Composed Image Retrieval with Text Feedback via Multi-grained Uncertainty Regularization

Yiyang Chen, Zhedong Zheng, Wei Ji,
Leigang Qu, Tat-seng Chua

Yiyang Chen
Tsinghua University
ICLR2024

Outline

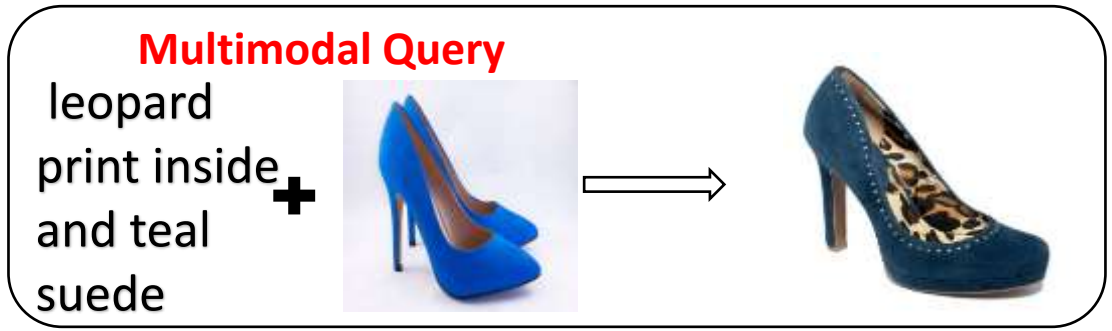
- Background
- Method
- Experiment

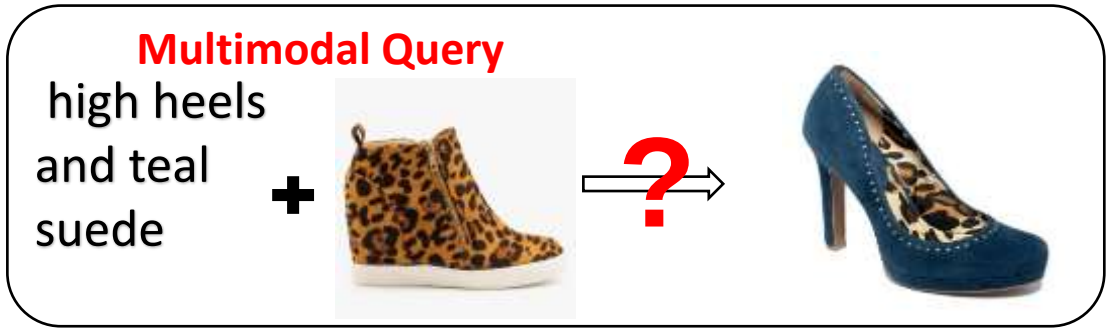
Outline

- Background
- Method
- Experiment

Multimodal Query for Fashion Search

- Multimodal query is easy and feasible for users to express their intention, since users are involved to refine the query.
- Query becomes : " I want similar [image] but with [text]." But the target image is not necessarily what the user wants, especially query for different styles.



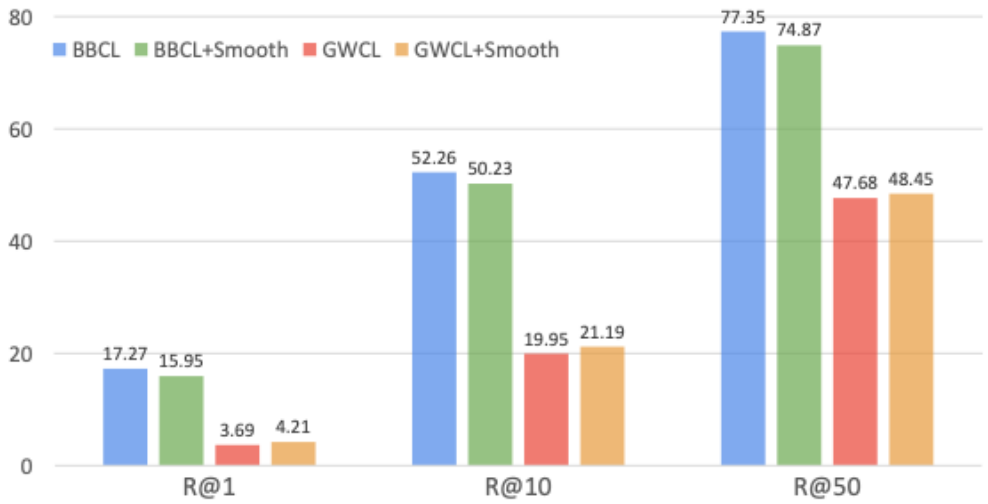


The Ambiguity Problem in Image Retrieval with Text

- This could be sensor noise, motion noise and labeling error in dataset.
- The existing methods usually mistake these candidate images as negative samples.
- These problems result in uncertainty which cannot be reduced even if more data were to be collected.

