Deep Learning for Document Layout Generation: A First Reproducible Quantitative Evaluation and a Baseline Model

Romain Carletto, Hubert Cardot, Nicolas Ragot

Deep Learning for Document Layout Generation: A First Reproducible Quantitative Evaluation and a Baseline Model

Romain Carletto[1], Hubert Cardot[1], and Nicolas Ragot[1]

Université de Tours, LIFAT EA 6300, Tours, France
{romain.carletto,hubert.cardot,nicolas.ragot}@univ-tours.fr

Abstract. Deep generative models have been recently experimented in automated document layout generation, which led to significant qualitative results, assessed through user studies and displayed visuals. However, no reproducible quantitative evaluation has been settled in these works, which prevents scientific comparison of upcoming models with previous models. In this context, we propose a fully reproducible evaluation method and an original and efficient baseline model. Our evaluation protocol is meticulously defined in this work, and backed with an open source code available on this link : https://github.com/romain-rsr/quant_eval_for_document_layout_generation/tree/master

Keywords: Document Layout Generation · Quantitative Evaluation · Generative Adversarial Network.

1 Introduction

1.1 Document Layout Generation

For decades, developments in information and communication technologies leveraged interest in automated document layout generation. This application usually consists in automatically laying out elements on a canvas of given dimensions. It can take as inputs a random vector and additional optional features such as elements categories, reading order of the elements, element contents (texts or images) and geometric constraints such as aspect ratios or areas of the elements. While the outputs can take many forms, it is usually a list of bounding boxes, one for each element, with for each of these boxes its category, its dimensions and positions on the canvas. While former automated attempts in this field faced lack of both functionality and flexibility, recent solutions based on deep generative models reached interesting possibilities and provided encouraging visual and user study results. However, quantitative evaluations of these later solutions show important deficiencies, which prevents any scientific comparison with upcoming works.

1.2 A Tricky Quantitative Evaluation

Quantitative evaluation of generative models has always been a delicate matter, and is even more complicated in the specialized field of document layout generation. Real world layout guidelines are mainly implicit, even to layout designers themselves, and while metrics from image generation field are still being adapted, standard metrics from document layout community do not apply. As an example, the IoU score which is commonly used in layout analysis has limited utility in layout generation, where different positions for a same element can be of similar quality.

While in related works interesting quantitative evaluation methods have emerged and are progressively converging to a common standard, none of these works, after a meticulous study, have provided a reproducible and reliable definition of a quantitative evaluation method. In this context, the comparison between a new model and previous works relies mainly on subjective interpretations of visual results, which are of high interest but not sufficient to match hard sciences principles. Therefore, this paper aims to propose, on main document layout datasets, a first reproducible quantitative evaluation of any document layout generation model, as well as an original baseline model.

2 Related Works

2.1 From Explicit Methods to Deep Learning Methods

Former works in automated document layout generation focused on turning layout guidelines into explicit and static quantitative rules. Earliest works were based on templates [1,2] or on interactive tools combining basic layout rules [3] and showed poor possibilities. Later works were based on geometrical objective functions to be optimized [4,5,6] faced qualitative biases and a strong lack of flexibility. To tackle these various shortcomings, recent researches focused on learning methods where guidelines are turned in a dynamic and data-driven way into implicit quantitative rules projected in a multidimensional feature space. More specifically, most recent researches have been based on deep learning generative models, which can learn very complex rules from simple objective functions, and are already successful in other generative tasks such as image or video generation.

In [7], a Generative Adversarial Network (GAN) generates layouts from only a random vector as input. It can generate layouts of documents, layouts of pixels subset (extracted from MNIST handwritten digits images), clipart scenes and geometric tangrams. In [8], authors extend [7] previous work to propose two main applications : layout generation and layout adjustment. Layout generation is split into three sub-applications. Image layout generation takes geometric constraints as input and generates several layouts propositions containing only the product image. Attribute-guided layout generation takes these first layout propositions as input and add additional elements, according to these elements attributes, to produce different propositions of complete layouts. Finally, grouping and ranking

application allows to select the best layout among different layout propositions, through overlap, alignment and discriminator scores.

The model presented in [9] is based on graph neural network and Variational Auto-Encoders (VAE) and generates a layout iteratively, element after element. It takes as inputs a set of elements, the order of elements to be iteratively laid out, and optional user-specified constraints. They experimented their models on generating application page layouts, magazine page layouts and private web advertising layouts.

