OCR-VQGAN: Taming Text-within-Image Generation

Juan A. Rodriguez[1], David Vazquez[2], Issam Laradji[2], Marco Pedersoli[3], Pau Rodriguez[2]
[1]Computer Vision Center, Barcelona, [2]ServiceNow Research, [3]ÉTS Montréal
joanrg.ai@gmail.com

Abstract

Synthetic image generation has recently experienced significant improvements in domains such as natural image or art generation. However, the problem of figure and diagram generation remains unexplored. A challenging aspect of generating figures and diagrams is effectively rendering readable texts within the images. To alleviate this problem, we present OCR-VQGAN, an image encoder, and decoder that leverages OCR pre-trained features to optimize a text perceptual loss, encouraging the architecture to preserve high-fidelity text and diagram structure. To explore our approach, we introduce the Paper2Fig100k dataset, with over 100k images of figures and texts from research papers. The figures show architecture diagrams and methodologies of articles available at arXiv.org from fields like artificial intelligence and computer vision. Figures usually include text and discrete objects, e.g., boxes in a diagram, with lines and arrows that connect them. We demonstrate the effectiveness of OCR-VQGAN by conducting several experiments on the task of figure reconstruction. Additionally, we explore the qualitative and quantitative impact of weighting different perceptual metrics in the overall loss function. We release code, models, and dataset at https://github.com/joanrod/ocr-vqgan.

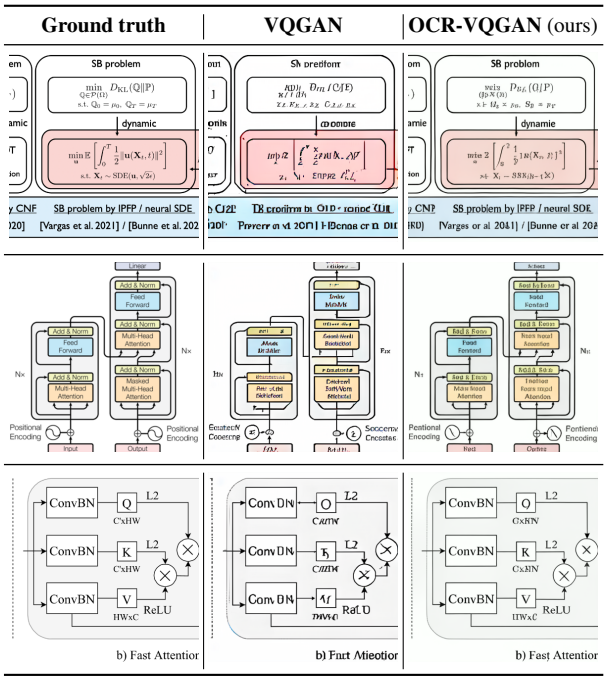


Table 1. Qualitative comparison for the task of figure reconstruction. OCR-VQGAN outperforms VQGAN at capturing text and symbol details.

1. Introduction

Image synthesis efforts in the recent literature have achieved impressive results in the domain of natural images. Some examples are face generation, landscapes, and art [1, 2, 3, 4, 5, 6]. Current methods can generate high-resolution and realistic images, allowing creators to guide the generation process using text descriptions and other conditioning modalities [7]. Within the next few years, text-to-image generation models are set to enhance and complement creative processes in many fields including art, design, video games, and content creation.

However, a common deficiency of current methods like Parti [4] or Imagen [3] is that they tend to fail at text rendering within images, as highlighted in [4] (limitations section). Current image generation systems trained on natural images do not produce the desired results for applications like structured diagram generation or automatic slide edition. This shortcoming makes recent text-to-image generation methods unsuitable to aid people to design and create figures and diagrams for their work. Thus, in this work, we tackle the problem of generating text within images. We focus on the generation of paper figures, which typically contain a visually summarized description of methods or architectures presented in academic papers. The use of figures in research publications has become widespread in many fields (we focus on Deep Learning (DL) and Computer Vision (CV)), allowing researchers to describe ideas compactly. People spend a significant amount of time building appealing and understandable figures. Therefore, tools

for assisting this process will be beneficial.

Although figures present a more precise structure and composition than natural images (i.e., connectivity between objects, textual descriptions, legends), Roy et al. [8] highlight that there is little agreement on how these figures should be created, as they do not follow a specific set of rules for formatting or structuring information (see Figure 1). Writers must define many parameters when generating figures, such as size, the position of shapes, colors, connections, or text styles. Hence, building image generation models in the space of figures is rather time-consuming, as there is significant variability. In this work, we tackle the problem of generating figures by (i) introducing a novel dataset of figure images and text pairs from research papers and (ii) addressing the text rendering problem of current image encoders.

We acquire a large database of figure images and texts from publicly available research papers on arXiv.org, which we call Paper2Fig100k. The dataset contains over 100k pairs of images and text captions (see Section 3), designed for the task of text-conditional generation of figures, which are important components in many real-life applications such as architectural blueprints, project designs, and research papers. The proposed dataset allows tackling the problem of text rendering in diagrams, which was not possible in previous scene-text datasets.