Aleatoric Uncertainty in Dataset*

$$\mathcal{L}_{\mathrm{NN}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2\sigma(\mathbf{x}_i)^2}||\mathbf{y}_i - \mathbf{f}(\mathbf{x}_i)||^2 + \frac{1}{2}\log\sigma(\mathbf{x}_i)^2$$



- **Aleatoric uncertainty** is the uncertainty arising from the natural stochasticity of observations, describes the inherent noise in the dataset. The unavoidable error cannot be reduced by increasing the number of samples.

- Global-wise classification deploys all categories, significantly degrades the performance since there are many false negative samples when calculating the loss due to aleatoric uncertainty in dataset.

* Kendall et al. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?". NeurIPS, 2017

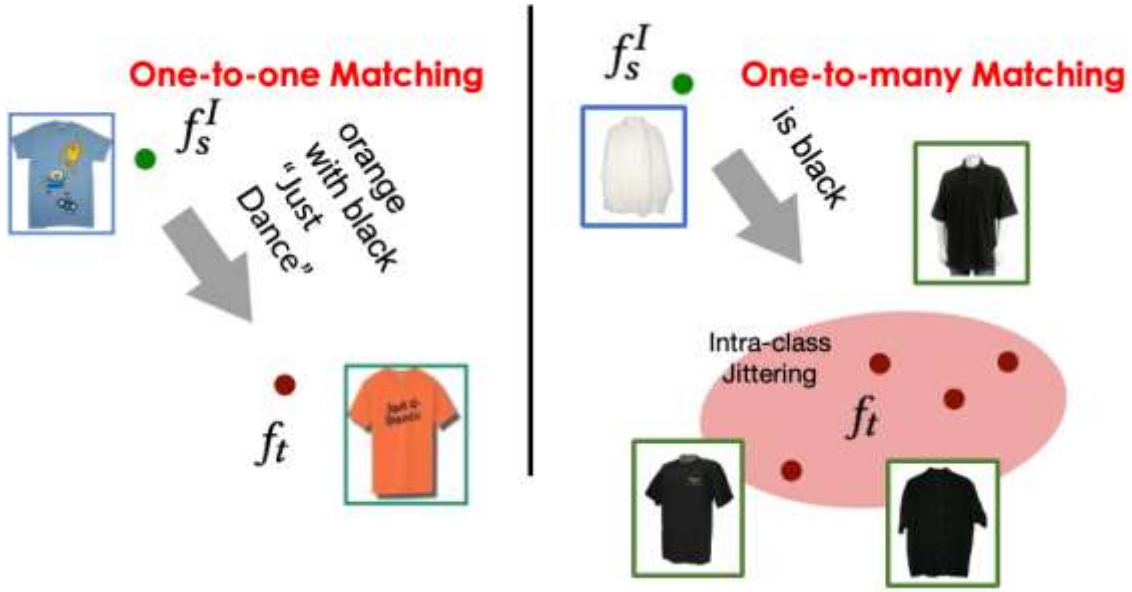
Multi-grained Retrieval

Integrate fine- and coarse-grained retrieval as matching data points with fluctuations

- **Fine-grained**: one-to-one matching
- **Coarse-grained**: matching between one query point and a point with an uncertain range

Match with an Uncertain Range

- Leverage uncertain fluctuation to build the multi-grained area in the representation space
- The uncertain range is a feature space, include-ing multiple potential candidates due to the imprecise query images or the ambiguous textual description



Multi-grained Retrieval via Uncertainty Regularization

Integrate fine- and coarse-grained retrieval as matching data points with fluctuations

- Introduce identically distributed fluctuations in the feature space to uncertainty modeling
- Introduce uncertainty regularization to adapt the matching objective according to the fluctuation range