In [12], authors merge GAN and VAE architectures to propose a model that encodes element attributes at different steps : when encoding a real layout into a latent vector, when generating a layout from a latent vector and when classifying a layout to be real or generated. These attributes can thus be learned in association with real training layouts during the training step and added as inputs, aside with a random vector, when using the model. They contain particularly rich and diversified information, such as image contents, text and label values than can be both continuous (such as aspect ratios) or discrete. Author architecture also allows the users to add optional soft constraints as inputs, such as reserved area for specific types of elements. Experiments are focused on generating magazine page layouts.

While not exactly focusing on document layout generation, [13] proposes a model based on VAE to generate real scene layouts. Generation steps are divided into several sub-steps : a first VAE generates a list of elements, then another VAE generates bounding boxes from it. The model is applied on a dataset adapted from MNIST, where handwritten digits are laid out on a black screen, and is also applied on COCO dataset, containing real life scenes with for each scene a labeled bounding box for every important object and person in the scene. Other moderately related but yet interesting works, [14] and [15], consider a potential sub-task of layout generation and generate bounding boxes from graphs in which strong relational and geometrical constraints are already indicated, e.g. which element must be at the right of which element, or which element has to be bigger than which element.

2.2 Deficiencies in Existing Quantitative Evaluation Methods

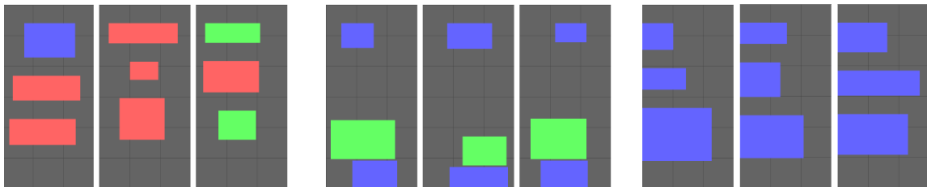
Previous works usually display visual results and often provide user studies of their experiments, but while this information can draw attention on the presented models, it can't allow any objective comparison with upcoming methods. In parallel, quantitative evaluation methods are also presented in the course of these works, but are not defined in a valid and reproducible way.

While some authors [12] do not provide any quantitative evaluation at all, others laid incomplete groundwork to define a valid quantitative metric. In [7], absolute alignment and overlap scores are provided for generated document layouts but the dataset on which these scores have been obtained, referred only as "the document layout dataset", has not been made available. The exact same problem is found in [8] where alignment and overlap scores are given for a private advertising layout dataset, not available to the scientific community.

In [9], authors propose an alignment evaluation on two public datasets but while our understanding of their alignment evaluation method has been confirmed by the authors, our results differ from theirs by orders of magnitude when we apply this method on the ground truth layouts of both evaluated datasets. Aside from the alignment evaluation method, [9] proposes an evaluation based on the Fréchet Inception Distance (FID). As described in their works, this method relies on the feature distribution of the penultimate layer of the discriminator : the distribution obtained on real layouts is compared with the distribution obtained on generated layouts and produces a score ranging from zero to positive infinity, with lower values indicating better performances. Given a fixed discriminator architecture with fixed feature parameters and fixed hyperparameters, this evaluation method allows for a quantitative comparison between two different generative models. But in [9], these parameters are missing.

3 Experimental Protocol

3.1 Datasets



(a) (b) (c)

Fig. 1. Layouts from the first synthetic dataset. (a) layouts have been produced through general rules only, while (b) and (c) layouts have been produced by two different specific rules.

Synthetic Dataset As explained in introduction, real world layout guidelines are mainly implicit which makes it difficult to quantitatively evaluate the quality of a layout. In this context, [10] proposes two synthetic datasets. The first dataset contains 100,000 synthetic document layouts, in which elements have fictitious semantic categories and are laid out following a combination of arbitrary and fictitious layout rules, that are both hard to learn and easy to quantitatively evaluate. A second dataset contains 100,000 other synthetic document layouts, in which elements have this time similar categories as in the web advertising industry : Product Image, Text, Call-To-Action (CTA) and Logo. In this second dataset, elements are laid out according to basic rules and distributions of the web advertising industry (e.g. logos are always on the top or on the bottom of the layout, in a majority of layouts images are bigger than CTAs and logos, ...). These

layout rules are thus both realistic and easy to quantitatively evaluate. Any learning document generation model can thus be trained on these two synthetic datasets and then be evaluated in an exact and quantitative way, through reusing the explicit rules that were used in the first place to produce each of these datasets.

