

## Appendix A. Additional qualitative results

In this section, we give additional qualitative results on both Paper2Fig100k and ICDAR13 datasets. We perform a random sampling of the generated images in the test sets and display the comparison with different methods. We aim at finding the limitations of the proposed image encoder when rendering texts within figures and text-in-the-wild.

Regarding results on Paper2Fig100k dataset, shown in tables 7 and 8, OCR-VQGAN outperforms the other methods in almost all scenarios (challenging shapes, text sizes, orientations, and colors). It gives qualitatively better results concerning both VQGAN alternatives (Imagenet pre-trained and Paper2Fig100k finetuned), showing that the OCR loss is beneficial when reconstructing text-within-figures. VQVAE, trained for DALLÉ, gives acceptable results when conditions are favorable, such as having a simple background-color combination or when texts have sufficiently large sizes. With long words or sentences, OCR-VQGAN can display sharper characters, whereas other methods tend to merge them. Arrows, straight lines, or dashed lines are sharper in OCR-VQGAN. We can observe a trade-off between the blurriness and clearness of text in VQVAE and VQGAN, where VQVAE generates more blurry samples. The limitations of our method are shown when complex color-background combinations are present, when the text is very small (low resolution) and when the text is displayed in a vertical orientation.

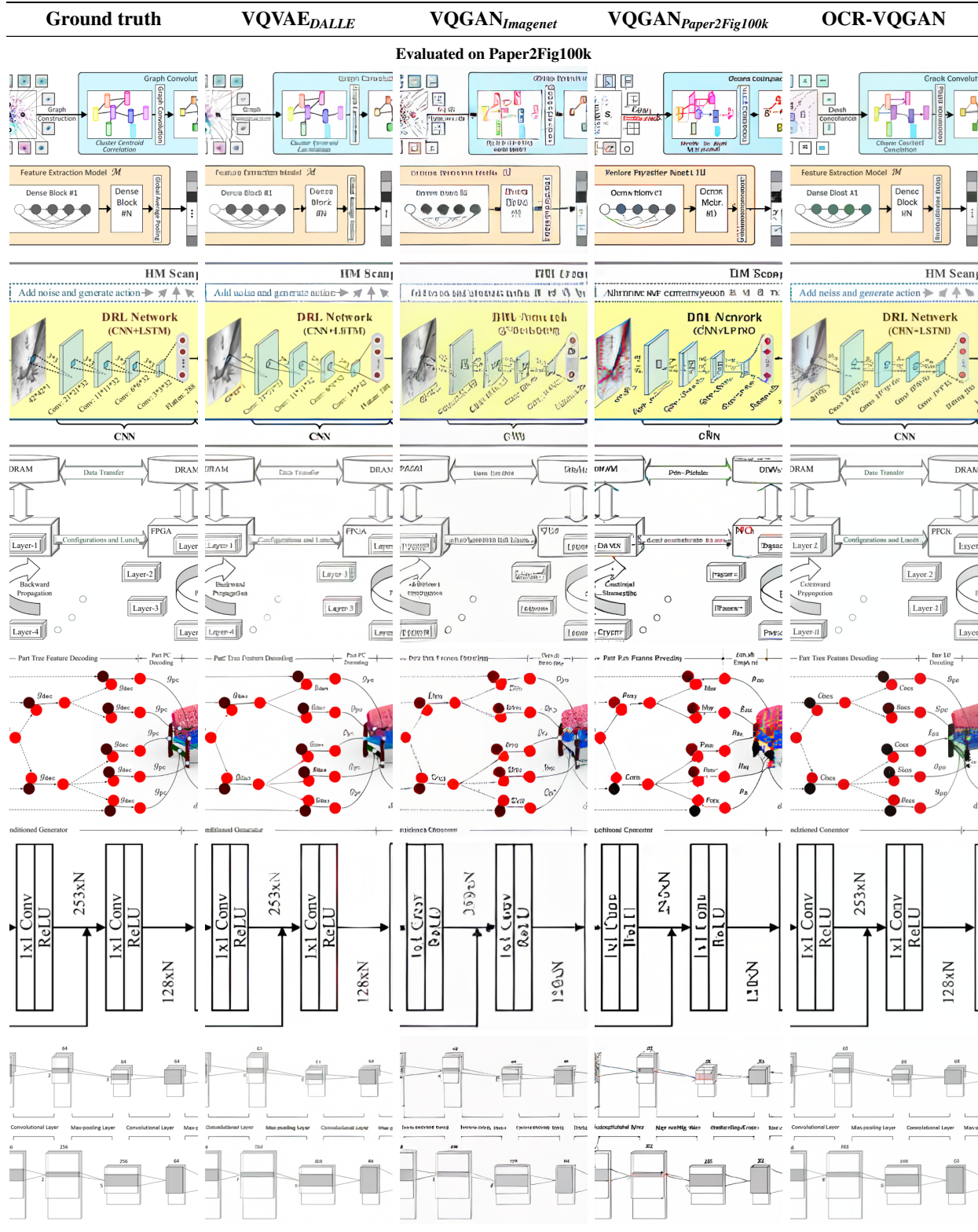
Qualitative results on ICDAR13 dataset are presented in tables 9 and 10. None of the methods were trained using these images, therefore the goal is to test how they perform in never-seen images of text-in-the-wild. The proposed OCR-VQGAN, even though it was trained with 80k images of figures, can reconstruct appealing natural images with in-the-wild texts. As shown in most of the samples, it is transferring the style of figures, tending to smooth out the textures, and highlighting the texts. However, it is sensitive to complex lighting, textures, and image quality conditions. Text is not readable in some cases. VQGAN approaches fail at the task of text reconstruction, where texts are mostly unreadable. VQVAE gives good reconstruction results in ICDAR13, generating natural-looking images and mostly readable text. This is because VQVAE was trained using natural images and it can handle challenging in-the-wild conditions. OCR-VQGAN tends to focus the attention on the texts, while VQVAE gives a more smooth generation in ICDAR13.

