Language-Driven Artistic Style Transfer

Tsu-Jui Fu[1] Xin Wang[2] William Wang[1]

[1]UC Santa Barbara, [2]UC Santa Cruz

https://tsujuifu.github.io

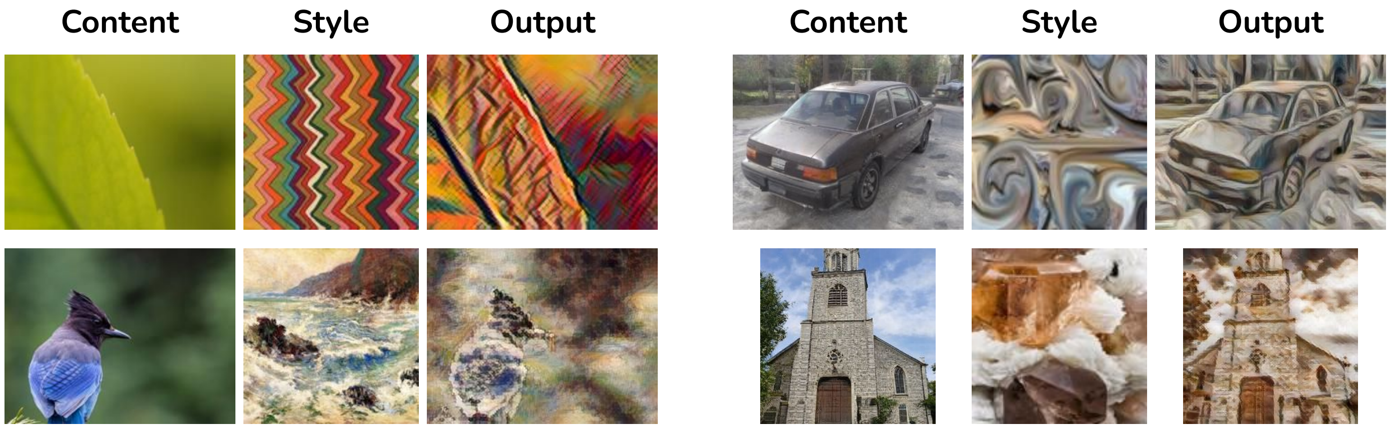
Artistic Style Transfer

- Render a photograph with an **arbitrary artwork style**
    - Preserve **content structures** yet present **style patterns**
- Content ($\mathcal{C}$) + Style ($\mathcal{S}$) $\rightarrow$ Stylized Output ($\mathcal{O}$)

Content **Style** **Output** **Content** **Style** **Output**

Artistic Style Transfer

- Render a photograph with an **arbitrary artwork style**
  - Preserve **content structures** yet present **style patterns**
- Content ($\mathcal{C}$) + Style ($\mathcal{S}$) $\rightarrow$ Stylized Output ($\mathcal{O}$)



- **Prepare collections of style image** in advance
- **Redraw new references first** if there is no expected style

Language-Driven Artistic Style Transfer (LDAST)

- **Language** is the most natural way for humans to communicate
    - **Follow textual descriptions** to perform style transfer
    - Improve **accessibility** and **controllability**
- Content ($\mathcal{C}$) + Instruction ($\mathcal{X}$) → Stylized Output ($\mathcal{O}$)

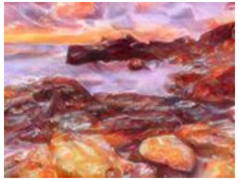
out on a lovely day with the water, sketching, and painting

