JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining

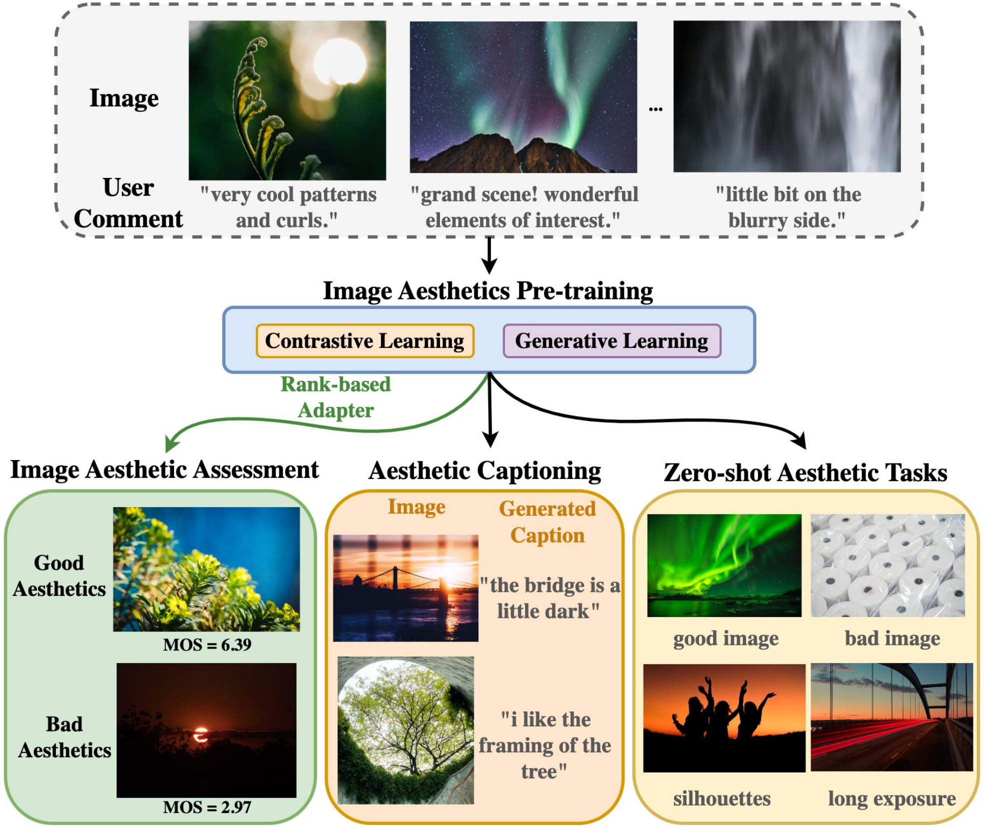
Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, Feng Yang

Google Research

WED-AM-173

VILA: **VI**sion **L**anguage **A**esthetics Learning Framework

- Pretrain an image aesthetic model with noisy image-comment pairs
- Efficiently adapt the model for downstream IAA tasks
    - Tunes only 0.1% params



Google

Motivation: Score-based IAA is Limited

- Image Aesthetic Assessment (IAA) methods are based on human ratings, but a single score does not capture the diverse aesthetic factors
  - E.g. composition, color, style, high-level semantics