We construct a rich image encoder for figures. We build upon the VQGAN [9] approach, based on learning a codebook of patch embeddings, exploiting perceptual and patch-based adversarial objectives. Our method adds an OCR perceptual loss term to the training objective that minimizes the distance between original and reconstructed images in the feature space. To this end, we add a pre-trained OCR detection model, encouraging the learned embeddings to represent clear text and diagrams (see Table 1).

Since we focus on compressing images in the domain of figures, we need to impose the model to learn the most frequent patches and local patterns present in the dataset, building a rich discrete latent representation (i.e., a codebook of patches). For instance, our encoder needs to represent a variety of text possibilities, such as different sizes, styles, fonts, and the orientation of characters. Text and background colors also need to be considered, as well as the sharpness of arrows, lines, or geometric shapes.

Traditional image reconstruction metrics heavily rely on L2 similarities in the pixel space, which tend to fail at quantifying high-level perceptual similarity. For example, a simple shift of the image in the horizontal axis makes a pointwise L2 distance give incorrect results. In contrast, humans can detect patterns and the overall structure of images, allowing them to make a better perceptual assessment. This can be achieved by minimizing a perceptual loss in the feature space, using models pre-trained for image recognition.

Our contributions can be summarized as follows. We propose:

- Paper2Fig100k, a novel dataset for the task of text-to-image generation, composed of texts and images acquired from publicly available research papers;
- OCR-VQGAN, an image encoder focused on synthesizing images of figures, preserving text-within-images and diagram structure; and
- an OCR perceptual loss and OCR similarity metric (OCR-SIM), devoted to measuring the perceptual distance of images with respect to an OCR pre-trained model.

The rest of the paper is structured as follows. Section 2 gives an overview of the recent literature on deep generative models, image encoders, and diagram-based tasks and datasets. Section 3 describes Paper2Fig100k, a novel dataset of research figures and texts. In Section 4 we propose OCR-VQGAN, an image encoder focused in preserving textual clarity within images. Experiments are conducted in Section 5. In Section 6 conclusions are drawn. Finally, in the final section, we reflect on the ethical and social impact of text-to-image generation.

2. Related Work

Text-to-image synthesis. Recent text-to-image generation methods have achieved outstanding image generation quality even for unseen concept compositions. Works like DALLE [10], DALLE-mini [11], or Parti [4] pose the task as a language modeling problem, using Transformer [12] architectures to learn the relationships between text and visual information. This family of methods relies on image encoders (or tokenizers) such as VQ-VAE [13] or VQGAN [9], that convert images into a sequence of tokens, allowing to treat both text and image modalities as a sequence-to-sequence task. Another family of methods that have gained popularity use diffusion methods [14] to directly generate images from text embeddings. Works like DALLE2 [2] and Imagen [3], use CLIP [15] and [16] text encoders to condition the diffusion process. Latent Diffusion [6] incorporates a BERT tokenizer and a more flexible cross-attention mechanism to condition the diffusion process. A weakness of current text-to-image synthesis methods is that they struggle to generate text [3, 4]. In this work, we improve the image encoder module to support text generation within images.

Image tokenizers. Vector Quantized Variational Autoencoder (VQVAE) [13, 17] is a popular approach for learning discrete representations of images. VQ-based methods learn a codebook of discrete latent embeddings and use a

nearest neighbor algorithm to map continuous latent features to discrete embeddings. They propose to model the data distribution by means of autoregressive density estimation, using causal convolutional kernels. VQGAN uses the quantization procedure of VQVAE and improves the richness of VQVAE's learned codebook. The authors of VQGAN modify the training objective using a VGG perceptual loss [18] and a patch-based adversarial module to obtain high-quality embeddings. They demonstrate how learning a rich codebook of image patches is crucial in order to perform high-resolution image synthesis. Although VQGAN improves the reconstruction quality of VQVAE, the model still struggles to draw text (see Table 6. In this work, we include an additional perceptual loss that improves text reconstruction.

Perceptual-based reconstruction. Perceptual similarity losses are common in the field of Style Transfer [19, 20, 21]. Zhang et al. [18] prove the effectiveness of using learned perceptual losses as a similarity objective, which is evaluated in the feature space. Their work shows how VGG16 [22] pre-trained on Imagenet [23] can be used as reconstruction loss, capturing differences in perceptual similarity. Perceptual losses pre-trained on Imagenet are adequate for measuring differences between natural images. However, they are inadequate to measure distances in text generation, since text recognition is not an objective of the Imagenet. Here we alleviate this problem by introducing an additional perceptual loss obtained from a model trained for text detection Optical Character Recognition (OCR) [24].