i feel chaotic and confused due to the black and gray tones

peaceful green colors and shading of the branches, feel content

Content

reflective, orange, purple, and red bubble



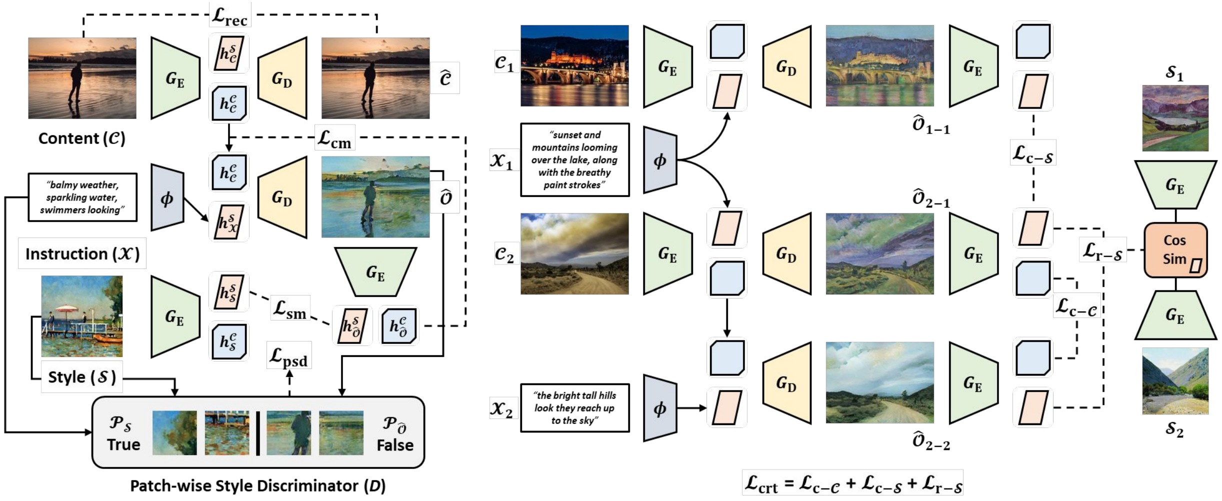
salt deposits forming around brown golden frosted crystal

large blackouts on rough off white, jute cotton surface

Contrastive Language Visual Artist (CLVA)

- For **training**, there are content images ($\mathcal{C}$), style images ($\mathcal{S}$), and instructions ($\mathcal{X}$)
- During **inference**, **only $\mathcal{C}$ and $\mathcal{X}$** are provided

- Learn the **latent style patterns** from the instruction
- Further compare contrastive pairs of **relative $\mathcal{C}$ and $\mathcal{X}$**



Language Visual Artist (LVA) Contrastive Reasoning (CR)

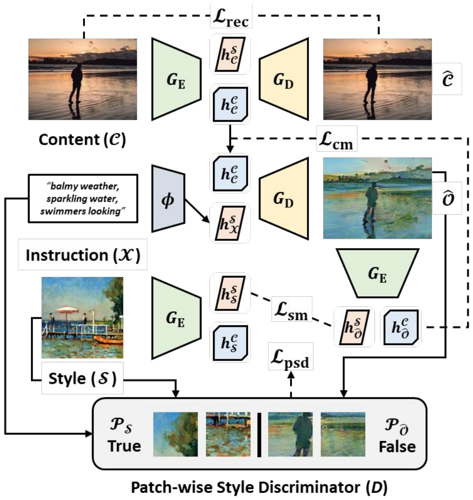
Language Visual Artist (LVA)

- Visual Encoder (**G_E**), Text Encoder (**Φ**), and Visual Decoder (**G_D**)
  - Extract content feature (**$h^{\mathcal{C}}$**), style feature (**$h^{\mathcal{S}}$**), and instruction feature (**$h^{\mathcal{X}}$**)
  - Compose **$h^{\mathcal{C}}$** and **$h^{\mathcal{X}}$** / **$h^{\mathcal{S}}$** to produce the stylized result

- **Structure Reconstruction ($\mathcal{L}_{\text{rec}}$)**
  - **Reproduce $\mathcal{C}$** from the original content style
- **Patch-wise Style Discrimination ($\mathcal{L}_{\text{psd}}$)**
  - ***D* distinguishes the patch ($\mathcal{P}$)** is from $\mathcal{S}$ or $\hat{\mathcal{O}}$
  - Optimize **G_E, Φ**, and **G_D** to **fool *D***
- **Content Matching ($\mathcal{L}_{\text{cm}}$)** and **Style Matching ($\mathcal{L}_{\text{sm}}$)**
  - Further **enhance the alignment** with the input

$$\mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{psd}} = ||\hat{\mathcal{C}} - \mathcal{C}||_2, \log(1 - D(\mathcal{P}_{\hat{\mathcal{O}}}, \mathcal{X})) + \log(D(\mathcal{P}_{\mathcal{S}}, \mathcal{X}))$$

$$\mathcal{L}_{\text{cm}}, \mathcal{L}_{\text{sm}} = ||h^{\mathcal{C}}_{\hat{\mathcal{O}}} - h^{\mathcal{C}}_{\mathcal{C}}||_2, ||h^{\mathcal{S}}_{\hat{\mathcal{O}}} - h^{\mathcal{S}}_{\mathcal{S}}||_2$$



