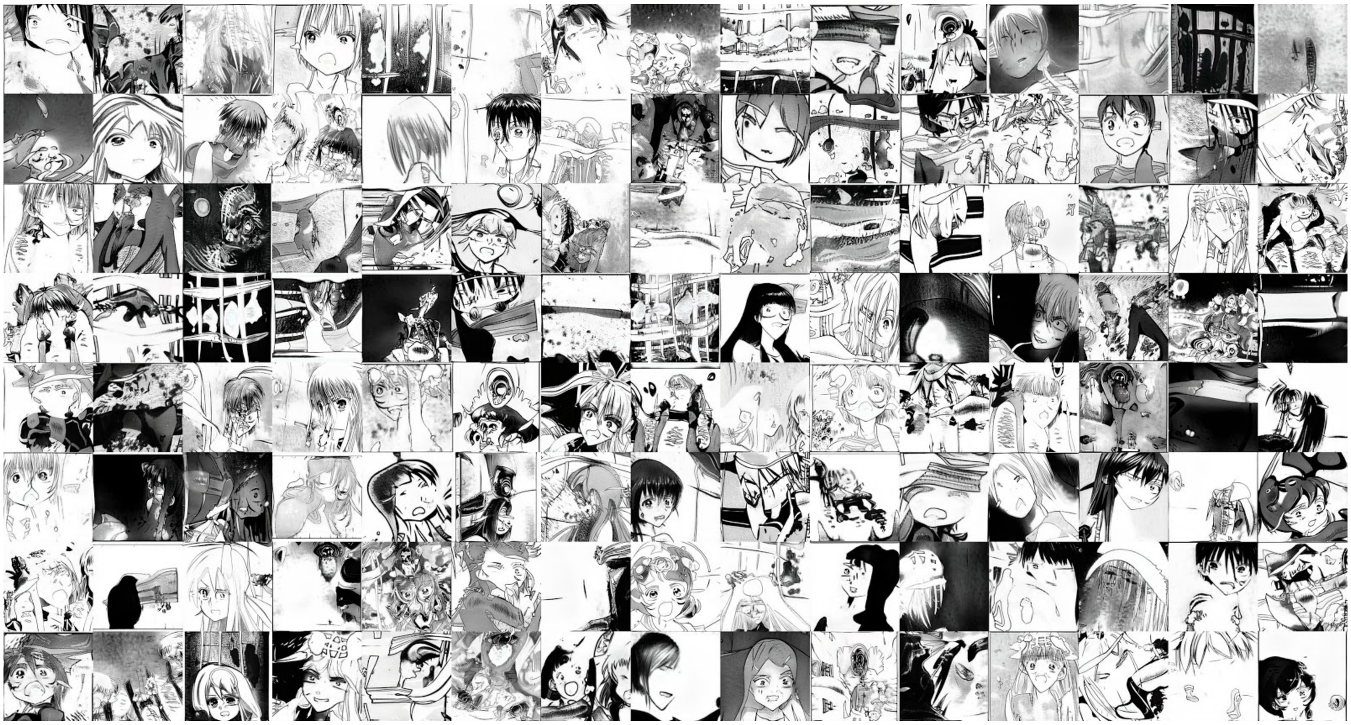

Figure 3: Progress of transfer learning from a model pre-trained on Manga female faces @ kimg=1490.

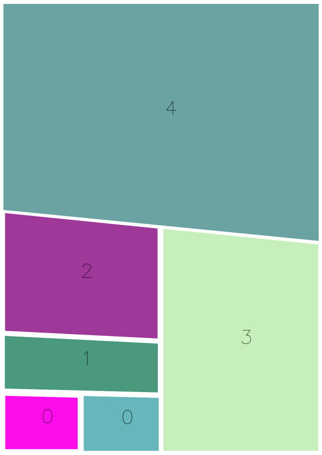

Figure 4: Panels and their associated levels.

5 sequences per page in the given order. This plays the role of action or character unfolding.