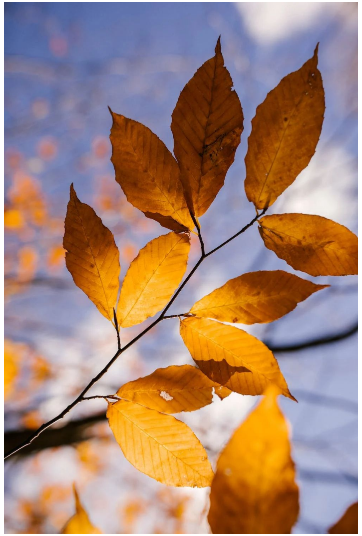
4.70

5.67

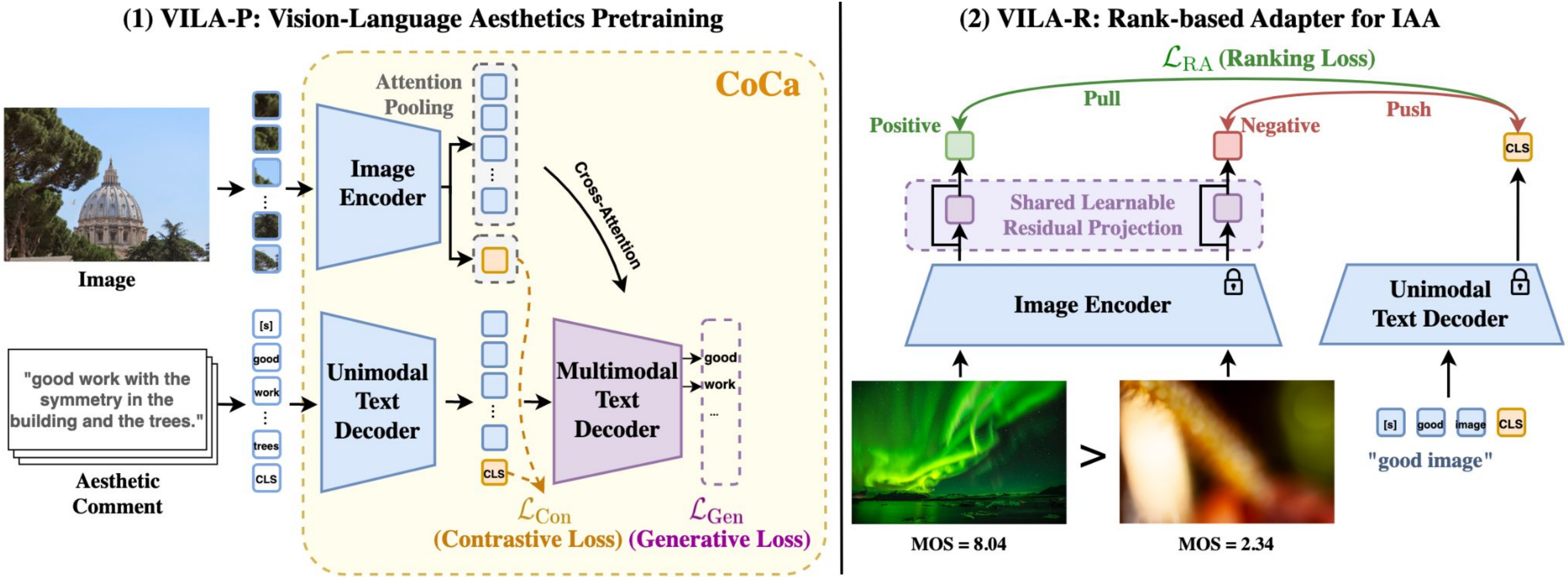
6.66

Proprietary + Confidential

Motivation: User Comments Provide Rich Aesthetic Semantics

Image			
User Comment	"there's a bit too much of the frame, and therefore not enough of the background here, imo"	"simple and nice composition, i like it"	"the idea is good here but the photo is too blurry."