Language Visual Artist (LVA)

Contrastive Reasoning (CR)

- Compare transferred results of **contrastive pairs** ($\{\mathcal{C}_1, \mathcal{X}_1, \mathcal{S}_1\}$ and $\{\mathcal{C}_2, \mathcal{X}_2, \mathcal{S}_2\}$)
  - Transfer to **various styles** while preserving the **same structure**
  - Apply **analogous style patterns** from **related style instructions**

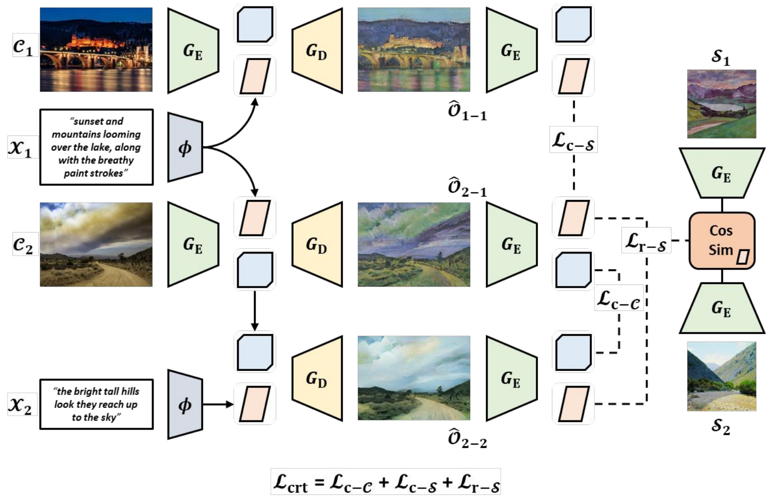
- **Consistent Matching ($\mathcal{L}_c$)**
  - Similar content structure from $\mathcal{C}_2$
  - Similar style patterns from $\mathcal{X}_1$
- **Relative Matching ($\mathcal{L}_r$)**
  - Relative style patterns from $\mathcal{X}_1$ **and** $\mathcal{X}_2$

$$\mathcal{L}_{\text{c}-\mathcal{C}} = ||h^{\mathcal{C}}_{\hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_1}} - h^{\mathcal{C}}_{\hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_2}}||_2 + ||h^{\mathcal{C}}_{\hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_1}} - h^{\mathcal{C}}_{\hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_2}}||_2$$

$$\mathcal{L}_{\text{c}-\mathcal{S}} = ||h^{\mathcal{S}}_{\hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_1}} - h^{\mathcal{S}}_{\hat{\mathcal{S}}_{2-1}}||_2 + ||h^{\mathcal{S}}_{\hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_2}} - h^{\mathcal{S}}_{\hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_2}}||_2$$

$$\begin{aligned}\mathcal{L}_{\text{r}-\mathcal{S}} = (&||h^{\mathcal{S}}_{\hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_1}} - h^{\mathcal{S}}_{\hat{\mathcal{O}}_{\mathcal{C}_1\text{-}\mathcal{X}_2}}||_2 + \\ &||h^{\mathcal{S}}_{\hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_1}} - h^{\mathcal{S}}_{\hat{\mathcal{O}}_{\mathcal{C}_2\text{-}\mathcal{X}_2}}||_2) \cdot r\end{aligned}$$

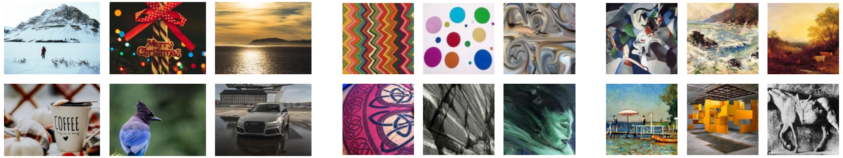
$$\mathcal{L}_{\text{ctr}} = \mathcal{L}_{\text{c}-\mathcal{C}} + \mathcal{L}_{\text{c}-\mathcal{S}} + \mathcal{L}_{\text{r}-\mathcal{S}}$$



Contrastive Reasoning (CR)