Outline

- Background
- Method
- Experiment

Content and Style Compositor*

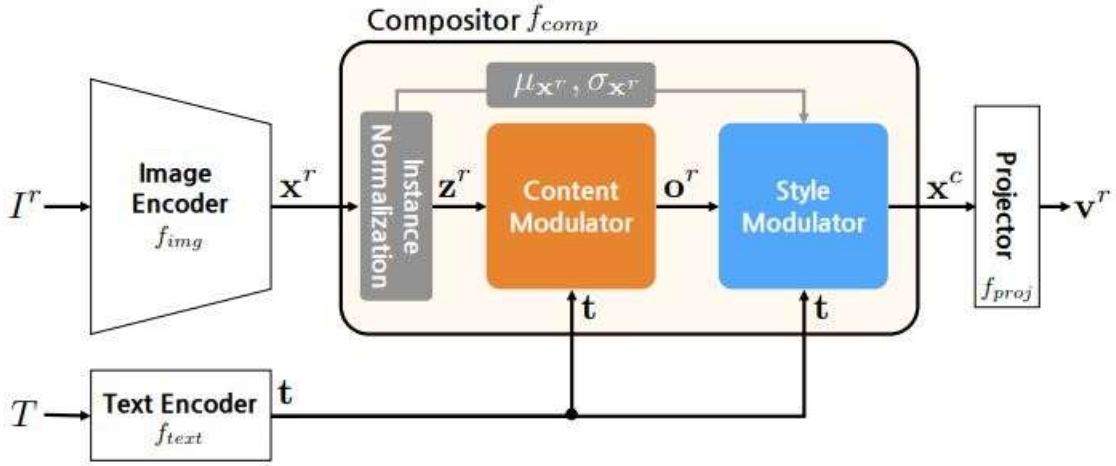
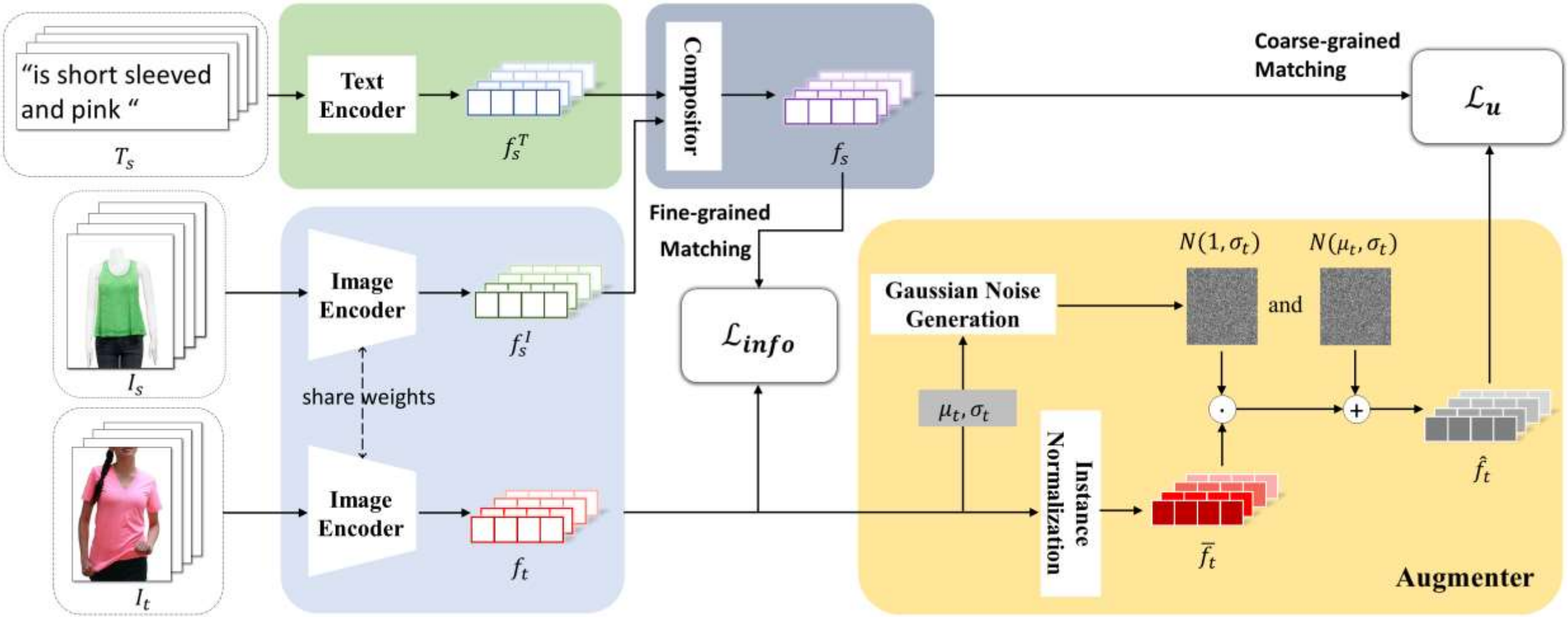


