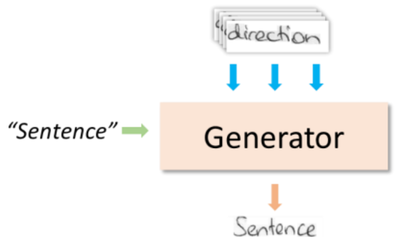
DoodleFormer: Creative Sketch Drawing with Transformers

Ankan Kumar Bhunia[1] Salman Khan[1,2] Hisham Cholakkal[1] Rao Muhammad Anwer[1,4]
Fahad Shahbaz Khan[1,3] Jorma Laaksonen[4] Michael Felsberg[3]
[1]MBZUAI, UAE [2]Australian National University, Australia [3]Linköping University, Sweden [4]Aalto University, Finland

ECCV TEL AVIV 2022

Problem formulation

We are given (a) set of handwritten word images as few-shot calligraphic **style examples** of one writer, (b) **query text** from an unconstrained set of vocabulary, our model strives to generate handwritten images with the same text in the writing style of the given writer.

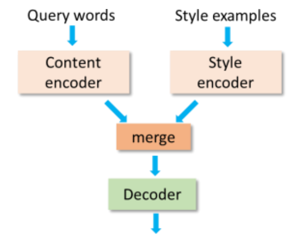


Motivation

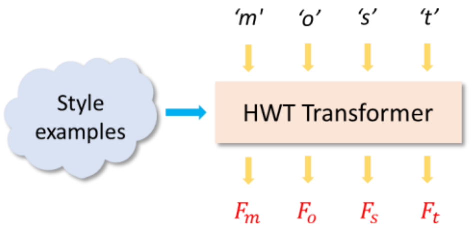
Limitation of existing frameworks

We distinguish the main architectural constraint that impede the quality of handwritten text image generation in the existing GAN-based methods [1,2].

- Separate processing of style and content: In these models, both Style and content are loosely connected as their representative features are processed separately by their respective encoders and then later concatenated.
- Global and Local style imitation: While such a scheme enables entanglement between style and content at the word-level, it does not explicitly enforce style-content entanglement at the character-level. *As a result, they struggle to accurately imitate local styles such as character shapes or ligatures.*



(1) Existing approaches



(2) Our proposed approach

Why Transformer-based Design?

We propose a transformer based design model (HWT).

- Our proposed HWT imitates the style of a writer for a given query content through *self- and encoder-decoder attention* that emphasizes relevant self attentive style features with respect to each character in that query.
- This enables us to *(a) capture style-content entanglement at the character-level,* and *(b) model both the global as well as local style features for a given calligraphic style.*
- Further, such a tight integration between style and content leads to a *cohesive architecture design.*

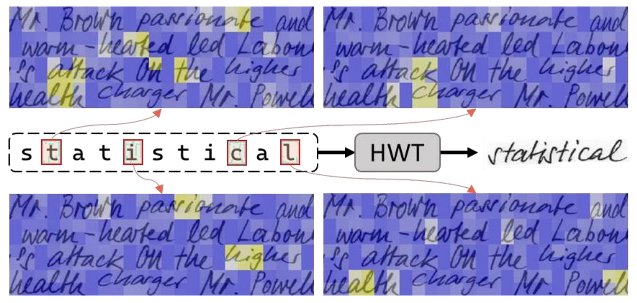
Methodology

- Our proposed generative model (G_θ) comprises an encoder-decoder Transformer network.

Our training algorithm follows the traditional GAN paradigm,

- where a discriminator network (D_ψ) is employed to ensure realistic generation of handwriting styles,
- A recognizer network (R_ϕ) aids in textual content preservation,
- A writer style classifier (S_γ) ensures satisfactory transfer of the calligraphic styles.
- In addition, we use cycle loss. that ensures the original style feature sequence can be reconstructed from the generated image.

Visualization of Attention maps


s t a t i s t i c a l
HWT
statistical
Mr. Brown passionate and warm-hearted led Labour's attack on the higher health charges Mr. Powell

The attention maps are computed for each character in the query word (*statistical*) which are then mapped to spatial regions in the given example style images.