Experimental Setup

- **Datasets**
  - **Content**: Wallpaper
  - **Style**: DTD^2 / ArtEmis

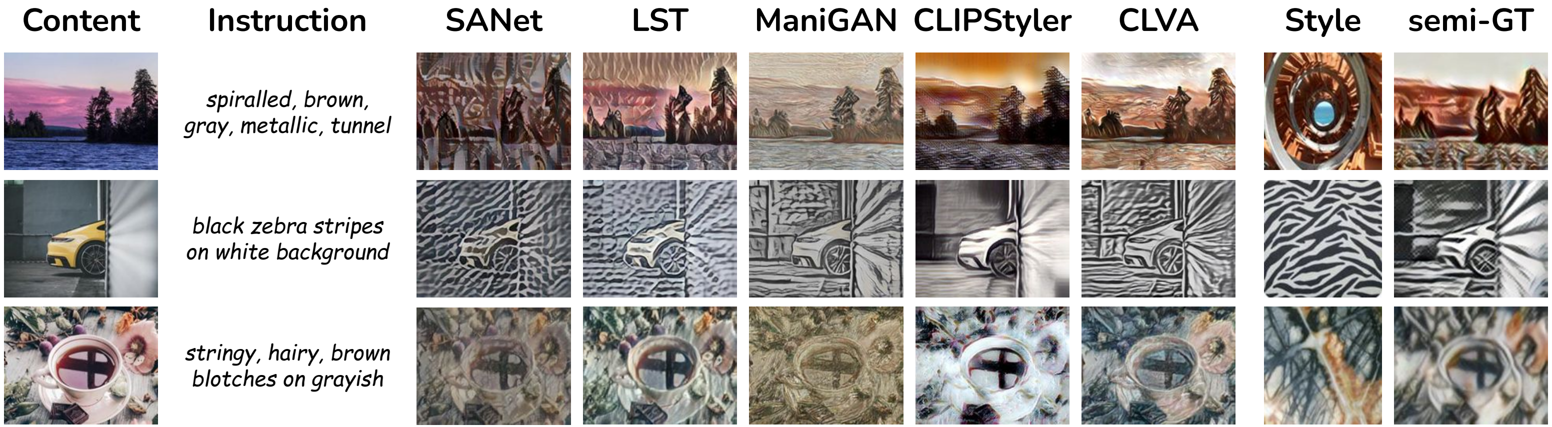


- **Evaluation Metrics (semi-GT** from AdaAttN**)**
  - **Percept (↓)**: distance of **gram matrix** from visual features (*vs.* style image)
  - **FAD (↓)**: distance of **InceptionV3 features** (*vs.* semi-GT)
  - **VLS (↑)**: **relative visual-text similarity** from CLIP (*vs.* semi-GT | instruction)

- **Baselines**
  - **Style Transfer**: SANet / LST
  - **Language-based Image Editing**: ManiGAN
  - **CLIP-based Optimization**: StyleCLIP / NADA / CLIPStyler

Instruction with Visual Attributes (DTD^2)

	Automatic Metrics			Human Evaluation			
Method	Percept ↓	FAD ↓	VLS ↑	Content ↑	Instruction ↑	Style ↑	semi-GT ↑
SANet	0.2129	0.1627	23.57	2.701	2.477	2.738	2.630
LST	0.2129	0.1533	23.16	2.743	2.831	2.651	2.528
ManiGAN	0.2401	0.1663	23.25	2.757	2.562	2.937	2.922
CLIPStyler	0.2598	0.1818	**24.62**	2.948	3.388	3.073	3.265
CLVA	**0.2033**	**0.1493**	24.00	**3.852**	**3.742**	**3.603**	**3.655**



Instruction with Emotional Effects (ArtEmis)

	Automatic Metrics			Human Evaluation			
Method	Percept ↓	FAD ↓	VLS ↑	Content ↑	Instruction ↑	Style ↑	semi-GT ↑
SANet	**0.0352**	<u>0.1548</u>	19.30	<u>3.170</u>	2.978	2.980	2.890
LST	0.0386	0.1595	19.92	2.967	2.714	2.614	2.757
ManiGAN	0.0500	0.1554	19.69	2.729	2.583	2.879	<u>3.192</u>
CLIPStyler	0.0659	0.1759	**21.04**	2.777	<u>3.140</u>	<u>2.998</u>	2.952
CLVA	<u>0.0357</u>	**0.1418**	<u>20.11</u>	**3.357**	**3.586**	**3.530**	**3.208**