VILA: Pretrain + Adapting



Overview of VILA

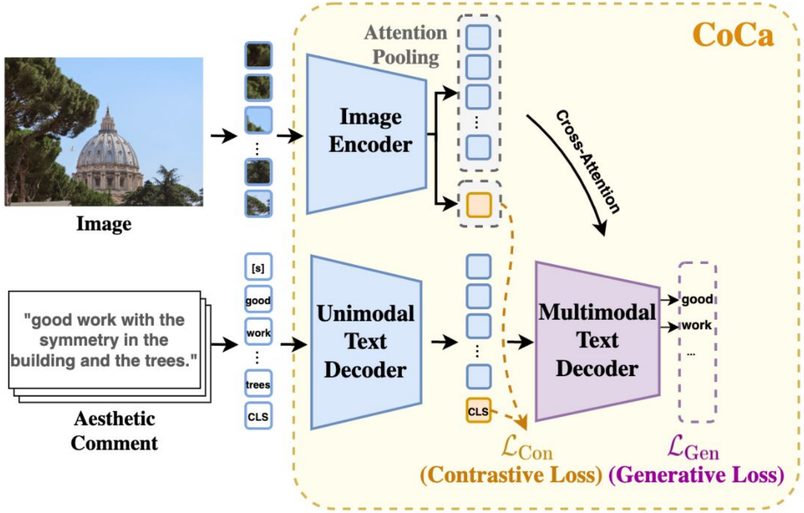
Proprietary + Confidential

VILA-P: Pretraining using Image-Comment Pairs

1. **General pretraining** with a filtered 650M subset of LAION-5B-EN

2. **Aesthetic pretraining** with 250K Image-Comment pairs from AVA-Captions, which is crawled from a professional photograph sharing website

(1) VILA-P: Vision-Language Aesthetics Pretraining



VILA-P: Experiment Results

- SOTA on image aesthetics captioning over AVA-Captions

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr
CWS [11]	**0.535**	0.282	0.150	0.074	0.254	0.059
Yeo *et al*. [58]	0.464	0.238	0.122	0.063	**0.262**	0.051
VILA	0.503	**0.288**	**0.170**	**0.113**	**0.262**	**0.076**

Table 5. Results on AVA-Captions dataset.

"pretty colors. the bright flowers on the trees add interest to anything."

"cute kitty is the best pose for this picture."

"color, focus and saturation are good. the image seems a little dark."

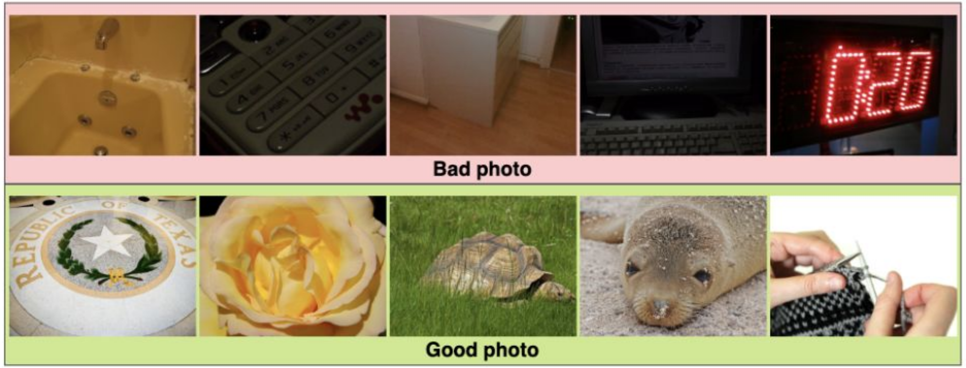
"maybe could have cropped a bit more on top of the birches."

"great perspective and colors in this shot. love the beautiful sky ."

"lovely shooting with excellent colour, great composition."

VILA-P: Experiment Results

- ZSL for Image Aesthetic Assessment



Top-5 Retrieved Images

	Prompts	
	$\boldsymbol{p}_g$	$\boldsymbol{p}_b$
Single Prompt	*"good image"*	*"bad image"*
Ensemble of Prompts	*"good image"* *"good lighting"* *"good content"* *"good background"* *"good foreground"* *"good composition"*	*"bad image"* *"bad lighting"* *"bad content"* *"bad background"* *"bad foreground"* *"bad composition"*

$$r = \frac{e^{\boldsymbol{v}^\top \boldsymbol{p}_g}}{e^{\boldsymbol{v}^\top \boldsymbol{p}_g} + e^{\boldsymbol{v}^\top \boldsymbol{p}_b}}$$

VILA-P: Experiment Results

- ZSL for Image Aesthetic Assessment
  - Surpasses many **supervised** baselines

Method	SRCC	PLCC
Kong *et al.* [24]	0.558	-
NIMA (Inception-v2) [43]	0.612	0.636
AFDC + SPP [2]	0.649	0.671
MaxViT [46]	0.708	0.745
AMP [31]	0.709	-
Zeng *et al.* (resnet101) [55]	0.719	0.720
MUSIQ [19]	0.726	0.738
Niu *et al.* [33]	0.734	0.740
MLSP (Pool-3FC) [15]	0.756	0.757
TANet [13]	0.758	**0.765**
$GAT_{\times 3}$-GATP [12]	**0.762**	0.764
Zero-shot Learning		
VILA-P (single prompt)	0.605	0.617
VILA-P (ensemble prompts)	0.657	0.663