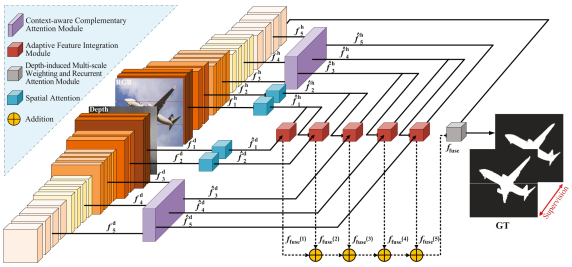
Related datasets and tasks. Some works have been proposed in the domain of document and figure analysis, mainly focused on classification [25] and object detection [8] tasks or visual question answering [26]. Most of the available datasets contain many types of figures such as tables, flow charts, and different types of plots. Hsu et al. [27] introduces a dataset of figures and texts for the task of image captioning, which contains $60,000$ samples of all available figures in scientific papers (e.g., scatter plots, bar plots, flowcharts, equations). Chen et al. [28] also approaches figure captioning by introducing the FigCAP dataset, containing samples from many types of figures. To the best of our knowledge, there are no publicly available datasets focused on diagram figures. We propose the construction of a new dataset of figure diagrams and texts from research articles.

3. Paper2Fig100k dataset

In order to accomplish text-to-figure generation, and to address the lack of publicly available datasets for the task, we present the Paper2Fig100k dataset. It consists on $102,453$ pairs of images and texts from $69,413$ papers. The data is split into a training set of $81,194$ samples and a test set of $21,259$ samples. While the proposed dataset is meant for text-to-figure generation, it can also be used to train the first stage of a text-to-image pipeline (image encoder). Paper2Fig100k can also be used for image-to-text generation (reverse process) and multi-modal vision-language tasks. Samples from the dataset are shown in Figure 1.

Paper2Fig100k contains images of architectures, diagrams, and pipelines (generally referred to as figures), with detailed text captions acquired from public research papers at arXiv.org. It also includes OCR-detected bounding boxes and text transcriptions of figures, that can be used for handcrafted attention and fine-grained text conditioning. As shown in Figure 2, the dataset is expected to increase exponentially over the years.

Input image



Prompt Modality	Examples
Caption	*Figure 2: An overview of our network. We propose a Complementary Attention and Adaptive Integration Network*
References	*Fig. 2 shows an overview of CAAI-Net, which is based on a two-stream structure for RGB images and depth maps. As can be observed, (...).*
OCR keywords	*Context-aware Complementary Attention Module, (...).*

Table 2. Example of our captioning system for the task of text-conditional image synthesis. **Note** that, the figure shown at the top of the table is a sample from the dataset [2].

3.1. Data acquisition pipeline

The dataset was acquired using the API and metadata offered by *arXiv dataset* [29], which includes both paper metadata (e.g., title, abstract, authors, or research fields) and tools for downloading the papers in pdf format via Google

[1]The example figure was extracted from the paper *Towards Accurate RGB-D Saliency Detection with Complementary Attention and Adaptive Integration*