As a next step we allocate a page and a panel with a custom amount of bubbles for each level, which we will then fill with text, both in the form of dialogue and of narration. For the purpose of dialogue we use three different types of text bubbles, which would represent the three different voices of our previously fine-tuned language models and a (final) fourth one which plays the role of narration. In the case (a) of randomized panels, we allow at most one voice-over per panel and up to four bubbles, the amount of which is randomly selected from a range of values related to their level. In the case of (b), we evenly and randomly distribute the narration text boxes and the bubbles of a given number inside the page. Also, in both cases, we bound the total amount of generated text per level, also taking into account its function - as narration or as dialogue. We implement a recursive function to best fit our text inside a given rectangle area and then re-scale our bubbles, so that their inscribed rectangle would properly fit our text. To locate the area, inside which we would place those bubbles, we construct a Boolean mask which we incrementally update, starting by adding the narration and afterwards by placing each dialogue bubble, while adding a constraint to the total area which can be covered by a bubble inside the panel or inside the page (in cases (a) and (b) respectively).

In terms of content, we use a poetry language model for the purpose of narration, as it provides a sense of poetic continuity throughout the pages. Moreover, for the dialogue bubbles we use three different types of language models which we associate with different qualities and emotions. First we have a Buddhist model, that with a more spiritual, zen and wise vibe. Then we have a Sci-Fi model that is futuristic, exhibiting a certain amount of complexity in situations and techniques. Finally, a True-Crime literature model that creates a sense of action and suspense is included. To conclude, a visual example of method (a) can be seen on Figure 6 and one of method (b) on Figure 7, which was the one that made it to the ”final cut”.

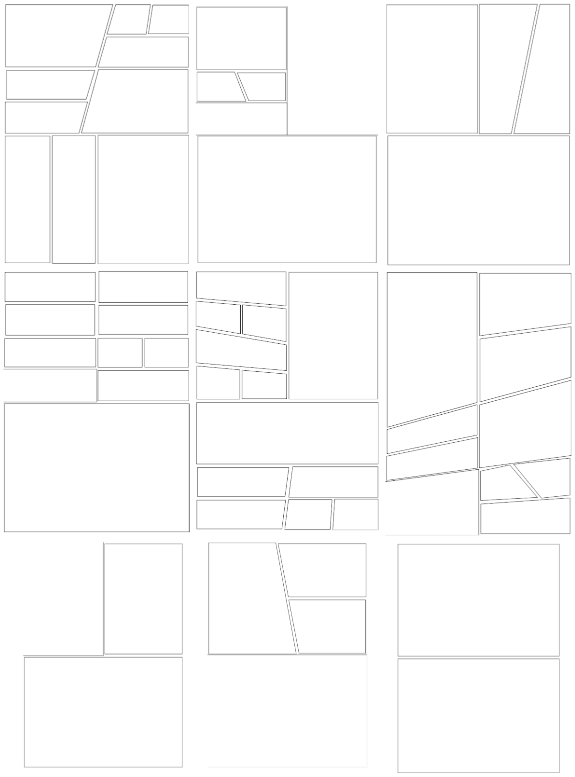

Figure 5: Examples of randomly sampled panel layouts.