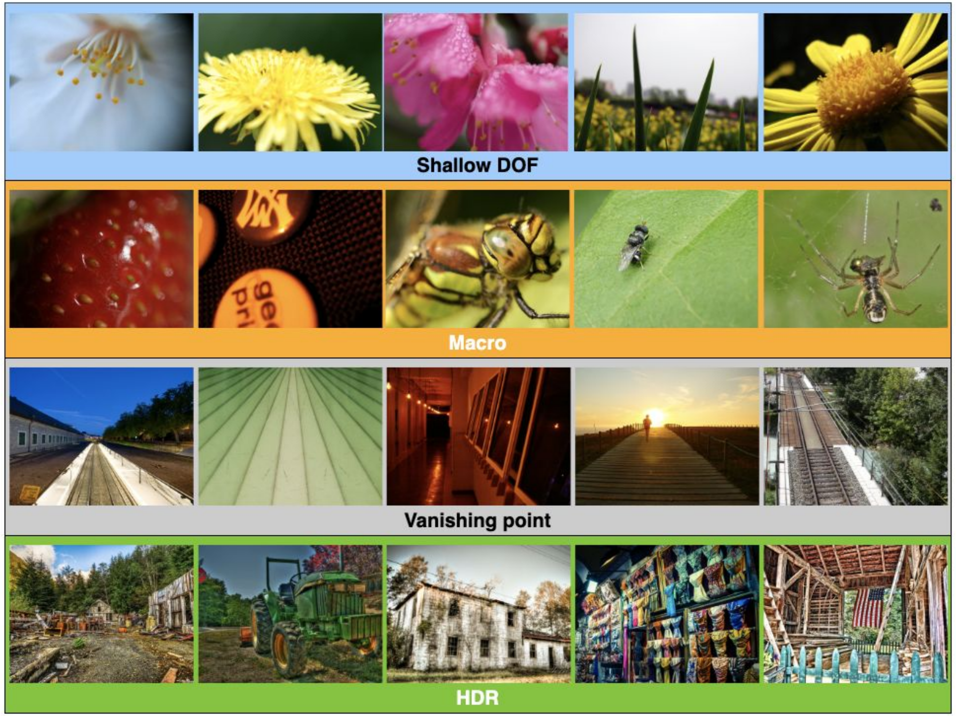
Image Aesthetic Assessment on AVA

VILA-P: Experiment Results

- ZSL for Style Classification

Method	mAP (%)
Murray *et al.* [36]	53.9
Karayev *et al.* [19]	58.1
Lu *et al.* [32]	64.1
MNet [46]	65.5
Sal-RGB [10]	71.8
Zero-shot Learning	
General Pretraining (single prompt)	29.3
General Pretraining (ensemble prompts)	32.6
VILA-P (single prompt)	62.3
VILA-P (ensemble prompts)	**69.0**

Table 4. Results on AVA-Style dataset. We gray out supervised baselines as they are not directly comparable to our unsupervised model which is not exposed to the training labels.