Experiments

Quantitative analysis of style imitation

	IV-S ↓	IV-U ↓	OOV-S ↓	OOV-U ↓
GANwriting [1]	120.07	124.30	125.87	130.68
Davis *et al* [2]	118.56	128.75	127.11	136.67
HWT (Ours)	**106.97**	**108.84**	**109.45**	**114.10**

Our HWT performs favorably in all four settings: In-Vocabulary words and seen style (**IV-S**), In Vocabulary words and unseen style (**IV-U**), Out of vocabulary content and seen style (**OOV-S**) and Out of vocabulary content and unseen style (**OOV-U**).

Quantitative analysis of Handwritten Text Generation

	FID ↓	GS ↓
ScrabbleGAN [3]	20.72	2.56×10^{-2}
Davis *et al* [2]	20.65	4.88×10^{-2}
HWT (Ours)	**19.40**	$\mathbf{1.01 \times 10^{-2}}$

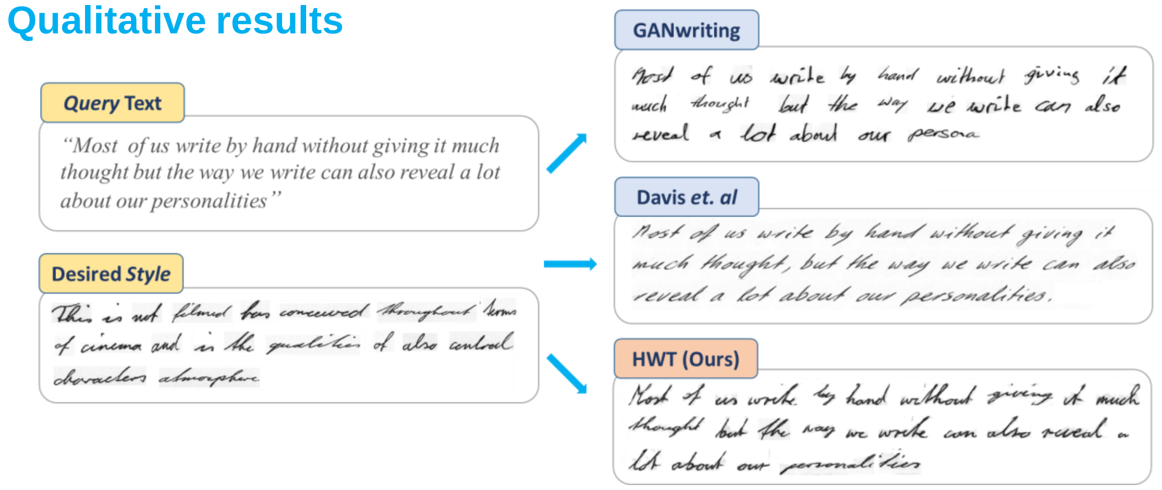
We evaluate the quality of the text image generated by our HWT following the same evaluation settings as used in ScrabbleGAN. Our HWT performs favorably against these methods in terms of both FID and GS score.

Handwritten Text Recognition (HTR)

Method	Training Data			CVL(%)		CVLoov(%)	
	GAN	CVL	IAM	WER	CER	WER	CER
—	✗	✓	✓	29.41	13.13	37.63	17.16
HiGAN [4]	✓	✓	✓	28.91	12.54	37.06	16.67
ScrabbleGAN [3]	✓	✓	✓	28.68	12.13	37.10	16.73
HWT (Ours)	✓	✓	✓	**27.81**	**11.84**	**36.47**	**15.95**

We utilize our generated samples for training HTR model to validate if the generated images can help improve text recognition performance.

Qualitative results


Query Text
"Most of us write by hand without giving it much thought but the way we write can also reveal a lot about our personalities"
Desired Style
This is not filmed has conceived throughout terms of cinema and is the qualities of also control characters atmosphere
GANwriting
Most of us write by hand without giving it much thought but the way we write can also reveal a lot about our persona
Davis et. al
Most of us write by hand without giving it much thought, but the way we write can also reveal a lot about our personalities.
HWT (Ours)
Most of us write by hand without giving it much thought but the way we write can also reveal a lot about our personalities

Conclusion

Qualitative, quantitative and human-based evaluations show that our HWT produces realistic styled handwritten text images with varying length and any desired writing style.