In the first dataset, while general rules alone have been used to create the majority of the layouts, additional specific rules have been used when creating certain layouts, as seen in Fig. 1. The use of these additional rules and their precedence on general rules is triggered by specific sequences of categories within the list of elements in each layout to be generated. Furthermore, some of the specific rules are deliberately in contradiction with the general layout rules to challenge the ability of a layout generation model to discern patterns within a complex and implicit combination of layout rules, similarly to what is expected from the model in a real case document generation application. The datasets are available at the following link : https://github.com/romain-rsr/synth_datasets_for_web_advertising_layout/tree/master

RICO This public dataset [11] contains 65,538 ground truth layouts extracted from various application pages. Numerous works have experimented on this dataset for its quality and for the high quantity of layouts it contains. Elements within the layouts can be nested and can present very different sizes and aspect ratios. RICO dataset can be found at this link : http://interactionmining.org/rico

Magazine Magazine dataset [12] contains 3,919 magazine page layouts, that can be more similar to advertising layouts than RICO application page layouts. This is of great interest since there is no public dataset available for advertising layouts but high industrial application in this field for sophisticated automated layout. The dataset is available at this link : https://xtqiao.com/projects/content_aware_layout/

3.2 Baseline Model

Architecture Our model is a GAN with both generator and discriminator based on a fully connected residual block architectures. Instead of the usual residual block containing two layers, residual blocks within our model are tangled and contain one layer each. As seen in Fig. 2, both generator and discriminator contain 5 layers assembled in tangled residual blocks, each containing 100 neurons except the last layer. In the generator, the number of neurons in the last layer is equal to the product of the number of elements to be laid out and the number of feature for each element. In discriminator, the last layer contain only one neuron to output the cagory of the input layout : real or fake. In both generator and discriminator, each layer is activated by a ReLU activation function except the last layer activated by a sigmoid function.

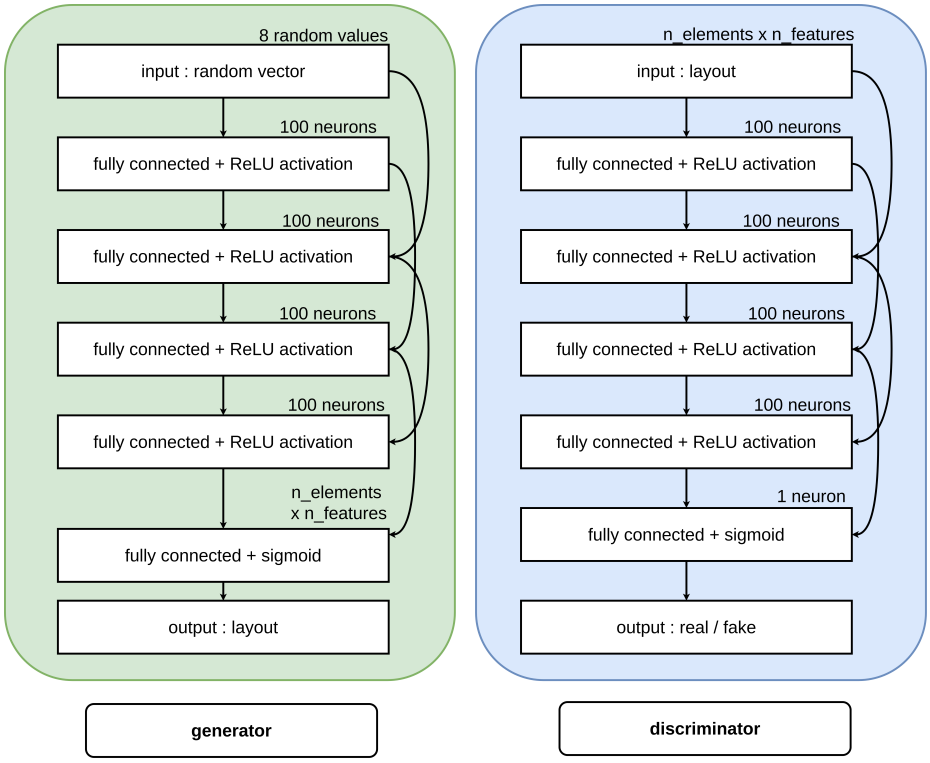


Fig. 2. Our generator and discriminator architectures.

Training Our experimental protocol can be applied on layouts of any number of elements. Yet, layouts have first to be grouped according to the number of elements they contain so the experiments are run on each group separately, with the number of elements per layout being given as an input parameter of our model and evaluation process. For greater simplicity in the analysis of our experiments, the presented results focus on layouts with three elements.