Figure 2. Overall pipeline of CoSMo.

- **Content Modulator** performs local updates to the reference image feature
- **Style Modulator** reintroduces global style information to the updated feature
- Focus on fine-grained matching

* Lee et al. "CoSMo: Content-Style Modulation for Image Retrieval with Text Feedback". CVPR, 2021

Pipeline


"is short sleeved and pink "
T_s
Text Encoder
f_s^T
Compositor
f_s
Coarse-grained Matching
$\mathcal{L}_u$
I_s
Image Encoder
f_s^I
share weights
Image Encoder
f_t
I_t
Fine-grained Matching
$\mathcal{L}_{info}$
Gaussian Noise Generation
$N(1, \sigma_t)$
and
$N(\mu_t, \sigma_t)$
μ_t, σ_t
Instance Normalization
$\bar{f}_t$
$\hat{f}_t$
Augmenter

Uncertainty Modeling

$$\mu,\ \sigma_t \leftarrow f_t$$

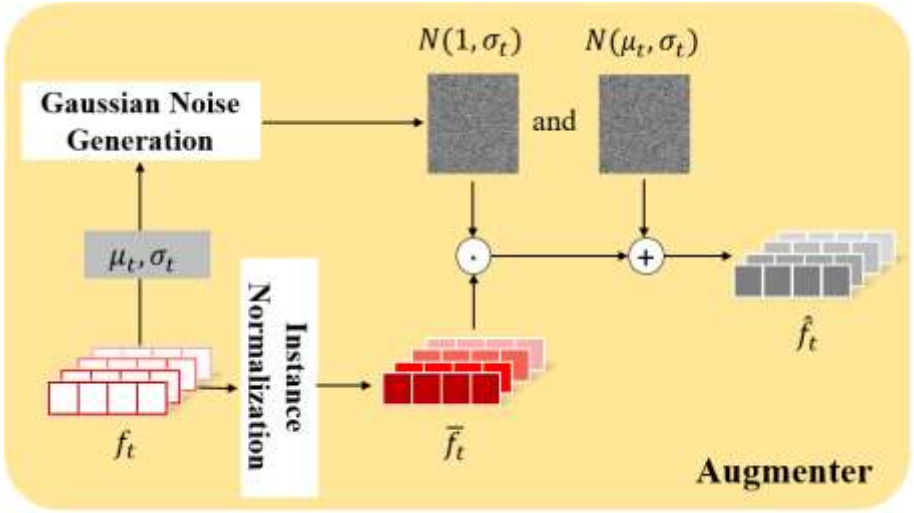
$$\bar{f}_t = \frac{f_t - \mu}{\sigma_t}$$

$$\alpha \sim N(1, \sigma_t),\ \beta \sim N(\mu, \sigma_t)$$

$$\hat{f}_t = \alpha \cdot \bar{f}_t + \beta$$

- f_t : Target Image Features

1. Get the mean and standard deviation.
2. Normalize f_t , use Batch Normal, Instance Normal or other methods.
3. Generate 2 Gaussian noisy.
4. Augment noisy to features.



The Augmenter generates the jittering and works on the final representation space directly.