1 Kang *et.al*, Ganwriting: Content conditioned generation of styled handwritten word images. In ECCV, 2020.
2 Davis *et al*., Text and style conditioned gan for generation of offline handwriting lines. BMVC, 2020.
3 Fogel *et al*, Scrabblegan: semi-supervised varying length handwritten text generation. In CVPR, 2020.
4 Gan *et. al.* HiGAN: Handwriting Imitation Conditioned on Arbitrary-Length Texts and Disentangled Styles. In AAAI, 2021

DoodleFormer: Creative Sketch Drawing with Transformers

Ankan Kumar Bhunia[1] Salman Khan[1,2] Hisham Cholakkal[1] Rao Muhammad Anwer[1,4]
Fahad Shahbaz Khan[1,3] Jorma Laaksonen[4] Michael Felsberg[3]
[1]MBZUAI, UAE [2]Australian National University, Australia [3]Linköping University, Sweden [4]Aalto University, Finland

ECCV TEL AVIV 2022

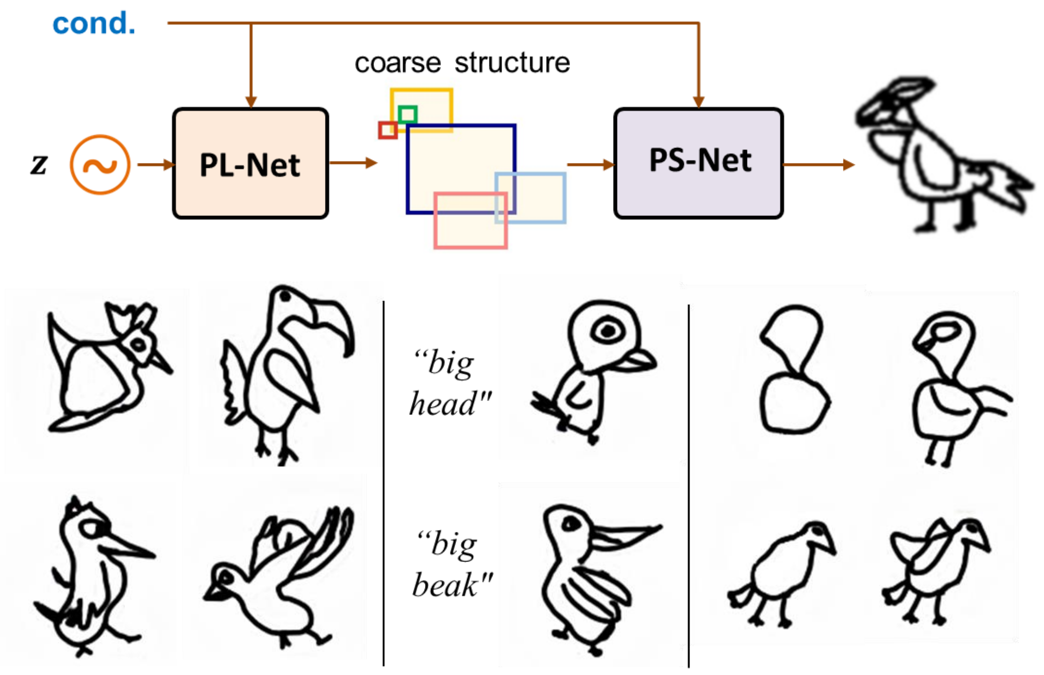
Motivation

Limitation of existing frameworks

- Poor Quality: The recent part-based method DoodlerGAN [1] doesn't employ an explicit mechanism to ensure that each body part is placed appropriately with respect to the rests. This leads to topological artifacts and connectivity issues.
- Lack of diversity: DoodlerGAN struggles to generate diverse sketch images, which is an especially desired property in creative sketch generation.

Why *Coarse-to-fine* Design?

