StyleBabel: Artistic Style Tagging and Captioning

[1]Dan Ruta, [1]Andrew Gilbert, [2]Pranav Aggarwal, [2]Naveen Marri, [2]Ajinkya Kale, [3]Jo Briggs, [4]Chris Speed, [2]Hailin Jin, [2]Baldo Faieta, [2]Alex Filipkowski, [2]Zhe Lin, [1,2]John Collomosse

[1]CVSSP University of Surrey, [2]Adobe Research, [3]University of Northumbria, [4]University of Edinburgh

Background:

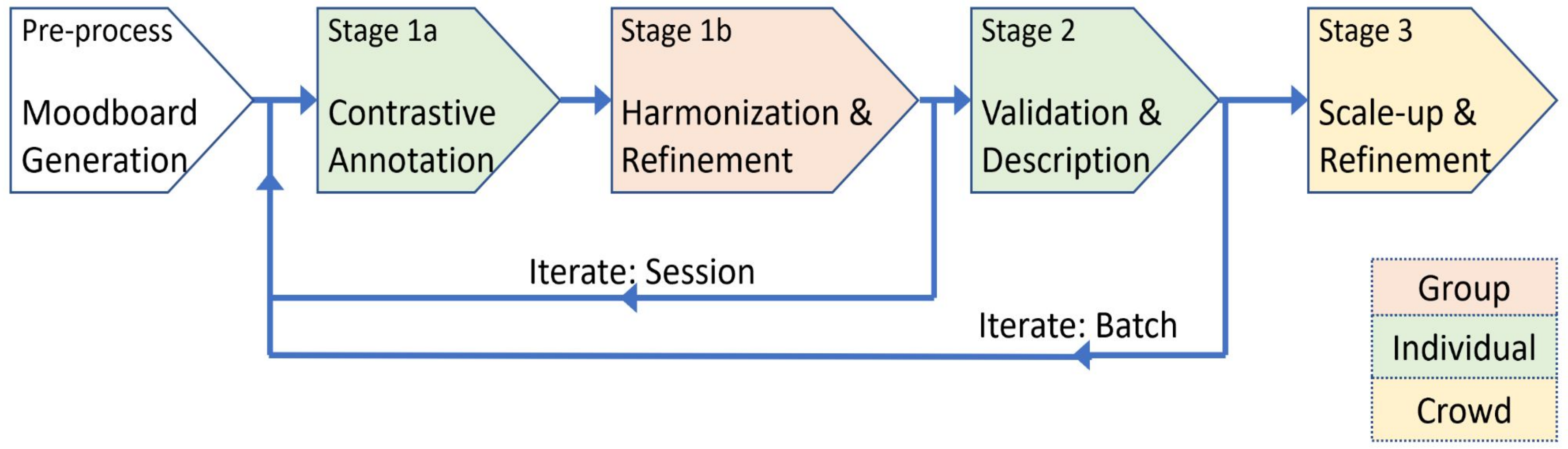
We set out to create a large scale multi-modal dataset, annotating a wide variety of artistic style, with both tags and natural language captions.

We present the first such dataset, covering a much wider variety of artistic styles compared to existing works, and we focus very specifically on the visual attributes of style, ignoring context, emotions, and meanings.

We present several experiments showcasing uses of our dataset, and we push SOTA on style representation, by expanding ALADIN with ViT.