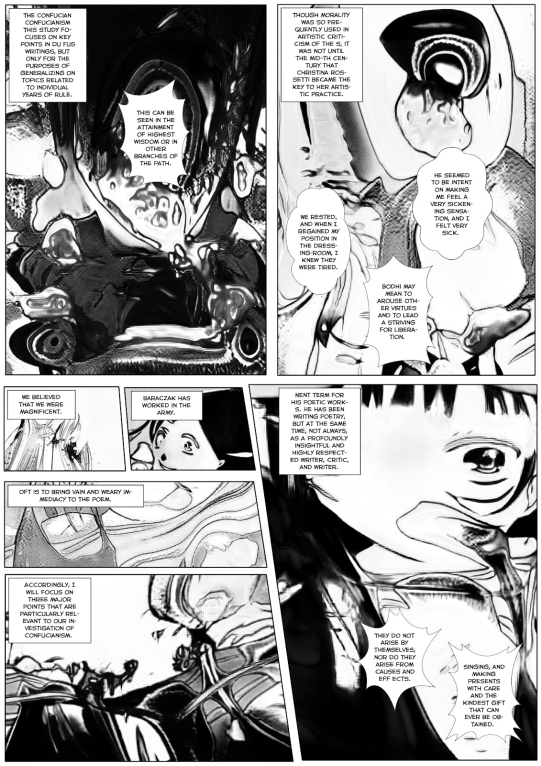
THE CONFUCIAN CONFUCIANISM THIS STUDY FOCUSES ON KEY POINTS IN DU FUS WRITINGS, BUT ONLY FOR THE PURPOSES OF GENERALIZING ON TOPICS RELATED TO INDIVIDUAL YEARS OF RULE.
THIS CAN BE SEEN IN THE ATTAINMENT OF HIGHEST WISDOM OR IN OTHER BRANCHES OF THE PATH.
THOUGH MORALITY WAS SO FREQUENTLY USED IN ARTISTIC CRITICISM OF THE S, IT WAS NOT UNTIL THE MID-TH CENTURY THAT CHRISTINA ROSSETTI BECAME THE KEY TO HER ARTISTIC PRACTICE.
HE SEEMED TO BE INTENT ON MAKING ME FEEL A VERY SICKENING SENSATION, AND I FELT VERY SICK.
WE RESTED, AND WHEN I REGAINED MY POSITION IN THE DRESSING-ROOM, I KNEW THEY WERE TIRED.
BODHI MAY MEAN TO AROUSE OTHER VIRTUES AND TO LEAD A STRIVING FOR LIBERATION.
WE BELIEVED THAT WE WERE MAGNIFICENT.
BARACZAK HAS WORKED IN THE ARMY.
NENT TERM FOR HIS POETIC WORKS. HE HAS BEEN WRITING POETRY, BUT AT THE SAME TIME, NOT ALWAYS, AS A PROFOUNDLY INSIGHTFUL AND HIGHLY RESPECTED WRITER, CRITIC, AND WRITER.
OFT IS TO BRING VAIN AND WEARY IMMEDIACY TO THE POEM.
ACCORDINGLY, I WILL FOCUS ON THREE MAJOR POINTS THAT ARE PARTICULARLY RELEVANT TO OUR INVESTIGATION OF CONFUCIANISM.
THEY DO NOT ARISE BY THEMSELVES, NOR DO THEY ARISE FROM CAUSES AND EFF ECTS.
SINGING, AND MAKING PRESENTS WITH CARE AND THE KINDEST GIFT THAT CAN EVER BE OBTAINED.

Figure 6: A page synthesized with the randomized assembly procedure.

Future work

In this work, we manage to construct a pipeline which synthesizes pages that resemble the style of Manga comics. The generation of image and text content, however, has been done independently and their assembly has been aesthetically pleasant and engaging but not necessarily indicative of our dataset. Associating images with text in a meaningful way and arranging them in a semantically consistent order is a critical next step, that will allow us to research creative narration in the multi-modal setting. Relevant tasks have been well researched, such as image captioning (Herdade et al. 2020) or image generation conditioned on text descriptions (Ramesh et al. 2021). However, it can be argued that the text and image interrelation, in the setting of comics, is not as straightforward and strict as that of an image and its caption.

An architecture that emerged recently and stirred up interest regarding the connection of visual representations to natural language is CLIP (Radford et al. 2021). To sum up its basic functionality, CLIP learns joint text and image embeddings, that allow the computation of a (cosine similarity based) matching score between images and text. We believe that analogous approaches should definitely be explored for the task of end-to-end comics synthesis. Ideally, a graphic novel generator is a multi-modal system that learns from and generates parallel streams of image and text, structured accordingly. As we previously mentioned, the function of images and their relation to text in graphic novels is unique and unexplored. Image and text are equally important parts of the same narrative, with no strict hierarchical relationship between them being implied. Therefore, simply generating one modality conditioned on the other is an approach limiting to the comics medium. For example, images can serve the purpose of capturing the reader's attention for the text to find its meaning, while in other cases text can simply include a non-important dialogue as what is happening in the image is more important for the story.

Thus, in order to fully model a graphic novel, one could model narrative as a latent and discrete time series underlying both the image and text data, from which new images and text can emerge. This work does not claim to have tackled this problem, but rather it is the first applied and complete study towards this direction, using contemporary tools and suggesting their limitations.

Discussion

The comics industry has been quite reticent in embracing the complex nature of technological developments in artificial intelligence. But this situation might soon change. The online abundance of digitized media content, available through third-party groups of comics fans, the increasing convenience of programming language frameworks and machine learning libraries, the secularization of knowledge through e-learning and the plummeting prices in specialized hardware might contribute to reach a critical point where artifi-