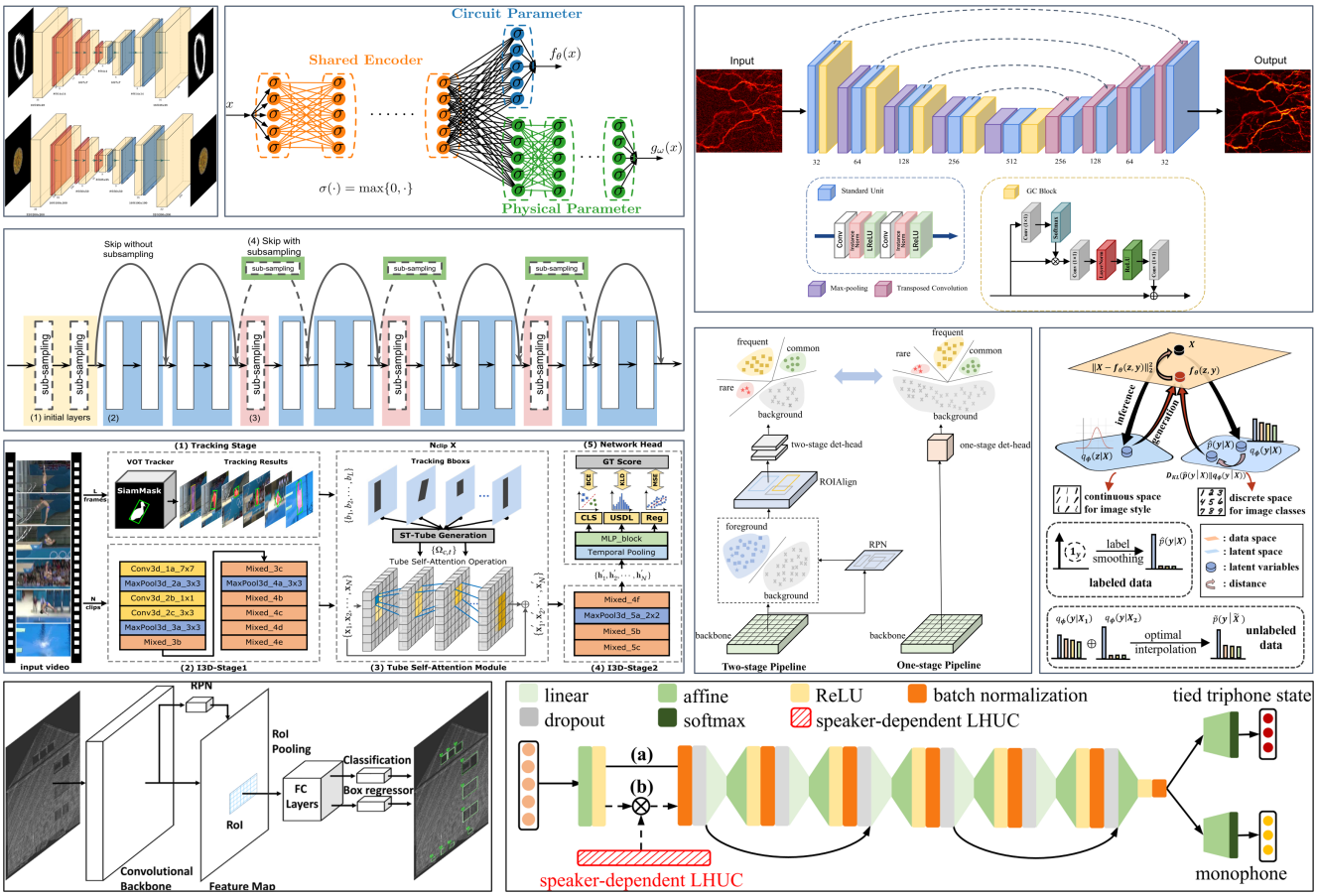


Figure 1. Samples from Paper2Fig100k dataset. We show how the samples have high variability in aspect ratio, image resolution, text and diagram sizes, or amount of information displayed. Unlike natural images, figures contain fine-grained information details that are relevant for a complete understanding of the diagram.

Cloud. *arXiv dataset* is updated weekly, offering more than 1.7 million papers across all STEM fields.

The complete set of papers available covers a vast amount of fields in the arXiv taxonomy[2], therefore we filter and keep papers in the categories of Machine Learning (cs.LG), Artificial Intelligence (cs.AI), Computer Vision and Pattern Recognition (cs.CV), and Computation and Language (cs.CL). We downloaded all papers published after January 2010, which represent a total of $183,427$ papers.

Papers in pdf format are processed and parsed using the GROBID [30] open-source library, which allows to extract and organize texts and images of pdf files, mostly focused on technical or scientific documents. It is based on a cascade of models for object detection such as conditional random fields (CRF). The software is production ready and very competitive in terms of processing speed (it is capable of processing 10.6 pdf per second). We release the pipeline at https://github.com/joanrod/paper2figure-dataset.

3.2. Heuristics for obtaining figures

The $183,427$ downloaded papers contain $\sim 1.6M$ images. However, many of them contain qualitative results or other kinds of natural images that we want to avoid. Since we are interested in figure diagrams, we apply simple heuristics to keep only figures describing architectures or methods, and remove figures related to results or examples. The result is a set of 102,453 images.

We use text-based heuristics using figure captions. We keep figures containing strings in the caption such as "architecture", "model diagram" or "pipeline". We remove figures with words like "table", "results" or "example", which may not correspond with the desired figures. Gray-scale histograms are used to remove outliers such as blank (all white) or natural images (almost no white).

We process images with an OCR detection and recognition system (EasyOCR), based on the CRAFT [24] OCR

[2]arXiv follows a standardized taxonomy to encode field categories, https://arxiv.org/category_taxonomy

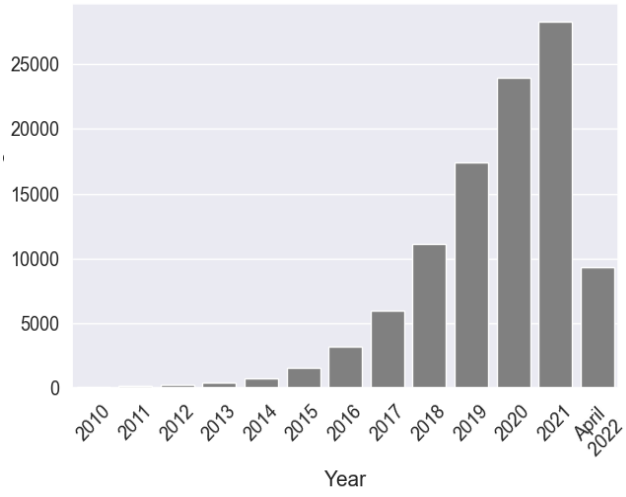


Figure 2. In this plot we show the number of figures extracted per year, ending in April 2022. This figure shows the exponential increase in research publications.

detector and the CRNN [31] text recognizer. Note that this process is applied once and it is independent of the OCR-VQGAN method later introduced in this work. The goal is to obtain textual tags and captions to automatically annotate samples and use them for text conditioning.