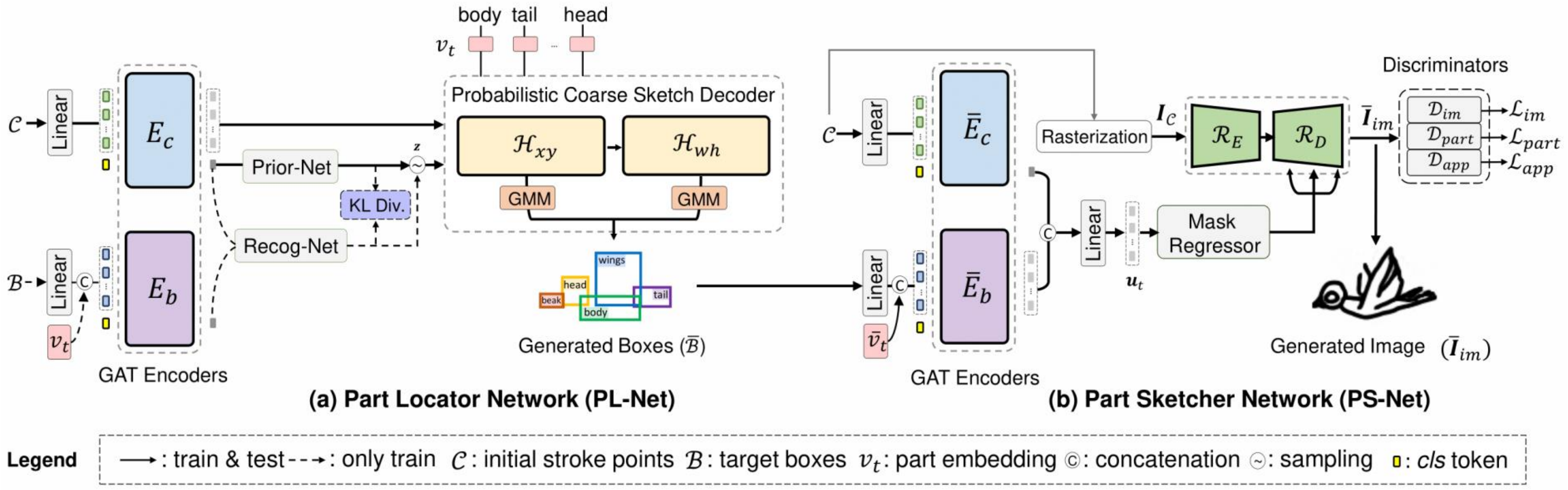
- Generally, a human artist *(i) first draws the holistic coarse structure* of the sketch and then *(ii) fills the fine-details* to generate the final sketch. By first drawing the holistic coarse structure of the sketch aids to appropriately decide the location and the size of each sketch body part to be drawn.



(a) cond. : initial strokes (b) cond. : text (c) cond. : partial sketch

- **1st stage**: PL-Net, takes the initial stroke points as the conditional input and learns to return the bounding boxes corresponding to each body part to be drawn;
- **2nd stage**: PS-Net, takes the predicted box locations as inputs and generates the final sketch image;

Methodology

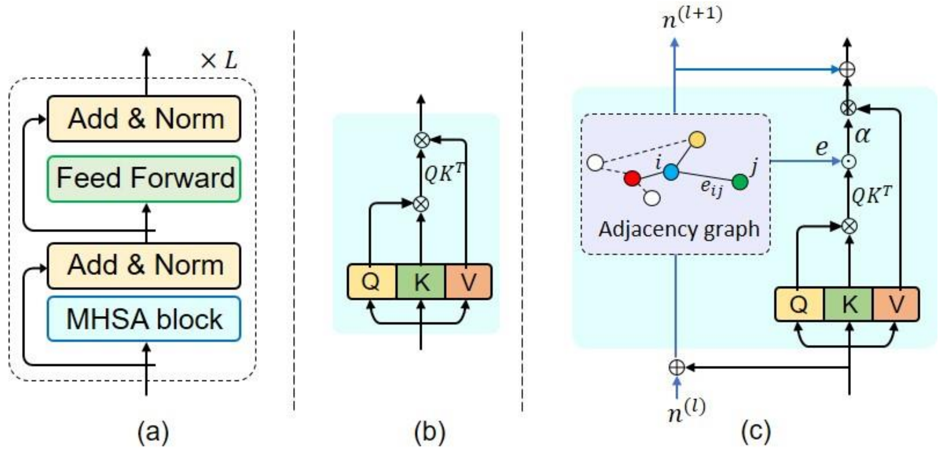


(a) Part Locator Network (PL-Net) (b) Part Sketcher Network (PS-Net)

Legend ⟶ : train & test - - → : only train $\mathcal{C}$: initial stroke points $\mathcal{B}$: target boxes v_t : part embedding ©: concatenation ~: sampling ▯: *cls* token

We propose a novel two-stage *transformer-based encoder-decoder* framework, DoodleFormer, for creative sketch generation. DoodleFormer decomposes the creative sketch generation problem into the construction of holistic coarse sketch composition followed by injecting fine details to generate final sketch image.

- GAT Encoder blocks: Our framework comprises of *graph-aware transformer* (GAT) block-based encoders to capture structural relationship between different regions within a sketch.