Model's performance is monitored through the binary cross-entropy loss function during training, and the model is trained through Adam optimizer, which extends rmsprop optimizer on one hand, by adapting its learning rate to each parameter, and extends adagrad optimizer on an other hand, by adapting its learning rate to first and second moment (respectively mean and uncentered variance) of recent gradient magnitudes. In each experiment, we select a subset of the complete dataset and split it in train, validation and test sets, respectively of 60%, 20% and 20% of the subset.

3.3 Quantitative Evaluation Metrics

Our quantitative evaluation is based on four metrics : the Fréchet Inception Distance (FID), the Comparative Alignment Score, the Comparative Overlap Score and the Comparative Diversity Score. Some of them are inspired and revised version of incomplete yet interesting quantitative evaluations described in related works. Both pseudo-code and ready-to-use code of these evaluations are publicly available in our git, aside with real and generated evaluated samples.

Fréchet Inception Distance This metric, initially defined in [16], is applied as specified in [17] to the output feature distribution within the discriminator penultimate layer. The mean and the covariance matrix of these features are first computed when applying the discriminator on generated layouts, then the same operation is carried out with real layouts. Finally, mean μ_X and covariance matrix $\sum_X$, obtained on generated layouts, are compared with mean μ_Y and covariance matrix $\sum_Y$, obtained on real layouts, following equation :

$$d^2(F,G) = |\mu_X - \mu_Y| + tr(\sum_X + \sum_Y - 2(\sum_X \sum_Y)^{1/2}) \quad (1)$$

The resulting score is an absolute value which must be as low as possible. As specified earlier, this evaluation is of interest only if fixed architecture, feature parameters and hyperparameters are settled for the discriminator used in the evaluation, so that when comparing two models, the shift in the FID score comes only from the shift in the similarity between real and generated layouts. Also, using a discriminator of a specific existing work in layout generation would give an unfair advantage to this work when comparing new models with it, so a standard and non specialised discriminator architecture has to be used instead. Therefore, we used the pre-classification layers of the open source inception v3 discriminator, whom fixed parameters are indicated in our git.

Absolute Alignment Score In related works, alignment score is often mentioned, as an absolute measure applied on generated layouts only, and measures the closest possible alignment of elements on each layout, according to one of the possible vertical alignment axes (element lefts, centers or rights). Here is the [9] definition of this absolute alignment score :

$$alignment_{gen} = \frac{1}{N} \sum_k \sum_i \underset{j,i \neq j}{min} \{min(l(e_i^k, e_j^k), m(e_i^k, e_j^k), r(e_i^k, e_j^k)\}) \quad (2)$$

Where N is the total number of generated layouts, e_i^k and e_j^k are the i_{th} and j_{th} elements of the k_{th} generated layout and where l,m and r are respectively the distances between lefts, centers and rights of two considered elements.

Comparative Metric Absolute score can be interpreted very differently according to the type of evaluated documents. On RICO application pages, for example, lower alignment score could be preferred by designers while on advertising layouts, greater misalignment can be a quality factor so absolute alignment score has no general objective value. Therefore, our evaluation goes further and compares the absolute alignment score obtained on generated layouts with the absolute alignment score obtained on real layouts. While simply computing the ratio between two absolute scores would have been a straightforward comparison, it is not adapted if the score used as the divider is equal to zero. To overcome this problem and obtain similar score scales as in the FID evaluation, the following comparative metric has been defined to compare absolute scores $score_{gen}$ and $score_{real}$, obtained respectively on generated and real layouts :

$$comp(score_{gen}, score_{real}) = |\log \frac{score_{gen} + 1e{-}10}{score_{real} + 1e{-}10}| \tag{3}$$

Comparative Alignment Score We apply the comparative metric on the couple of absolute alignment scores obtained on real and generated layouts to get the comparative alignment score, which is finally a robust and objective indication of how similar generated layouts are to the real layouts, independently of the subjective interpretation of what is a good absolute alignment value within each dataset :

$$alignment_{comp} = comp(alignment_{gen}, alignment_{real}) \tag{4}$$

Absolute Overlap Score In parallel to the alignment measure, absolute overlap score is used in several related works and measures the ratio of overlapping areas over the total canvas area :

$$overlap_{gen} = \frac{1}{N} \sum_{k} \sum_{i} \sum_{j,j<i} \frac{intersection(area_i^k, area_j^k)}{area_c^k} \tag{5}$$