Contribution #1 - Grounded Annotation Methodology:

- Iterative, collaborative labelling methodology based on *Grounded Theory* (GT)
  - Qualitative methodology often used in humanities and social sciences
  - Unconstrained, multi-stage data clustering exercises
  - Simultaneously evolving shared vocabulary
  - Undertaken by experts at art universities, with open discussions
  - Final cleaning stage undertaken by trained crowd annotation team

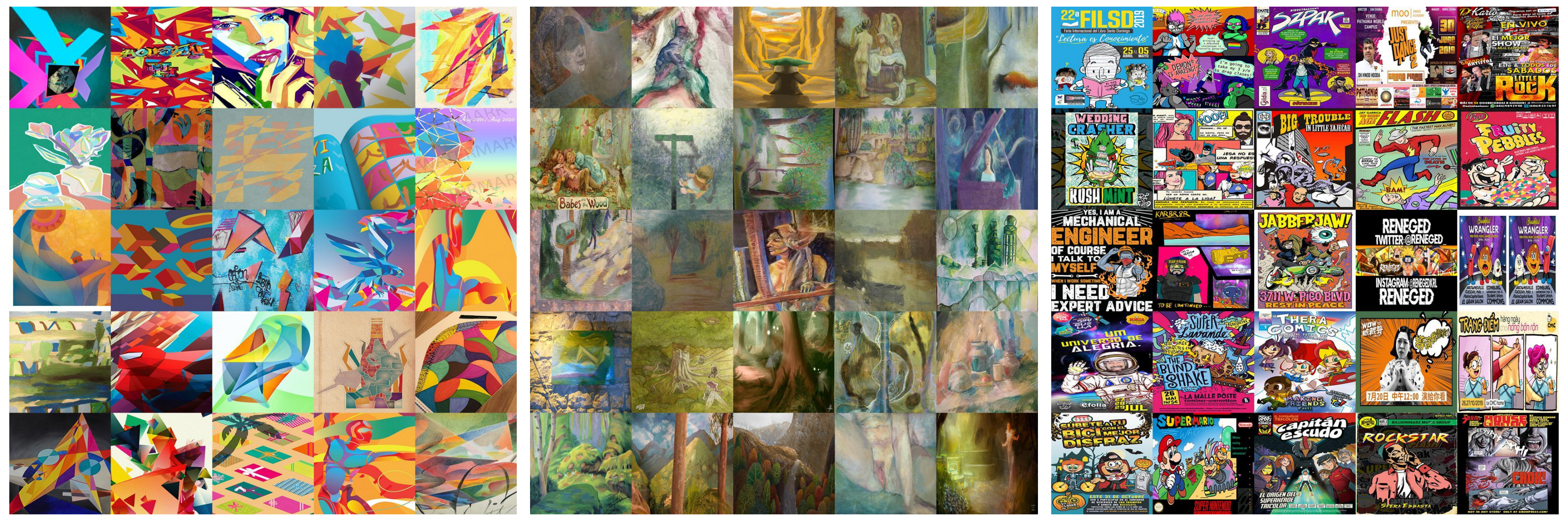


[

Changed (Removed or added tags in Stage 1b)	line, both have the same white backgrounds with very little tonal variations/ just block colours, no difference in tones, bright white background, b+w, constructed by lines, no colour, hand drawing linear, pen and ink	b+w black and white, pen and ink, fantasy, intricate, white space, central composition, linear, illustration
Final tags after Stage 1b cleaning	ink work, sketches, bright, black and white, white background, clean, monotonal, drawing, simple, illustration, linear, pen and ink	graphic, drawing, black and white, pen and ink, fantasy, intricate, white space, central composition, linear, illustration

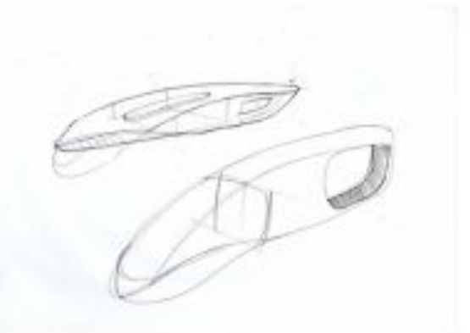
Data:
https://cvssp.org/projects/danruta/stylebabel.csv

Contribution #2 - StyleBabel dataset:



Example style groups from the StyleBabel dataset annotation

- First diverse, large scale dataset for natural language captions and free-form tags, for diverse artistic style
  - 135k digital artworks
  - Images collected from the Behance platform
  - Both tags and caption labels collected via GT
  - Labels are individualised to each image

Image				
Tags	dim, concept, action, fantasy, powerful, digital, photography, animated, prototype, masculine, detailed, professional, lighting	abstract, moody, portrait, oil, painting, drawing, artistic, melancholic, pleasing	cold, digital, book, bright, colors, drawing, child, stroke, busy, clear, illustration, festive, blue	experimental, analog, line, development, black, drawing, sketch, figure, commercial, white, scamp, stroke, product, pencil, rough, thin, isometric
Caption	Fantasy themed digital illustration featuring an animated male character, dim highlighting and a hazy, dark and cluttered background. The illustration highlights the powerful masculine character with sharp objects around.	Portrait oil painting of a female character featuring abstract shapes and psychedelic patterns against a dark background. The artistic artwork is melancholic and using thin repetitive strokes and shades.	Digital bright fantasy anthropomorphism cartoon illustration created with soft diffused blended hues, brush strokes, lines, and geometric forms in neutral and cool tones.	Analog experimental sketches with thin pencil strokes and lines. The isometric drawing expresses commercial product development.

- Excerpt of the dataset - four images and their associated tags and natural language caption labels
- Estimated total cost of annotation: $160k
- Dataset is released freely as CC-BY 4.0

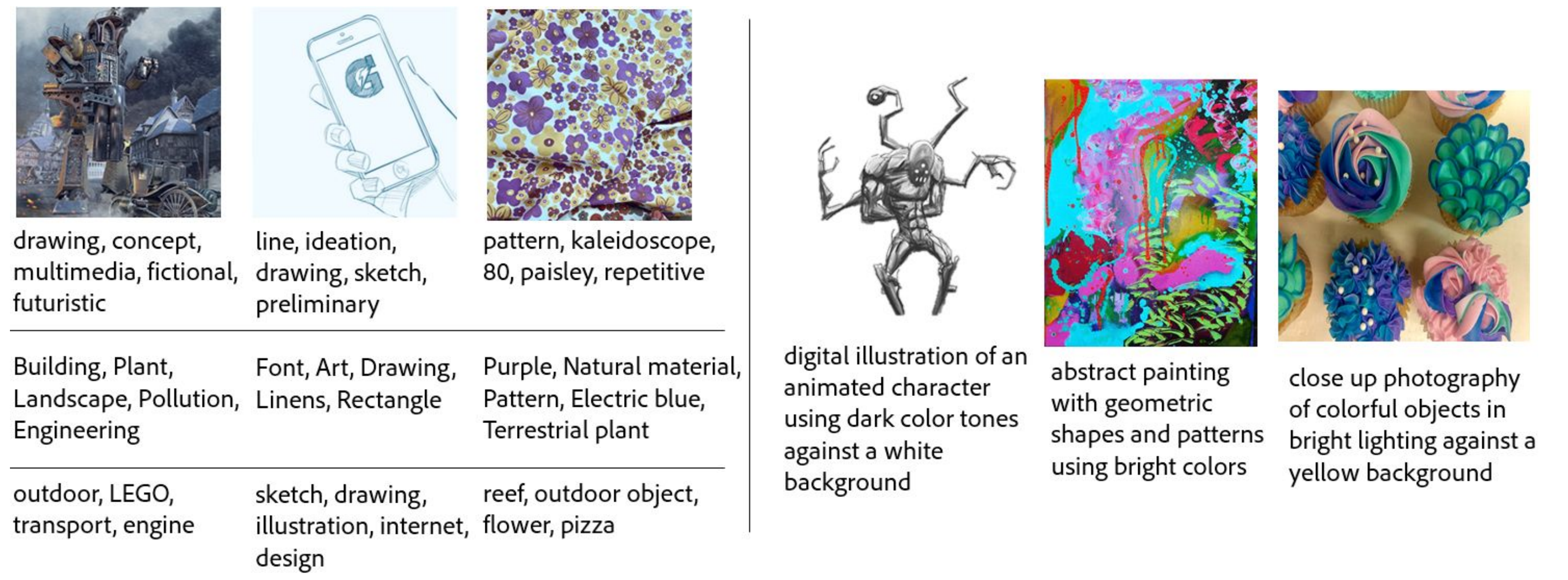
TAGS: contrast, kaleidoscope, psychedelic, complex, intricate, complicated, paisley,colored, colorful,repetitive, surreal