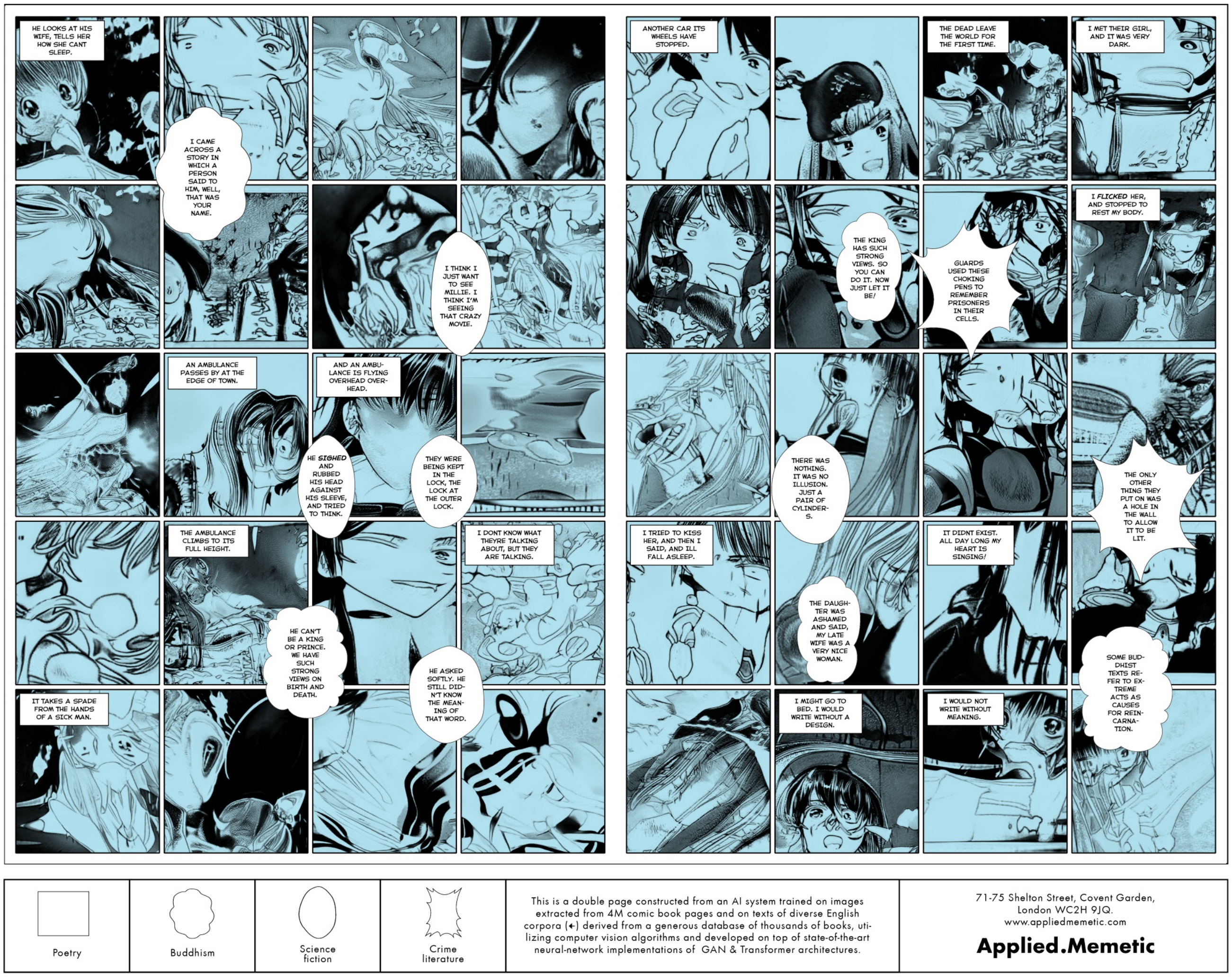
HE LOOKS AT HIS WIFE, TELLS HER HOW SHE CANT SLEEP.
I CAME ACROSS A STORY IN WHICH A PERSON SAID TO HIM, WELL, THAT WAS YOUR NAME.
I THINK I JUST WANT TO SEE MILLIE. I THINK I'M SEEING THAT CRAZY MOVIE.
AN AMBULANCE PASSES BY AT THE EDGE OF TOWN.
AND AN AMBULANCE IS FLYING OVERHEAD OVERHEAD.
HE *SIGHED* AND RUBBED HIS HEAD AGAINST HIS SLEEVE, AND TRIED TO THINK.
THEY WERE BEING KEPT IN THE LOCK, THE LOCK AT THE OUTER LOCK.
THE AMBULANCE CLIMBS TO ITS FULL HEIGHT.
I DONT KNOW WHAT THEYRE TALKING ABOUT, BUT THEY ARE TALKING.
HE CAN'T BE A KING OR PRINCE. WE HAVE SUCH STRONG VIEWS ON BIRTH AND DEATH.
HE ASKED SOFTLY. HE STILL DIDN'T KNOW THE MEANING OF THAT WORD.
IT TAKES A SPADE FROM THE HANDS OF A SICK MAN.
ANOTHER CAR ITS WHEELS HAVE STOPPED.
THE DEAD LEAVE THE WORLD FOR THE FIRST TIME.
I MET THEIR GIRL, AND IT WAS VERY DARK.
THE KING HAS SUCH STRONG VIEWS. SO YOU CAN DO IT. NOW JUST LET IT BE!
GUARDS USED THESE CHOKING PENS TO REMEMBER PRISONERS IN THEIR CELLS.
I *FLICKED* HER, AND STOPPED TO REST MY BODY.
THERE WAS NOTHING. IT WAS NO ILLUSION. JUST A PAIR OF CYLINDER-S.
THE ONLY OTHER THING THEY PUT ON WAS A HOLE IN THE WALL TO ALLOW IT TO BE LIT.
I TRIED TO KISS HER, AND THEN I SAID, AND ILL FALL ASLEEP.
IT DIDNT EXIST. ALL DAY LONG MY HEART IS SINGING!
THE DAUGHTER WAS ASHAMED AND SAID, MY LATE WIFE WAS A VERY NICE WOMAN.
SOME BUDDHIST TEXTS REFER TO EXTREME ACTS AS CAUSES FOR REINCARNATION.
I MIGHT GO TO BED. I WOULD WRITE WITHOUT A DESIGN.
I WOULD NOT WRITE WITHOUT MEANING.
Poetry
Buddhism
Science fiction
Crime literature
This is a double page constructed from an AI system trained on images extracted from 4M comic book pages and on texts of diverse English corpora (←) derived from a generous database of thousands of books, utilizing computer vision algorithms and developed on top of state-of-the-art neural-network implementations of GAN & Transformer architectures.
71-75 Shelton Street, Covent Garden, London WC2H 9JQ. www.appliedmemetic.com
Applied.Memetic