Uncertainty Regularization

$$\mathcal{L}_{\text{info}}\left(f_s,\ f_t\right)=\frac{1}{B}\sum_{i=1}^{B}-\log\frac{\exp\left(\kappa\left(f_s^i,f_t^i\right)\right)}{\sum_{j=1}^{B}\exp\left(\kappa\left(f_s^i,f_t^j\right)\right)}$$

$$\mathcal{L}_{\text{u}}\left(f_s,\ \hat{f}_t,\sigma\right)=\frac{\mathcal{L}_{\text{info}}\left(f_s,\ \hat{f}_t\right)}{2\sigma^2}+\frac{1}{2}\log\sigma^2.$$

$$\mathcal{L}_{\text{total}}=\gamma\mathcal{L}_{\text{u}}\left(f_s,\ \hat{f}_t,\sigma_t\right)+(1-\gamma)\mathcal{L}_{\text{info}}\left(f_s,\ f_t\right).$$

- Fine-grained matching: InfoNCE loss function
- Coarse-grained matching: inspired by Aleatoric Uncertainty
- γ is a dynamic weight hyperparameter to balance the ratio of the fine- and coarse-grained retrieval

Outline

- Background
- Method
- Experiment

R@50 on Three Datasets

R@50	FashionIQ	Shoes	Fashion200k
Baseline	57.23	76.46	67.8
Ours	61.39	79.84	70.2
	+4.03	+3.38	+2.40

Results on FashionIQ

Method	Visual Backbone	Dress		Shirt		Toptee		Average	
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
MRN (Kim et al., 2016)	ResNet-152	12.32	32.18	15.88	34.33	18.11	36.33	15.44	34.28
FiLM (Perez et al., 2018)	ResNet-50	14.23	33.34	15.04	34.09	17.30	37.68	15.52	35.04
TIRG (Vo et al., 2019b)	ResNet-17	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39
Pic2Word (Saito et al., 2023)	ViT-L/14	20.00	40.20	26.20	43.60	27.90	47.40	24.70	43.70
VAL (Chen et al., 2020)	ResNet-50	21.12	42.19	21.03	43.44	25.64	49.49	22.60	45.04
ARTEMIS (Delmas et al., 2022)	ResNet-50	27.16	52.40	21.78	54.83	29.20	43.64	26.05	50.29
CoSMo (Lee et al., 2021)	ResNet-50	25.64	50.30	24.90	49.18	29.21	57.46	26.58	52.31
DCNet (Kim et al., 2021)	ResNet-50	28.95	56.07	23.95	47.30	30.44	58.29	27.78	53.89
FashionViL (Han et al., 2022)	ResNet-50	28.46	54.24	22.33	46.07	29.02	57.93	26.60	52.74
FashionViL* (Han et al., 2022)	ResNet-50	33.47	59.94	25.17	50.39	34.98	60.79	31.21	57.04
Baseline	ResNet-50	24.80	52.35	27.70	55.71	33.40	63.64	28.63	57.23
Ours	ResNet-50	**30.60**	**57.46**	**31.54**	**58.29**	**37.37**	**68.41**	**33.17**	**61.39**
CLVC-Net (Wen et al., 2021)	ResNet-50×2	29.85	56.47	28.75	54.76	33.50	64.00	30.70	58.41
Ours	ResNet-50×2	**31.25**	**58.35**	**31.69**	**60.65**	**39.82**	**71.07**	**34.25**	**63.36**
CLIP4Cir (Baldrati et al., 2022)	ResNet-50×4	31.63	56.67	**36.36**	58.00	38.19	62.42	35.39	59.03
Ours	ResNet-50×4	**32.61**	**61.34**	33.23	**62.55**	**41.40**	**72.51**	**35.75**	**65.47**

Results on Shoes and Fashion200k

Method	Shoes			Fashion200k		
	R@1	R@10	R@50	R@1	R@10	R@50
MRN(Kim et al., 2016)	11.74	41.70	67.01	13.4	40.0	61.9
FiLM(Perez et al., 2018)	10.19	38.89	68.30	12.9	39.5	61.9
TIRG(Vo et al., 2019b)	12.6	45.45	69.39	14.1	42.5	63.8
VAL(Chen et al., 2020)	16.49	49.12	73.53	21.2	49.0	68.8
CoSMo(Lee et al., 2021)	16.72	48.36	75.64	23.3	50.4	69.3
DCNet(Kim et al., 2021)	-	53.82	79.33	-	46.9	67.6
ARTEMIS(Delmas et al., 2022)	18.72	53.11	79.31	21.5	51.1	70.5
Baseline	15.26	49.48	76.46	19.5	46.7	67.8
Ours	18.41	53.63	79.84	21.8	52.1	70.2

THANKS

Yiyang Chen
Tsinghua University
ICLR2024

Code[github]