CAPTION: kaleidoscopic illustration featuring repetitive psychedelic pattern with abstract shapes against a plain background. The colorful artwork is complex and intricate.

TAGS: abstract, chaotic, complex, texture, colors, pastel, sparse, scale, large, colorful, expressionism

CAPTION: Abstract oil painting featuring chaotic and complex linear patterns. This expressionism also features sparse textures using pastel and muted colors.

Contribution #3 - Multi-modal experiments:


drawing, concept, multimedia, fictional, futuristic
line, ideation, drawing, sketch, preliminary
pattern, kaleidoscope, 80, paisley, repetitive
Building, Plant, Landscape, Pollution, Engineering
Font, Art, Drawing, Linens, Rectangle
Purple, Natural material, Pattern, Electric blue, Terrestrial plant
outdoor, LEGO, transport, engine
sketch, drawing, illustration, internet, design
reef, outdoor object, flower, pizza
digital illustration of an animated character using dark color tones against a white background
abstract painting with geometric shapes and patterns using bright colors
close up photography of colorful objects in bright lighting against a yellow background

Example tags and captions generated automatically by models trained with StyleBabel

- Automatic tag generation (img -> tag)
  - We beat SOTA CLIP model on tag retrieval when using the cleanest level of our StyleBabel dataset
  - We further push SOTA with our improvements to the ALADIN style model, where we integrate ViT into its model backbone

Data	Model	WordNet score
CLIP Webscale	CLIP [36] baseline	0.168
StyleBabel-mturk	ALADIN-ViT	0.164
StyleBabel (coarse)	CLIP [36]	0.187
StyleBabel (coarse)	ALADIN-ViT	0.225
StyleBabel (FG)	CLIP	0.215
StyleBabel (FG)	ALADIN-ViT	**0.352**

- Tag-based image retrieval (tag->img)
  - Inverted tag retrieval experiment, where images are retrieved
  - Both experiments are restricted to style information only

Hyper-realism
vivid

- Automatic caption generation (img -> caption)
  - Greatly exceed SOTA for style information in captions
  - Built using a Virtex+AttentionOnAttention model
  - Focused on global style features, rather than localized semantic features, as standard in captioning literature
  - Tables with examples can be found in the supplementary materials

Data	Model	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	Rouge-L	CIDEr
MS-COCO baseline	VirTex	0.162	0.053	0.016	0.005	0.037	0.145	0.022
StyleBabel (CL)	VirTex	0.127	0.049	0.022	0.010	0.054	0.135	0.076
StyleBabel (IL)	VirTex	0.331	0.187	0.113	0.071	0.129	0.288	0.350
StyleBabel (CL+IL)	VirTex	**0.335**	**0.189**	**0.118**	**0.078**	**0.131**	**0.288**	**0.372**
StyleBabel (CL+IL)	ResNet LSTM	0.087	0.021	0.008	0.002	0.033	0.080	0.017
StyleBabel (CL+IL)	ALADIN-ViT LSTM	0.094	0.030	0.013	0.006	0.042	0.089	0.034
VirTex	Artemis	0.185	0.083	0.041	0.023	0.081	0.182	0.146
VirTex	Artemis (SB)	0.120	0.031	0.013	0.005	0.034	0.108	0.029