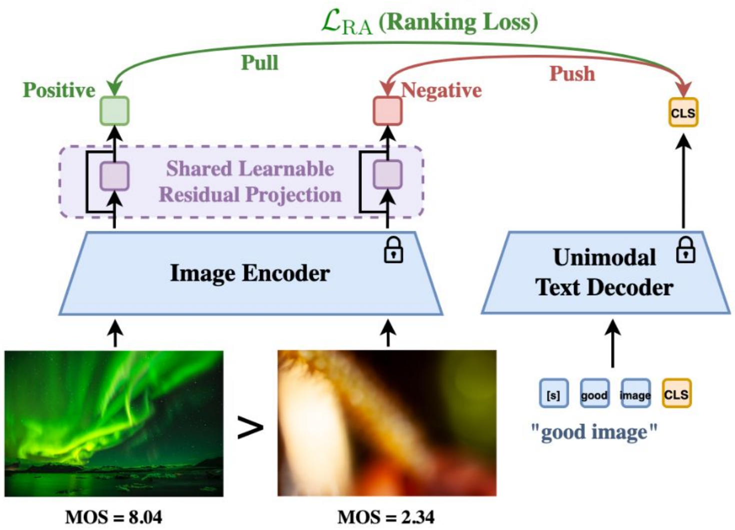
Top-5 Retrieved Images

Proprietary + Confidential

VILA-R: Rank-based Adapter for IAA

- Inspired from ZSL setting, using text prompts to score images
  - Use the frozen text embedding of “good image” as an anchor to score images
  - Adjust image representation (w/ a learnable residual projection) to optimize the relative ranking between two images
- Tunes only **0.1%** of the total parameters

(2) VILA-R: Rank-based Adapter for IAA



$$\tilde{\boldsymbol{v}} = \mathrm{normalize}(\boldsymbol{v}^\top \boldsymbol{H} + \boldsymbol{v}),$$

$$r = \tilde{\boldsymbol{v}}^\top \boldsymbol{w}_p$$

$$\mathcal{L}_{\mathrm{RA}} = \frac{1}{P} \sum_{i,j,i\neq j,l_i>l_j} \max\left(0, m - \tilde{\boldsymbol{v}}_i^\top \boldsymbol{w}_p + \tilde{\boldsymbol{v}}_j^\top \boldsymbol{w}_p\right)$$

VILA-R: Experiment Results

- State-of-the-art performance on image aesthetics assessment over AVA

Method	SRCC	PLCC
Kong *et al.* [24]	0.558	-
NIMA (Inception-v2) [43]	0.612	0.636
AFDC + SPP [2]	0.649	0.671
MaxViT [46]	0.708	0.745
AMP [31]	0.709	-
Zeng *et al.* (resnet101) [55]	0.719	0.720
MUSIQ [19]	0.726	0.738
Niu *et al.* [33]	0.734	0.740
MLSP (Pool-3FC) [15]	0.756	0.757
TANet [13]	0.758	**0.765**
$\mathrm{GAT}_{\times 3}$-GATP [12]	**0.762**	0.764
Zero-shot Learning		
VILA-P (single prompt)	0.605	0.617
VILA-P (ensemble prompts)	0.657	0.663
VILA-R	**0.774**	**0.774**

Image Aesthetic Assessment on AVA

Ablation: Necessity of Aesthetic Pretraining

- Aesthetic related information is **under-represented** in general image-text pairs from the Web
- Learning on noisy image-comment pairs from photo sharing website captures the **rich aesthetic semantics**

	ZSL Ens. Prompts		
General Pretraining	✓		✓
Aesthetic Pretraining		✓	✓
SRCC	0.228	0.265	**0.657**
PLCC	0.228	0.276	**0.663**

ZSL performance on AVA Image Aesthetic Assessment

	ZSL Single Prompt		ZSL Ens. Prompt	
General Pretraining	✓	✓	✓	✓
Aesthetic Pretraining		✓		✓
mAP	29.3	**62.3**	32.6	**69.0**

ZSL performance on AVA-Style classification

Ablation: Effectiveness of the Rank-based Adapter

- **Using text anchor is better:** it leverages the rich textual aesthetic information from pretraining
- **Learning a residual is better:** we only need to slightly adjust the image embedding
- **Finetune can further improve performance, but disturbs the generic pretrained weights**
  - E.g. AVA-Style mAP drops from 69% to 26%

Method	SRCC	PLCC
VILA-P w/ L2 Loss	0.757	0.756
VILA-P w/ EMD Loss [43]	0.759	0.759
VILA-R w/o Text Anchor	0.763	0.764
VILA-R w/o Residual	0.766	0.766
VILA-R (Ours)	**0.774**	**0.774**
VILA-R Finetune Image Encoder	0.780	0.780

Table 3. Ablation for the proposed rank-based adapter (Sec. 4) on AVA. First two groups use frozen pretrained image encoder.

Thanks!