Content	Instruction	SANet	LST	ManiGAN	CLIPStyler	CLVA	Style	semi-GT
	*this creature looks like something out of a nightmare, with the dark portrayed*							
	*sunset and mountains looming over the lake, along with the breathy paint strokes*							
	*charmed by the beautiful bright day, and the person at the side of the pale water*							

Specific Content Domain (Car & Church)

	Automatic Metrics			Human Evaluation			
Method	Percept ↓	FAD ↓	VLS ↑	Content ↑	Instruction ↑	Style ↑	semi-GT ↑
ManiGAN	0.2329	0.1672	23.44	2.861	2.894	2.978	2.893
StyleCLIP	0.2609	0.1812	21.55	**3.459**	2.845	2.930	2.829
NADA	0.2733	0.1876	23.38	2.542	2.798	2.846	2.932
CLIPStyler	0.2493	0.1826	**24.16**	2.986	3.067	3.003	3.032
CLVA	**0.1957**	**0.1544**	23.68	3.153	**3.465**	**3.344**	**3.315**

Ablation Study

- Reconstruction ($\mathcal{L}_{rec}$) + Patch-wise style ($\mathcal{L}_{psd}$) makes promising LDAST
- Content matching ($\mathcal{L}_{cm}$) helps the **structure similarity**
- Style matching ($\mathcal{L}_{sm}$) aims at **analogous style patterns**
- Contrastive reasoning ($\mathcal{L}_{ctr}$) leads to a **comprehensive improvement**

$\mathcal{L}_{rec}$+$\mathcal{L}_{psd}$	$\mathcal{L}_{cm}$	$\mathcal{L}_{sm}$	$\mathcal{L}_{ctr}$	Percept ↓	FAD ↓	VLS ↑
✔	✘	✘	✘	0.2290	0.1568	23.29
✔	✔	✘	✘	0.2304	0.1512	23.27
✔	✘	✔	✘	<u>0.2049</u>	0.1508	<u>23.69</u>
✔	✔	✔	✘	0.2100	<u>0.1499</u>	23.54
✔	✔	✔	✔	**0.2033**	**0.1493**	**24.00**

Why CLVA is better than CLIP-based?

- Investigate via **instruction-to-style retrieval**
  - CLIP cannot capture **detailed patterns** well

	DTD[2]		ArtEmis	
Method	**R@1**	**R@5**	**R@1**	**R@5**
CLIP	13.9	30.7	9.8	20.7
CLVA	**19.3**	**45.1**	**13.9**	**30.7**

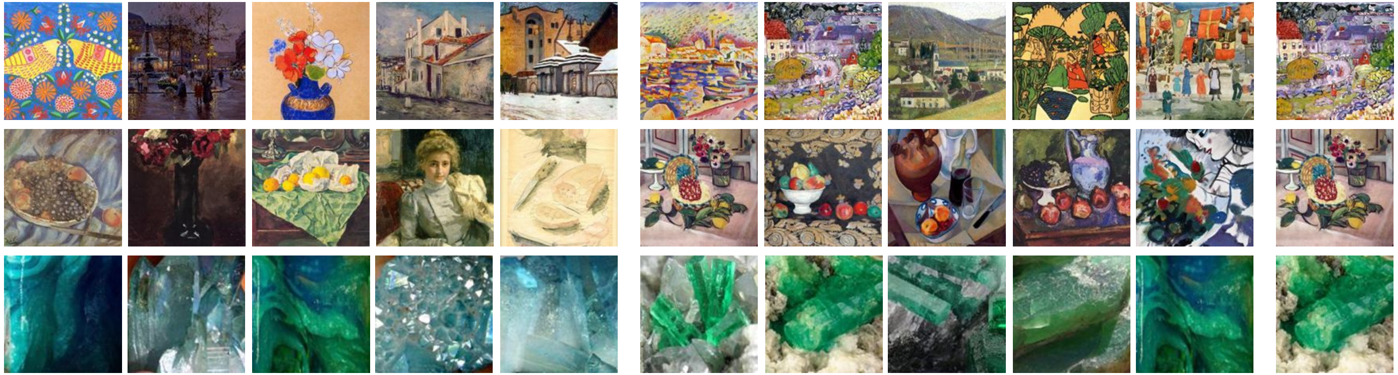
	Human Evaluation			
Method	**Content ↑**	**Instruction ↑**	**Style ↑**	**semi-GT ↑**
CLIPStyler (ft.)	1.208	1.347	1.292	1.333
CLVA	**1.792**	**1.653**	**1.708**	**1.667**