3.3. Prompt modalities

Text captions generally encode the same information that is represented in the figure, hence text-conditional image synthesis can be achieved. There are many options for conditioning figure generation based on the text of a paper. As figures usually describe the methods presented in their papers, we propose pairing figure images with the text in *methodology* sections. We also explore other conditioning information such as figure captions and keywords extracted from an OCR model.

Concretely, we define three prompting modalities for future research in text-to-image generation: caption, references, and OCR keywords. Captions are obtained directly from the figure caption. References are extracted from paragraphs in the paper referencing the figure. OCR keywords are a concatenation of the texts detected by the OCR model. Table 2 exemplifies the kind of captions in the dataset.

4. Method

Our goal is to train a figure-based image encoder (tokenizer) capable of transforming images into sequences of discrete tokens, and an image decoder (detokenizer) that reconstructs figures from tokens, preserving details of text and diagram structure. Tokens are indices of a learned codebook of patch embeddings in the discrete latent space, that encodes the patches in the original image. When encoding an image, individual patch embeddings are assigned to the nearest codebook entry. To solve the reconstruction task at hand, the method needs to learn the most relevant and realistic patches within our dataset, and assign discrete embeddings in the codebook. To this end, the proposed OCR-VQGAN encoding and decoding pipeline uses a patch-based adversarial procedure, a VGG-based perceptual (LPIPS) loss, and a novel OCR perceptual loss.

4.1. OCR-VQGAN

We leverage VQGAN's image encoder [9] to learn a mapping from the image space to a discrete latent representation of tokens. To this end, the VQGAN architecture is composed of an image encoder, a decoder, and a vector quantization stage. The encoder is devoted to downsample an image $x \in \mathbb{R}^{H\times W\times 3}$ into discrete codes $z_q \in \mathbb{R}^{h\times w\times n_z}$, where n_z is the size of the embedding space. One can simply describe each code with its codebook index and rearrange the discrete representation by a grid of shape $h \times w$.

Using the same architecture and notation as Esser et al. [9], the image encoder E and decoder G are convolutional neural networks aimed at learning a discrete codebook $\mathcal{Z} = \{z_k\}_{k=1}^{K} \subset \mathbb{R}^{n_z}$. In a forward pass, an image x is approximated as $\hat{x} = G(q(E(x)))$, where q is the quantization function, which performs a nearest neighbour operation. Instead of using an L2 loss, they use the LPIPS perceptual loss by Zhang et al. [18], which is better at capturing perceptually rich details of the image. Finally, VQGAN introduces a patch-based adversarial strategy with a discriminator D that learns to distinguish real from fake patches, therefore the decoder G gets better at generating realistic samples.

The original VQGAN loss can be expressed as follows,

$$\mathcal{L}_{VQGAN} = \mathcal{L}_{VQ}(E, G, \mathcal{Z}) + \lambda \mathcal{L}_{GAN}(\{E, G, \mathcal{Z}\}, D), \quad (1)$$

where $\mathcal{L}_{VQ}$ is the vector quantization loss, and $\mathcal{L}_{GAN}$ is a patch-based Hinge loss, and λ is an adaptive weight for $\mathcal{L}_{GAN}$ (refer to [9] for detailed derivation).

4.2. OCR Perceptual Similarity

We propose a new OCR perceptual loss for rendering clear texts in generated images. We use a frozen pre-trained CRAFT [24] model, which is a text detector trained to localize individual characters in natural images. The model is based on VGG16 [22] with batch normalization as backbone and uses a U-net [32] architecture with skip connections in the upsampling layers. The CRAFT model is kept frozen and adds 20M parameters to the VQGAN architecture, which is comparable to the 14M parameters of LPIPS. As introduced by [18], we forward input patches x and reconstructed patches x_0, which represent the input and reconstructed images, through the OCR model, and extract L feature maps from intermediate layers. Specifically, we

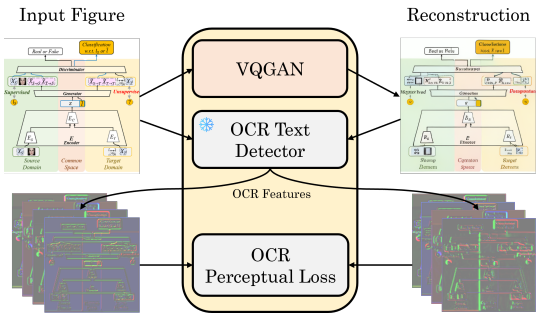


Figure 3. OCR Perceptual loss computation in OCR-VQGAN. OCR feature maps are extracted at intermediate layers and the OCR loss is computed as depicted in equation 4.2.