- While the standard self-attention module is effective towards learning highly contextualized feature representation, it does not explicitly emphasize on the *local structural relation*. However, creative sketches are structured inputs with definite connectivity patterns between sketch parts. To model this structure, we propose to encode an adjacency based graph implemented with spectral graph convolution.

- GMM-based probabilistic coarse sketch Decoder: we further introduce probabilistic coarse sketch decoders that utilize *GMM modelling for box prediction*. This enables our DoodleFormer to achieve diverse, yet plausible coarse structure for sketch generation.
- Different from the conventional box prediction that directly maps the decoder output features as deterministic box parameters, our GMM-based box prediction is modeled with M normal distributions

Loss Objectives: The PL-Net loss is the weighted sum of the *reconstruction loss*, and the *KL divergence* loss.

The training of PS-Net follows the standard GAN formulation where the PS-Net generator is followed by additional discriminator networks to obtain *image-level, part-level, and appearance adversarial losses*.

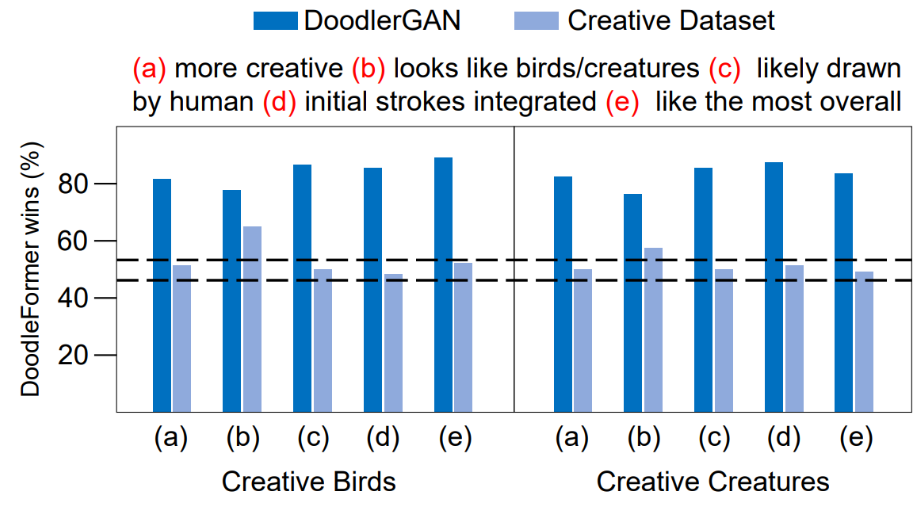
Experiments

Quantitative analysis of DoodleFormer

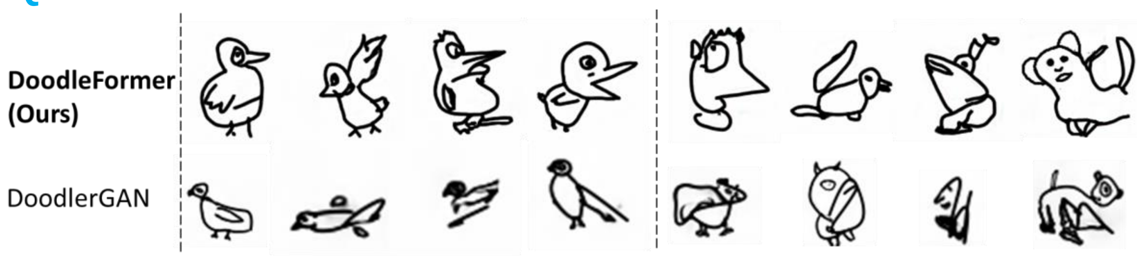
Methods	Creative Birds			Creative Creatures			
	FID(↓)	GD(↑)	CS(↑)	FID(↓)	GD(↑)	CS(↑)	SDS(↑)
Training Data	-	19.40	0.45	-	18.06	0.60	1.91
SketchRNN [12]	82.17	17.29	0.18	54.12	16.11	0.48	1.34
StyleGAN2 [17]	130.93	14.45	0.12	56.81	13.96	0.37	1.17
DoodlerGAN [10]	39.95	16.33	**0.69**	43.94	14.57	0.55	1.45
DoodleFormer (Ours)	**16.45**	**18.33**	0.55	**18.71**	**16.89**	**0.56**	**1.78**

User study analysis

Higher values indicate DoodleFormer is preferred more often over the compared approaches (DoodlerGAN [1] and human drawn datasets).