Instruction | **CLIP** | **CLVA** | **GT**

all of the bright colors in the town makes it a happy place to live

lovely still life that looks like a tropical table setting

light green shiny embedded in a white rough and raised surface



Efficiency

- Evaluate on a **single TITAN X** (12GB) with content image size **256x192**
  - CLIP-based methods require **numerous iterations for optimization**
  - CLVA further takes advantage of **parallelization**

	Time (sec ↓)			GPU (MB ↓)		
Method	**BS=1**	**32**	**50**	**BS=1**	**32**	**50**
ManiGAN	0.079	0.533	1.148	3,312	6,572	8,129
StyleCLIP	32.38	*	*	4,149	*	*
NADA	63.49	*	*	6,413	*	*
CLIPStyler	99.98	*	*	5,429	*	*
CLVA	**0.029**	**0.246**	**0.405**	**1,525**	**3,207**	**4,441**

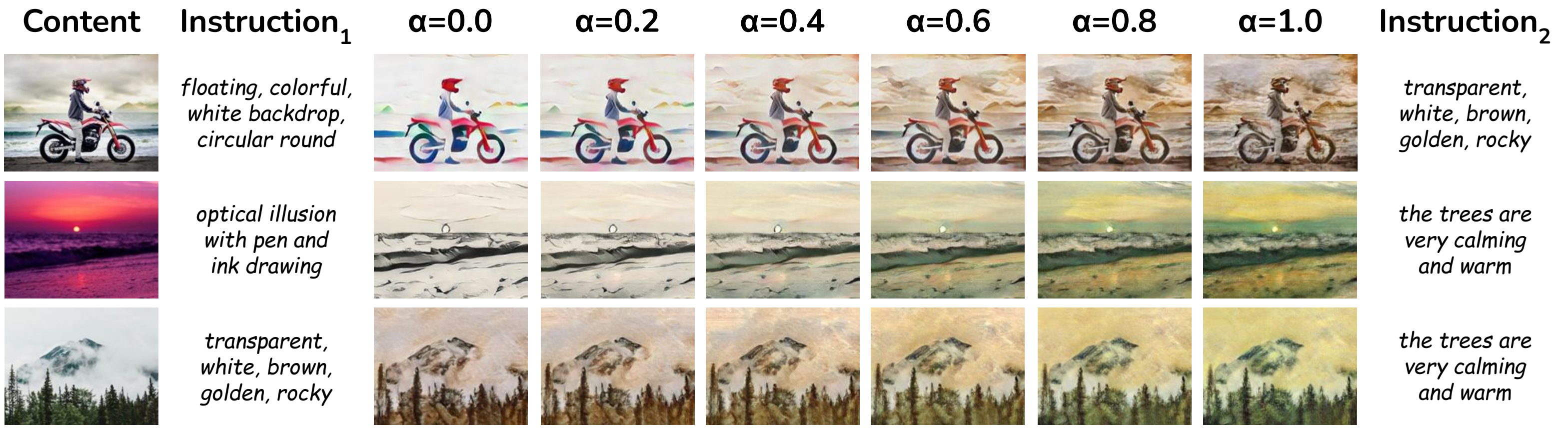
***** means this method can only **run one input at a time**

Linear Interpolation

- Consider two instructions $\mathcal{X}_1$ and $\mathcal{X}_2$
  - The **interpolated style feature** should be

$$h_{\mathrm{p}}^{\mathcal{S}} = (1-\alpha)h_{\mathcal{X}_1}^{\mathcal{S}} + \alpha h_{\mathcal{X}_2}^{\mathcal{S}}$$

  - Present a **smooth transformation** in between



Fine-grained Control

- Achieve fine-grained style control by **partial semantic editing**
  - The extracted patterns are **explicit** to reflect **each aspect of style semantic**

Super Resolution (2560x1440)

- **Borrow from SANet**, which supports content images with **any resolutions**

crystals like a krypton diamond in white and green

wrinkled leather, gray shiny surface

painted, rubbed, smeared yellow, green, blue and red

warm painting feels like the sun is setting behind

Conclusion

- Language-driven artistic style transfer (**LDAST**)
  - **Control artistic style transfer** via natural language
- Contrastive language visual artist (**CLVA**)
  - Learn to **extract explicit visual semantics** from style descriptions
  - Carry out instructions with **visual attributes / emotional effects**

Project

Code