Figure 7: A random two-pager generated from the website of this paper. It comes in a standard 5 by 4 square layout.

cial intelligence will be gradually integrated in the comics pipeline. Synthetic and generative processes might soon reshape the ways we produce, consume, archive and distribute comics artifacts. A more wide adoption of artificial intelligence in different strata of the industry might reconfigure existing readership(s) market(s).

Our research aims to explore the conditions for synthesizing graphic narratives and comics with the use of deep neural networks. This may result to a better understanding of creativity for comics artists and cartoonists but might also contribute to multiple applications of multi-modal expressive communication that has become our primary modality in sharing and shaping representation of our worlds. Modeling a graphic narrative still remains a challenging task, which has generally been ignored by the deep learning and computer science community. Researching and understanding this multi-modal, discrete and symbolic procedure involved in the production of comics, provides a very challenging task which can unite comics artists and deep learning engineers and potentially augment human creativity in ways which have never been experienced before. Finally, we suggest that reverse-engineering or modeling parts of this procedure can even provide us with mathematical and technological tools, which could profit other fields unrelated to artistic practices.

Acknowledgments

Ilan Manouach is the founder of Applied Memetic, a synthetic media company committed to explore the impact of deep learning in the comics industry. This research and the development of its early AI models has been materialized using the company's resources.

References

Anonymous; community, D.; and Branwen, G. 2021. Danbooru2020: A large-scale crowdsourced and tagged anime illustration dataset. https://www.gwern.net/Danbooru2020.

Aydao. 2021. This anime does not exist. https://aydao.ai/work/2021/01/18/stylegan2ext.html.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners.

Cao, Y.; Chan, A. B.; and Lau, R. W. H. 2012. Automatic stylistic manga layout. *ACM Trans. Graph.* 31(6).

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; Zhang, Z.; Cheng, D.; Zhu, C.; Cheng, T.; Zhao, Q.; Li, B.; Lu, X.; Zhu, R.; Wu, Y.; Dai, J.; Wang, J.; Shi, J.; Ouyang, W.; Loy, C. C.; and Lin, D. 2019. Mmdetection: Open mmlab detection toolbox and benchmark.

Dubray, D., and Laubrock, J. 2019. Deep cnn-based speech balloon detection and segmentation for comic books.