Method	samples/s	dim $\mathcal{Z}$	Params.
VQVAE	9.69	8192	97M
VQGAN	8.17	16834	92M
OCR-VQGAN	6.35	16834	112M

Table 3. Test time results using Paper2Fig100k test set. The test performs a forward pass of the samples and computes both LPIPS and OCR perceptual similarities, using only 1 V100 GPU.

store the activation map after each upsampling layer. Layers are then normalized in the channel dimension, denoted as $\hat{y}_{hw}, \hat{y}_{0hw} \in \mathbb{R}^{H_l \times W_l \times C_l}$ for each layer l. The OCR perceptual loss is expressed as

$$\mathcal{L}_{ocr} = \sum_{l} \frac{1}{H_l W_l} \sum_{h,w} ||\hat{y}^l_{hw} - \hat{y}^l_{0hw}||_2^2. \qquad (2)$$

Instead of using a network $\mathcal{F}$ with weights w_l to scale feature maps (proposed in LPIPS [18]), we simply average in the spatial dimension and sum over the channel dimension ($w_l = 1 \forall l$). The OCR perceptual loss is added to the loss function in equation 4.1, which then defines the OCR-VQGAN loss.

4.3. Evaluation Metrics

We use image reconstruction-based metrics that allow us to measure the similarity between input and reconstructed images in terms of high-level semantics. To this end, distances in the feature space are more suitable.

FID [33, 34] is a reconstruction metric related to the diversity of the generated images with respect to the original ones. LPIPS [18] is a learned perceptual similarity metric. Both of these metrics measure appearance in reference to natural images, as they use features pre-trained on an Imagenet task. However, these metrics are insensitive to the text generation quality, which plays an important role in figures and diagrams. We introduce a third metric, OCR similarity (OCR-SIM), that quantifies the similarity of images that include text. This metric is computed as in equation 4.2, and it is more adequate to assess the proposed method, where the quality of the generated text is important.

Method	LPIPS↓	OCR-SIM↓	FID↓
Paper2Fig100k			
VQVAE$_{DALLE}$	0.10	0.87	9.91
VQGAN$_{Imagenet}$	0.12	1.04	6.68
VQGAN$_{Paper2Fig100k}$	0.15	1.18	4.37
OCR-VQGAN	**0.07**	**0.42**	**1.69**
ICDAR13			
VQVAE$_{DALLE}$	0.23	1.17	71.84
VQGAN$_{Imagenet}$	0.22	1.61	37.06
VQGAN$_{Paper2Fig100k}$	0.29	1.97	133.97
OCR-VQGAN	0.36	1.26	84.77

Table 4. Quantitative comparison of methods in the reconstruction task. The first part of the table corresponds to the method evaluation on Paper2Fig100k test set, and the second part corresponds to test results on the full ICDAR13.

w_{ocr}	w_{vgg}	LPIPS↓	OCR-SIM↓	FID↓
0.0	1.0	0.15	1.18	4.37
1.0	0.0	0.10	0.50	2.23
1.0	**0.2**	**0.07**	**0.42**	**1.69**
1.0	0.5	0.08	0.46	1.84
1.0	0.8	0.09	0.50	2.03
1.0	1.0	0.08	0.49	2.10

Table 5. Results on Paper2Fig100k test set using OCR-VQGAN. We compare different weighting settings for VGG Perceptual loss (w_{vgg}) and OCR Perceptual Loss (w_{ocr}).

5. Experimental Results

In this section, we conduct experiments to assess the performance of the proposed method in the task of figure reconstruction. Our goal is to obtain the optimal configuration of our method and compare OCR-VQGAN with popular image encoders using reconstruction metrics.

5.1. Training setting

We use images of size 384×384, which we empirically set to maximize resolution and GPU memory. Images are resized to the smallest size between H and W, and randomly cropped. The baseline VQGAN architecture has 112M parameters, a codebook of size $16,384$ with embeddings vectors of size 256, and it is pre-trained on Imagenet (VQGAN$_{Imagenet}$). VQGAN encodes images with a downsampling factor of 16, resulting in grids of 24×24 (or