Where N is the number of generated layouts, where $area_c^k$ is the total canvas area of the k_{th} generated layout (in most datasets this value is constant) and where $area_i^k$ and $area_j^k$ are the respective areas of i_{th} and j_{th} elements of the k_{th} generated layout

Comparative Overlap Score Absolute overlap score encounters the same interpretation concerns as the absolute alignment score : RICO layouts, for example, present nested elements, which are fully overlapping, while in other types of document layouts, overlaps are unacceptable. Therefore, as for the alignment evaluation, our final overlap score is obtained by applying the comparative metric to the absolute overlap scores obtained on generated and real layouts, and

hence benefits of the same objectiveness and generalisation as our alignment comparative score :

$$overlap_{comp} = comp(overlap_{gen}, overlap_{real}) \tag{6}$$

Comparative Diversity Score Our evaluation method additionally incorporates the comparative diversity score, which compares standard deviations of real and generated layouts features. As an example, a standard deviation is computed for the left position of the first element, over all generated layouts, and one standard deviation is computed for each pair of element rank - element feature. The minimum value of these standard deviations is then computed for the generated layouts. The same operation is applied on real layouts, then real and generated obtained minimums are compared through the comparative metric.

$$diversity_{comp} = comp(\min_{i,j}\{\sigma^{ij}_{gen})\}, \min_{i,j}\{\sigma^{ij}_{real}\}) \tag{7}$$

Where σ^{ij}_{gen} (respectively σ^{ij}_{real}) is the standard deviation of the j_{th} feature of the i_{th} element over all generated layouts (respectively over all real layouts).

Applying a same comparative metric on different datasets Note that comparative alignment score is neither penalizing or rewarding alignment, it is only penalizing differences between the absolute alignment score obtained on generated layouts, and the one obtained on real layouts. The same reasoning applies for comparative overlap score and for comparative diversity score. Therefore these comparative scores can be easily compared from different datasets, even with high variation of any given property between and within those datasets (such as the number of layout elements).

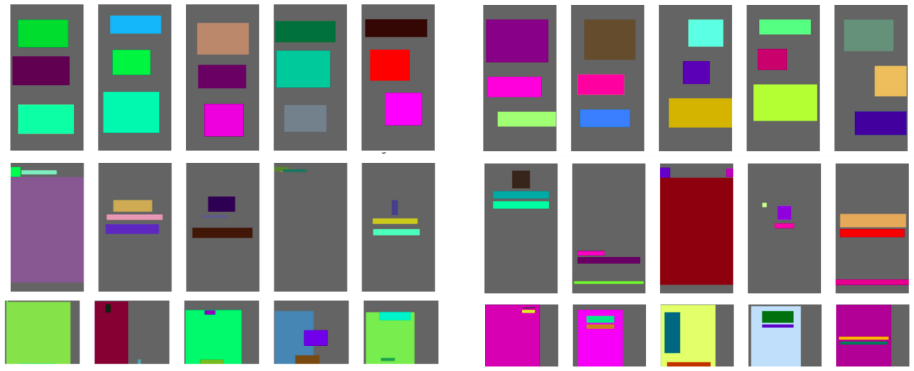
Independence between training metrics and comparative metrics Comparative metrics are used only during evaluation and are not used at all during training, so that evaluation scores remain independent of the training process. The only metric used during training is the binary cross-entropy loss function, which is agnostic to our evaluation metrics.

3.4 Baseline Evaluation Results

As specified earlier in the related work section, there is no reproducible quantitative baseline for document layout generation, which make impossible for us to compare our model to previous work. As an example, [9] unconstrained document layout generator achieves an FID score of 143.51 on RICO dataset while our model achieves an FID score of 66.96 on the same dataset. While these results could show that our model allow a significant performance gain, this actually cannot be asserted since the discriminator architecture and parameters used for the FID computation in [9] is not available.

Table 1. Quantitative evaluation results

Quantitative Metric	Synthetic dataset I	RICO dataset	Magazine dataset
Fréchet inception distance	33.86	66.96	90.15
comparative alignment score	0.25	0.29	0.32
comparative overlap score	12.32	17.69	$3.59e^{-4}$
comparative diversity score	19.83	0.65	1.37



(a) (b)

Fig. 3. Generated (a) and real (b) layouts from different datasets. First row layouts are from the first synthetic dataset, second row layouts are from RICO dataset and third row layouts are from Magazine dataset.