## Appendix B. OCR Perceptual loss in PyTorch

We present a Python implementation of the OCR Perceptual loss. The OCR perceptual loss operation accepts input and reconstructed images and has access to the OCR model that computes OCR features.

```
1 import torch
2
3 def normalize_tensor(x, eps=1e-10):
4     norm_factor=torch.sqrt(torch.sum(x**2, dim=1))
5     return x/(norm_factor+eps)
6
7 def ocr_perceptual_loss(image, reconstruction):
8     input_ocr_layers=ocr_model(image)
9     rec_ocr_layers=ocr_model(reconstruction)
10
11     ocr_loss=0
12     for l in layers:
13         in_feat=normalize(input_ocr_layers[l])
14         rec_feat=normalize(rec_ocr_layers[l])
15
16         diffs=(in_feat - rec_feat)**2
17         diffs=diffs.sum(dim = 1)
18         ocr_loss+=diffs.mean([2, 3])
19
20     return ocr_loss
```

Listing 1. Implementation of OCR Perceptual loss in Python (PyTorch)



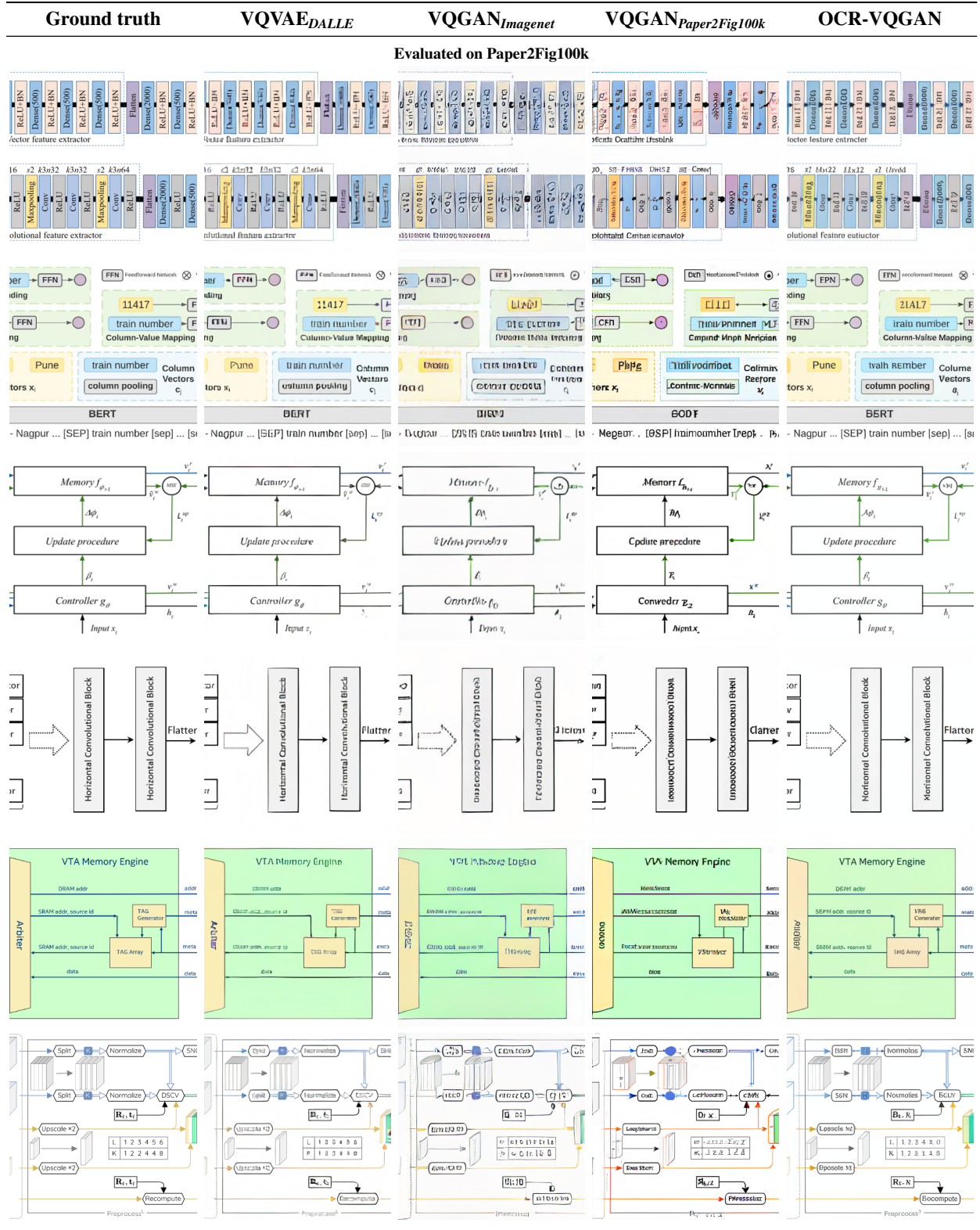


Table 8. Reconstructed images from Paper2Fig100k test set.


















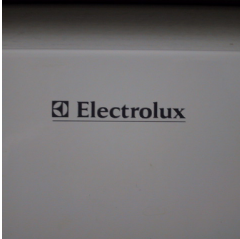

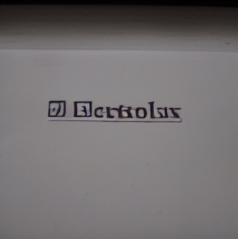

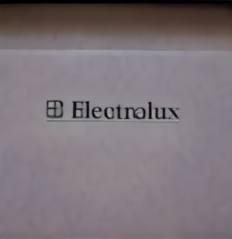





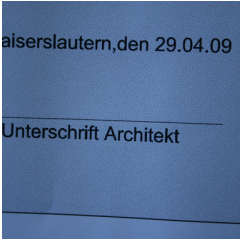
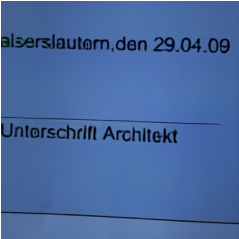
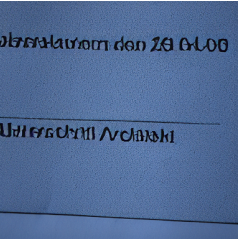
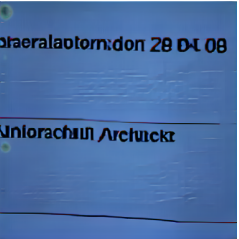
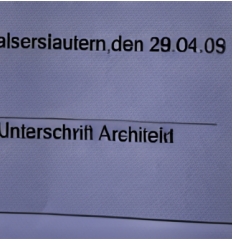
Ground truth	VQVAE <sub>DALLE</sub>	VQGAN <sub>Imagenet</sub>	VQGAN <sub>Paper2Fig100k</sub>	OCR-VQGAN
Evaluated on ICDAR13				
				
				
				
				
				
				

Table 9. Reconstructed images from ICDAR13 test set.



Ground truth	VQVAE <sub>DALLE</sub>	VQGAN <sub>Imagenet</sub>	VQGAN <sub>Paper2Fig100k</sub>	OCR-VQGAN
Evaluated on ICDAR13				

Table 10. Reconstructed images from ICDAR13 test set.