Qualitative results



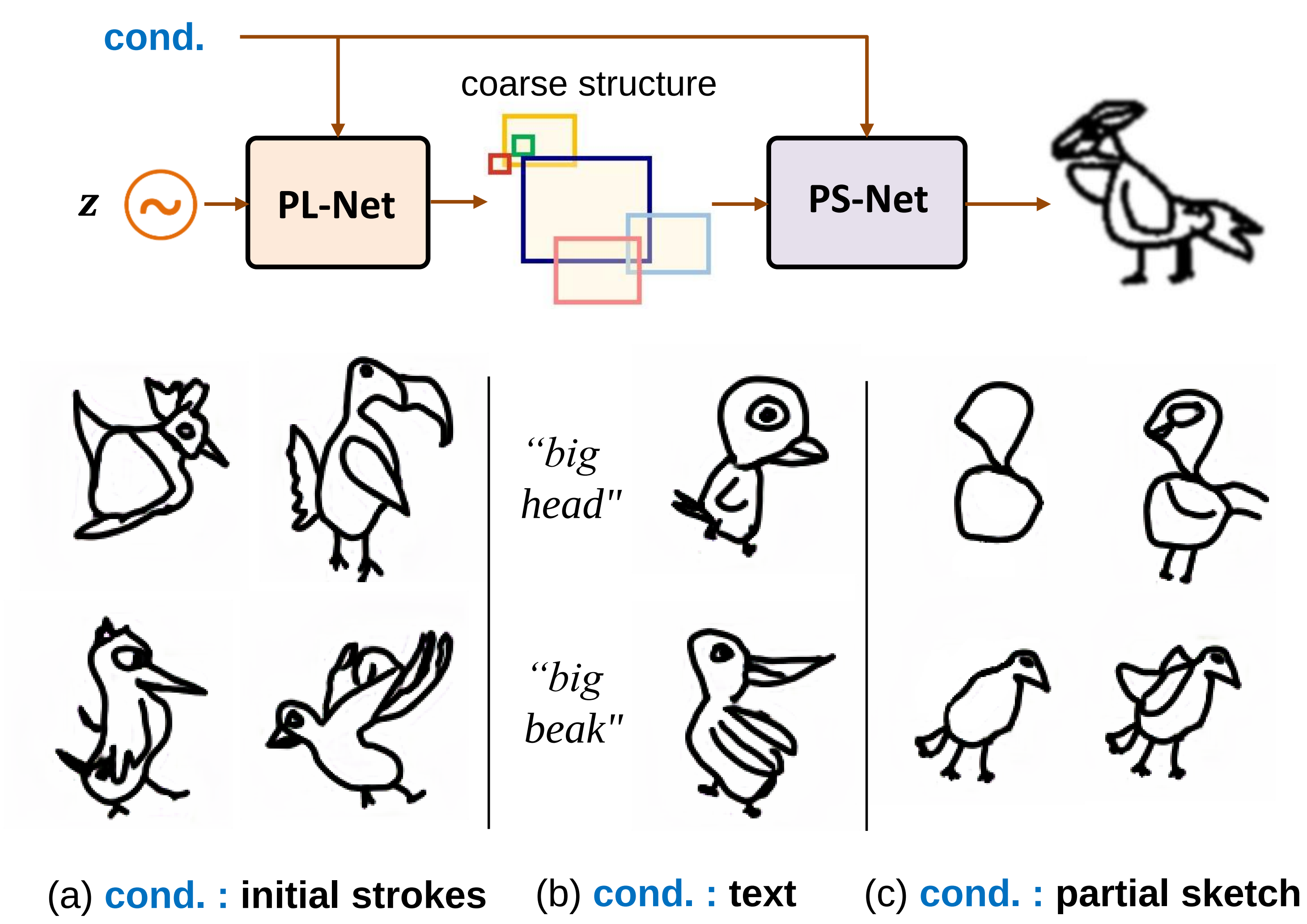
For additional results for different applications (Text-to-sketch, Sketch completion an House plan generation) of Doodlerformer see our paper.

Conclusions

Qualitative, quantitative and human-based evaluations show that our DoodlerFormer produces diverse, yet realistic creative sketches.

Scan code for project page

[1] Ge, Songwei, et al. "Creative sketch generation." arXiv preprint arXiv:2011.10039 (2020).



(a) cond. : initial strokes (b) cond. : text (c) cond. : partial sketch

(a) Creative Sketch Generation

(b) Text to Creative Sketch Generation

(c) Creative Sketch Completion