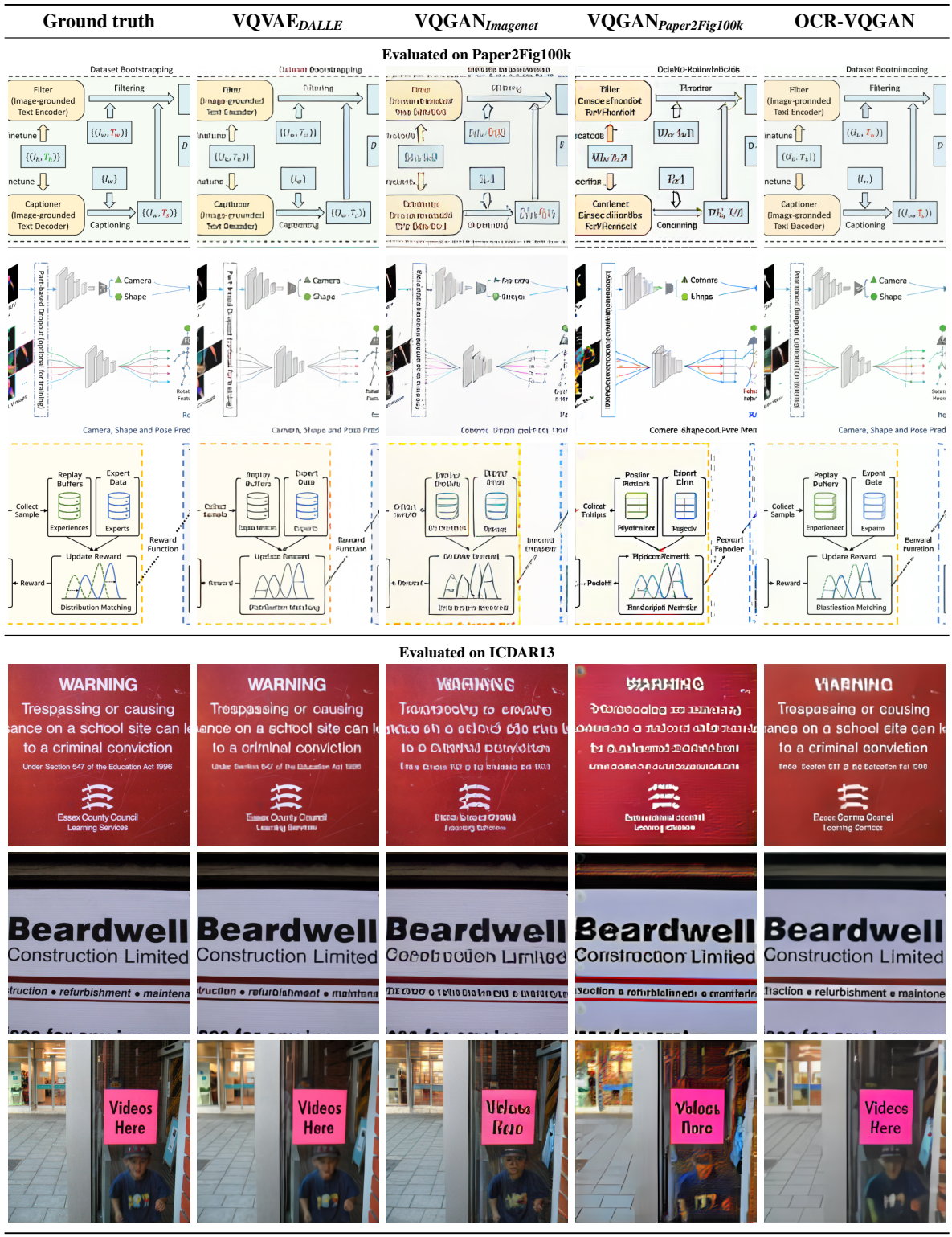


Table 6. Qualitative results on the reconstruction task of different image encoders. OCR-VQGAN outperforms other methods in figure-based images (Paper2Fig100k) both in the clarity of texts and diagrams. Results regarding ICDAR13 dataset (never seen in training) show that VQVAE gives similar results to OCR-VQGAN. It can be seen that OCR-VQGAN highlights text regions and adds "figure style" to natural images.

sequences of 576 tokens). We also use a pre-trained VQ-VAE model from DALLE [10], which consists of a codebook of size 8192.

For training the models, we use data parallelism with 4 V100 GPUs with a total effective batch size of 16 for 20 epochs. We perform an initial warm-up that does not use the discriminator, as has been empirically found beneficial for better reconstructions [9]. We use the Adam optimizer with a learning rate of 4.5×10^{-4}.

5.2. Training datasets

We train OCR-VQGAN using two datasets of images that contain text. **Paper2Fig100k**, presented in Section 3, contains $81,194$ training samples and $21,259$ test samples of figure diagrams with rendered text. **ICDAR13** [35], was presented during the ICDAR 2013 Robust Reading Competition, focused on the task of scene text detection. The dataset is composed of high-resolution natural images with texts in English to test methods when texts are displayed in natural scenes. The dataset contains 229 samples for train and 233 for testing, but we use all 462 samples as a test set, as we use this dataset only for evaluation.

5.3. Complementary Perceptual Losses

We perform a hyper-parameter search for OCR-VQGAN over different weighting configurations of the LPIPS and OCR perceptual losses. Specifically, we scale the two losses using weights w_{ocr} and w_{vgg}. Results in Table 4 show that w_{ocr} is more important than w_{vgg} when dealing with figures and diagrams. The use of a small weight for w_{vgg} (between 0.2 and 0.5) gives the best performance. This is because some figures present small regions with natural images. In those regions, the LPIPS features get more activated, and the loss improves.

OCR model overhead. Table 3 reports results on the performance overhead of adding the OCR model to VQGAN, in terms of network parameters and test time. We also show the parameters that define the latent space $\mathcal{Z}$. Test time is measured during the evaluation of Paper2Fig100k test dataset. We also evaluate test speed on the VQVAE model from DALLE [10], which uses a smaller discrete latent space. This result shows an acceptable increment of test time given the increase in parameters (20M) and the gain in qualitative performance.