Therefore, we propose our own quantitative baseline results, indicated in Table 1. These results are also available in our git along with the evaluation functions and the real and generated layouts that have been used to obtain them. While scores are globally of the same order of magnitude from one dataset to another, some score discrepancies remain noteworthy. FID is particularly lower on synthetic dataset, where size and location ranges are smaller and where there is no nested element. On this dataset, diversity score is also much higher, which is due to first element top position being constant on each real layout. Therefore, even a short deviation in generated layouts corresponding feature is highly penalized. Finally, due to a high number of nested elements, real layouts overlap score is not as tight in Magazine as in the other datasets. Therefore, the absolute difference in overlap score between generated and real layouts is less sensitive and penalized in this dataset, which explains such a low comparative overlap score for Magazine Dataset.

In order to put into perspective these quantitative results, generated layout representations have been randomly selected and are displayed along with real layout representations in Fig. 3. We can see on results obtained on synthetic dataset that only one layout, among all displayed generated layouts, contains an overlapping error. On results obtained on RICO dataset, we see strong similarity between generated and real layout patterns, e.g. a very large bounding box

covering the majority of the canvas with much smaller, horizontally aligned bounding boxes above it. We see in both generated and real RICO layouts that elements are generally extremely close to each other without overlapping, which shows that our model matches industrial precision standards.

3.5 Additional results : application of the quantitative evaluation metrics on specific examples

Table 2. Intensive training vs. poor training

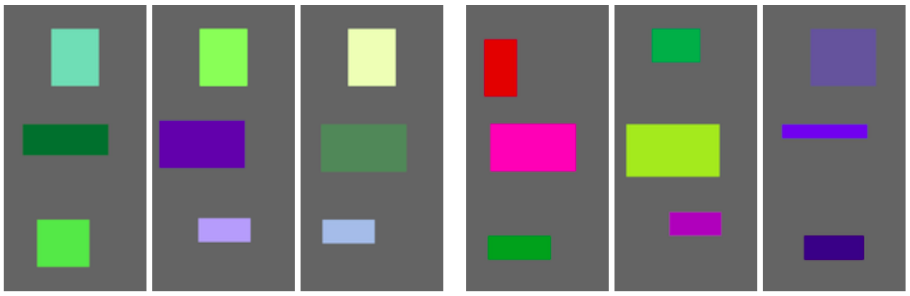
Intensive training on Synth II	**Alignment**	**Overlap**	**Diversity**
Absolute score on real layouts	0.112	0	0
Absolute score on generated layouts	0.116	0	0.003
Comparative score	0.031	0	17.166
Poor training on Synth II	**Alignment**	**Overlap**	**Diversity**
Absolute score on real layouts	0.112	0	0
Absolute score on generated layouts	0.088	1.15e-05	0.022
Comparative score	0.238	11.653	19.230
Intensive training on Rico	**Alignment**	**Overlap**	**Diversity**
Absolute score on real layouts	0.120	0.274	0.234
Absolute score on generated layouts	0.095	0.317	0.174
Comparative score	0.236	0.146	0.293
Poor training on Rico	**Alignment**	**Overlap**	**Diversity**
Absolute score on real layouts	0.120	0.274	0.234
Absolute score on generated layouts	0.019	0.157	0.014
Comparative score	1.830	0.558	2.843

In order to assert the resilience and the versatility of our quantitative evaluation metrics, an additional set of experiments focused on more particular examples. More specifically, these experiments aim to verify that the accuracy and the consistency of our metrics remain proportional to the level of training of the evaluated model and remain independent of datasets properties such as the number of element per layout or the degree of alignment, overlap and diversity within each dataset.

Evaluation scores after intensive training and after poor training We evaluated the consistency of our metrics with respect to the level of training by comparing the obtained results after 100 and 10.000 training epochs. This protocol was first ran on the second synthetic dataset, then ran again on Rico dataset. Results of Table 2 show that on both datasets and on each property (alignment, overlap and diversity), better training leads to better comparative scores. As expected, the consistency of these comparative scores is contrasting

Table 3. Results on datasets with high differences in terms of alignment, overlap and diversity