Fujimoto, A.; Ogawa, T.; Yamamoto, K.; Matsui, Y.; Yamasaki, T.; and Aizawa, K. 2016. Manga109 dataset and creation of metadata. In *Proceedings of the 1st International Workshop on CoMics ANalysis, Processing and Understanding*, MANPU '16. New York, NY, USA: Association for Computing Machinery.

Furusawa, C.; Hiroshiba, K.; Ogaki, K.; and Odagiri, Y. 2017. Comicolorization: Semi-automatic manga colorization. In *SIGGRAPH Asia 2017 Technical Briefs*, SA '17. New York, NY, USA: Association for Computing Machinery.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial networks.

Groensteen, T.; (Firm), P.; Nguyen, N.; and of Mississippi, U. P. 2007. *The System of Comics*. University Press of Mississippi.

Guérin, C.; Rigaud, C.; Mercier, A.; Ammar-Boudjelal, F.; Bertet, K.; Bouju, A.; Burie, J.-C.; Louis, G.; Ogier, J.-M.; and Revel, A. 2013. ebdtheque: A representative database of comics. 1145–1149.

Gwern. 2019a. Making anime faces with stylegan. https://www.gwern.net/Faces.

Gwern. 2019b. This waifu does not exist. https://www.gwern.net/TWDNE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition.

Herdade, S.; Kappeler, A.; Boakye, K.; and Soares, J. 2020. Image captioning: Transforming objects into words.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2018. Gans trained by a two time-scale update rule converge to a local nash equilibrium.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network.

Ho, A. K. N.; Burie, J.-C.; and Ogier, J.-M. 2012. Panel and speech balloon extraction from comic books. In *Proceedings of the 2012 10th IAPR International Workshop on Document Analysis Systems*, DAS '12, 424–428. USA: IEEE Computer Society.

Howard, J., and Gugger, S. 2020. Fastai: A layered api for deep learning. *Information* 11(2):108.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks.

Li, L.; Wang, Y.; Tang, Z.; and Gao, L. 2014. Automatic comic page segmentation based on polygon detection. *Multimedia Tools Appl.* 69(1):171–197.

Matsui, Y.; Ito, K.; Aramaki, Y.; Fujimoto, A.; Ogawa, T.; Yamasaki, T.; and Aizawa, K. 2016. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications* 76(20):21811–21838.

Nagadomi. 2018. waifu2x: Image super-resolution for anime-style art using deep convolutional neural networks. https://github.com/nagadomi/waifu2x.

Nguyen, N.-V.; Rigaud, C.; and Burie, J.-C. 2018. Digital comics image indexing based on deep learning. *Journal of Imaging* 4:89.

Oeldorf, C., and Spanakis, G. 2019. Loganv2: Conditional style-based logo generation with generative adversarial networks.

Ogawa, T.; Otsubo, A.; Narita, R.; Matsui, Y.; Yamasaki, T.; and Aizawa, K. 2018. Object detection for comics using manga109 annotations.

Pham, D. T.; Dimov, S. S.; and Nguyen, C. D. 2004. An incremental k-means algorithm. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 218(7):783–795.

Qin, X.; Zhou, Y.; He, Z.; Wang, Y.; and Tang, Z. 2017. A faster r-cnn based method for comic characters face detection. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, 1074–1080. IEEE.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation.

Rayar, F.; Oriola, B.; and Jouffrais, C. 2020. Alcove: An accessible comic reader for people with low vision. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, 410–418. New York, NY, USA: Association for Computing Machinery.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation.

Ross, D.; Lim, J.; Lin, R.-S.; and Yang, M.-H. 2008. Incremental learning for robust visual tracking. *International Journal of Computer Vision* 77:125–141.

Smith, R. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, 629–633.

Su, H.; Niu, J.; Liu, X.; Li, Q.; Cui, J.; and Wan, J. 2020. Mangagan: Unpaired photo-to-manga translation based on the methodology of manga drawing.

Sun, W., and Kise, K. 2013. Detection of exact and similar partial copies for copyright protection of manga. *Int. J. Doc. Anal. Recognit.* 16(4):331–349.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Yamada, M.; Budiarto, R.; Endo, M.; and Miyazaki, S. 2004. Comic image decomposition for reading comics on cellular phones. *IEICE Transactions* 87-D:1370–1376.

Zhou, Y.; Wang, Y.; He, Z.; Tang, Z.; and Suen, C. Y. 2020. Towards accurate panel detection in manga: A combined effort of cnn and heuristics. In *International Conference on Multimedia Modeling*, 215–226. Springer.