5.4. Evaluation of image tokenizers

We analyze model performance on the tasks of figure-based reconstruction (Paper2Fig100k) and natural text-based reconstruction (ICDAR13) (Tables 5 and 6). The proposed OCR-VQGAN model outperforms the other methods both quantitatively and qualitatively, being able to display almost all details in the figures. One limitation is that text can only be recovered when it is sufficiently large. This can be solved using larger resolutions or upscaling models. We found that vertical text is also challenging to reconstruct as well as uncommon background colors.

Results on IDCAR13 dataset show acceptable LPIPS and OCR-SIM, even though the models were not trained on that dataset. VQVAE and VQGAN show better FID scores because they were trained using natural images. VQVAE displays most of the natural text but fails with small text sizes. OCR-VQGAN, fine-tuned with figures, reconstructs appealing natural images and displays texts. It also reconstructs images with its own "figure style" by smoothing textures and highlighting the rendered texts. As expected, its main limitation is that it fails when text appears small, with orientation, and with infrequent color and background combinations (see the "Warning" example from Table 5).

6. Conclusion

We focused on generating diagrams with clear texts within images. We proposed OCR-VQGAN as an image encoder and decoder to improve within-image text generation. We add a loss term for OCR perceptual similarity as a complement to the default VGG LPIPS to a VQGAN architecture. In addition, we presented Paper2Fig100k, the first text-to-image dataset in the domain of research papers and figures. We conducted several experiments that demonstrate how the OCR perceptual loss is beneficial for generating clear texts and diagram shapes. Results show that a small weight for the VGG term is also beneficial. We hope our work constitutes a first stepping stone towards text-to-figure generation.

7. Ethics and social impact

In this work, we focus on generating paper figures and argue how this application could be useful for researchers in the process of generating understandable diagrams, and potentially help a broad audience when creating appealing and effective slide presentations. However, a central concern of this system is that it could be used for fake paper generation and for bypassing plagiarism detection systems.

Some steps that can be made to address this ethical concern is to build classifiers that allow for the detection of fake or plagiarized content. Experiments can be done in order to use the acquired knowledge learned by text-to-image models to train a discriminator. In the Parti paper [4], authors propose the use of watermarks in the generated images, in order to easily detect AI-generated samples. Also, the proposed dataset can be leveraged to train plagiarism detection systems for research publications.

Further research is needed to elucidate how these systems should be made public in order to align their behavior with ethical standards.

References

[1] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *arXiv*, abs/1912.04958, 2019.

[2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv*, abs/2204.06125, 2022.

[3] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv*, abs/2205.11487, 2022.

[4] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *arXiv*, abs/2206.10789, 2022.

[5] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. *arXiv*, abs/2105.13290, 2021.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv*, abs/2112.10752, 2021.

[7] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors, 2022. URL https://arxiv.org/abs/2203.13131.

[8] Aditi Roy, Ioannis Akrotirianakis, Amar V. Kannan, Dmitriy Fradkin, Arquimedes Canedo, Kaushik Koneripalli, and Tugba Kulahcioglu. Diag2graph: Representing deep learning diagrams in research papers as knowledge graphs. In *International Conference on Image Processing (ICIP)*, 2020.

[9] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *arXiv*, abs/2012.09841, 2020.

[10] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv*, abs/2102.12092, 2021.

[11] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Le Khac, Luke Melas, and Ritobrata Ghosh. Dall·e mini, 7 2021. URL https://github.com/borisdayma/dalle-mini.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv*, abs/1706.03762, 2017.

[13] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv*, abs/1711.00937, 2017.

[14] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv*, abs/1503.03585, 2015.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv*, abs/2103.00020, 2021.

[16] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv*, abs/1910.10683, 2019.

[17] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *NeurIPS*, 32, 2019.

[18] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *arXiv*, abs/1801.03924, 2018.

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv*, abs/1812.04948, 2018.

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[21] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric, 2015. URL https://arxiv.org/abs/1512.09300.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[24] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. *arXiv*, abs/1904.01941, 2019.

[25] Jobin kv, Ajoy Mondal, and C. Jawahar. Docfigure: A dataset for scientific document figure classification. pages 74–79, 09 2019. doi: 10.1109/ICDARW.2019.00018.

[26] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv*, abs/1710.07300, 2017.

[27] Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. SciCap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.findings-emnlp.277.

[28] Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, Tong Yu, Ryan Rossi, and Razvan Bunescu. Figure captioning with reasoning and sequence-level training, 2019. URL https://arxiv.org/abs/1906.02850.

[29] Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. On the use of arxiv as a dataset. *arXiv*, abs/1905.00075, 2019.

[30] Patrice Lopez and Luca Foppiano. Grobid. https://github.com/kermitt2/grobid, 2008–2022.

[31] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *arXiv*, abs/1507.05717, 2015.

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

[33] 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. *ArXiv*, abs/1706.08500, 2017.

[34] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in pytorch, 2020. URL https://github.com/toshas/torch-fidelity. Version: 0.3.0, DOI: 10.5281/zenodo.4957738.

[35] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazàn Almazàn, and Lluís Pere de las Heras. Icdar 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1484–1493, 2013. doi: 10.1109/ICDAR.2013.221.