Highly aligned real layouts (from Rico)	**metric**	**Score**
Absolute score on real layouts	alignment	0.0
Absolute score on generated layouts	alignment	0.00023
Comparative score	alignment	14.657
Poorly aligned real layouts (from Rico)	**metric**	**Score**
Absolute score on real layouts	alignment	0.365
Absolute score on generated layouts	alignment	0.340
Comparative score	alignment	0.070
Highly overlapping real layouts (from Rico)	**metric**	**Score**
Absolute score on real layouts	overlap	0.414
Absolute score on generated layouts	overlap	0.696
Comparative score	overlap	0.518
Poorly overlapping real layouts (from Rico)	**metric**	**Score**
Absolute score on real layouts	overlap	0
Absolute score on generated layouts	overlap	0.001
Comparative score	overlap	16.228
Highly diversified real layouts (from Synth II)	**metric**	**Score**
Absolute score on real layouts	diversity	0.005
Absolute score on generated layouts	diversity	4.39e-04
Comparative score	diversity	2.396
Poorly diversified real layouts (from Synth II)	**metric**	**Score**
Absolute score on real layouts	diversity	0.002
Absolute score on generated layouts	diversity	2e-04
Comparative score	diversity	2.516



(a) (b)

Fig. 4. Poorly diversified (a) and highly diversified real layouts (b) from the second synthetic dataset

with the divergence of some absolute scores. As an example, on the second synthetic dataset, more training leads to a higher absolute overlap score while on Rico, more training leads to a lower absolute overlap score. While the interpretation of these two absolute scores relies on the subjective understanding of their related datasets, a lower comparative score systematically implies a better performance, on any dataset.

Evaluating datasets with high differences in terms of alignment, overlap and diversity . The results in Table 3 have been obtained by experimenting the same model on specifically selected layouts of a same dataset, showing opposite values on a given property (alignment, overlap or diversity). As seen in the section focusing on highly and poorly diversified layouts, we see that comparative diversity scores for both data subsets are in the same order of magnitude, which is consistent with the fact that the same model has been experimented on both subsets of layouts, leading to similar performances. Additionally, the sections focusing on highly aligned and poorly overlapping layouts present very high comparative scores (which implies lower performance). These scores show the ability of the comparative metric to encompass the critical difference between a property equal to zero in real layouts (reflecting a hard constraint) and a very low non-zero value for the same property in generated layouts (which are then missing the hard constraint), by heavily penalizing the relative comparative scores.

Evaluating layouts with different numbers of elements Table 4 shows that the comparative metrics are consistent over layouts of different number of elements. Except the sections where generated layouts do not comply with hard constraints, comparative scores remain in the same order of magnitude, independently of the number of elements. Moreover, in the last section of the table we show that one comparative score can easily be applied to two set of layouts, each containing layouts with a distinct number of elements.

Table 4. Evaluating layouts with different numbers of elements

Layouts with 3 elements (from Synth II)	**Alignment**	**Overlap**	**Diversity**
Absolute score on real layouts	0.023	0	0
Absolute score on generated layouts	0.024	0	0.001
Comparative score	0.031	0	15.715
Layouts with 5 elements (from Synth II)	**Alignment**	**Overlap**	**Diversity**
Absolute score on real layouts	0.043	0	0
Absolute score on generated layouts	0.042	1.56e-05	2.14e-05
Comparative score	0.017	11.960	12.272
Layouts with 7 elements (from Synth II)	**Alignment**	**Overlap**	**Diversity**
Absolute score on real layouts	0.043	0	0
Absolute score on generated layouts	0.048	2.85e-05	2.38e-04
Comparative score	0.115	12.562	14.686
3-elements versus 7-elements layouts	**Alignment**	**Overlap**	**Diversity**
Absolute score on real layouts (3 elements)	0.023	0	0
Absolute score on real layouts (7 elements)	0.041	0	0
Comparative score	0.575	0	0

4 Conclusion

4.1 Contributions

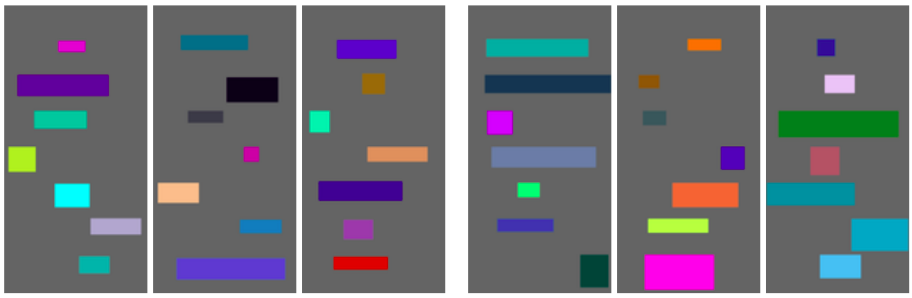
In a context where recent publications on automated document layout generation show off impressive model architectures and promising user studies, we aimed at setting a sorely missing quantitative basis for scientific comparison and cooperation in the field of document layout generation. We thus propose a first baseline and made our quantitative evaluation method fully reproducible, backing it with a turn-key git containing both data and evaluation metrics that led to the presented results. The model we propose is based on an original yet easy to implement adaptation of the residual block concept and shows satisfying results, in both quantitative and visual aspects.

4.2 Future Works

Now that a first reproducible quantitative evaluation is settled, it would be interesting to monitor quantitative performance shifts when adding related work modules to our baseline model, or when adding functionalities such as attribute-guided and constrained layout generation.

Another interesting approach would be to add a background image processing module to our layout generation model, since background and foreground graphical balance are a central problem when generating sophisticated graphical layouts.

Finally, a critical step to achieve in automated document layout generation would be to go beyond the bounding box model and consider generating layouts



(a) (b)

Fig. 5. Generated (a) and real layouts (b) with seven elements, from the second synthetic dataset

in a pixel-wise dimension. This could also have very interesting applications for image generation, allowing users to add specific constraints or attributes on reserved areas of an image to be generated.

5 Acknowledgements

This work is financed by Centre Val de Loire Region, in France, and by Madmix Digital, a creative studio based in Paris and New-York, who helped us to identify and scientifically match the major challenges of document layout generation.

References

1. B.-A. Myers. User Interface Software Tools. ACM Transactions on Computer-Human Interaction. 1994.
2. S. Lok, S. Feiner, G. Ngai. Evaluation of visual balance for automated layout. 9th international conference on Intelligent user interfaces. 2004.
3. S. Feiner, S. Nagy, A. Van Dam. An Experimental System for Creating and Presenting Interactive Graphical Documents. ACM Transactions on Graphics. 1982
4. P. Merell, E. Schkufza, Z. Li, M. Agrawala, V. Koltun. Interactive Furniture Layout Using Interior Design Guidelines. SIGGRAPH. 2011.
5. X. Lin. Active Layout Engine: Algorithms and Applications in Variable Data Printing. Computer-Aided Design. 2005
6. L. Purvis, S. Harrington, B. O'Sullivan, E. Freuder. Creating Personalized Documents: An Optimization Approach. ACM symposium on Document engineering. 2003
7. J. Li, J. Yang, A. Hertzmann, J. Zhang, T. Xu. LayoutGAN: Generating Graphic Layouts with Wireframe Discriminators. ICLR. 2019.
8. J. Li, J. Yang, J. Zhang, C. Liu, C. Wang, T. Xu. Attribute-conditioned Layout GAN for Automatic Graphic Design. IEEE Transactions on Visualization and Computer Graphics. 2020.
9. H.-Y. Lee, L. Jiang, I. Essa, P. B. Le, H. Gong, M.-H. Yang, W. Yang. Neural Design Network: Graphic Layout Generation with Constraints. ECCV. 2020.

10. R. Carletto, H. Cardot, N. Ragot. Automatic Generation of Web Advertising Layouts:A Synthetic Dataset and a Deep Learning Baseline Model. ICPRS. 2021.
11. B. Deka, Z. Huang, C. Franzen, J. Hibschman, D. Afergan, Y. Li, J. Nichols, R. Kumar. Rico: A Mobile App Dataset for Building Data-Driven Design Applications. ACM Symposium on User Interface Software and Technology. 2017.
12. X. Zheng, X. Qiao, Y. Cao, R. W. H. LAU. Content-aware Generative Modeling of Graphic Design Layouts. ACM ACM Transactions on Graphics. 2019.
13. A. A. Jyothi, T. Durand, J. He, L. Sigal, G. Mori. LayoutVAE: Stochastic Scene Layout Generation From a Label Set. ICCV. 2019
14. N. Nauata, K. H. Chang, C.-Y. Cheng, G. Mori, Y. Furukawa. House-GAN: Relational Generative Adversarial Networks for Graph-constrained House Layout Generation. ECCV. 2020.
15. B. Schroeder, S. Tripathi, H. Tang. Triplet-Aware Scene Graph Embeddings. Scene Graph Representation Learning workshop at ICCV. 2019.
16. D. C. Dowson, B. V. Landau. The Fréchet distance between multivariate normal distributions. Journal of Multivariate Analysis. 1982.
17. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. NIPS. 2017.