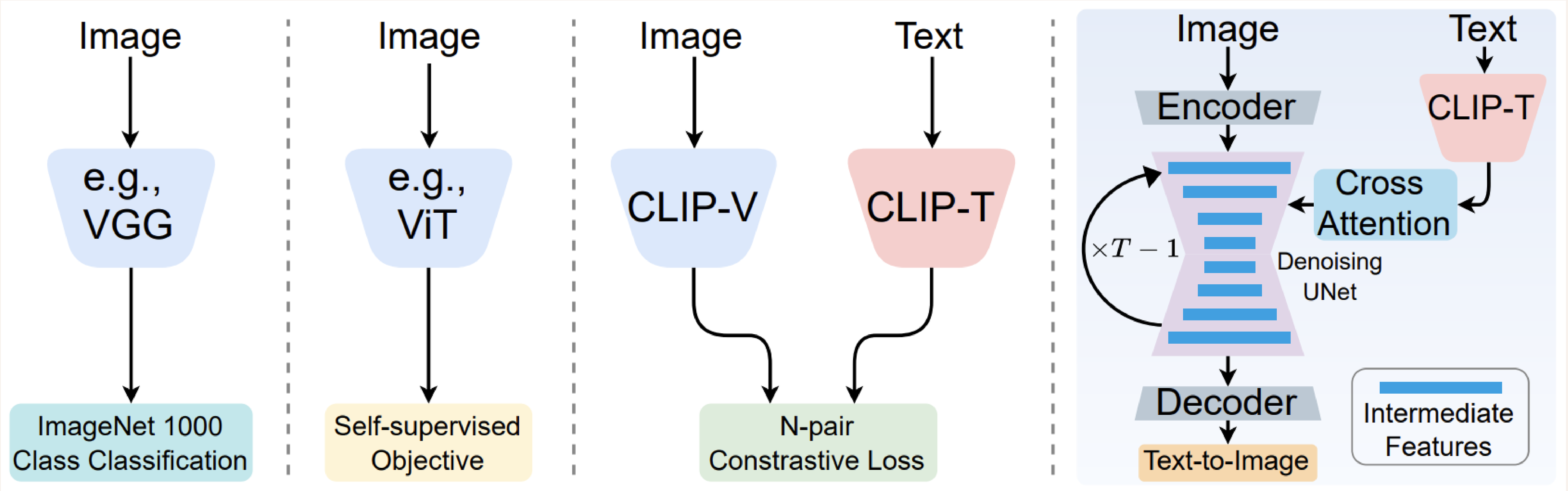
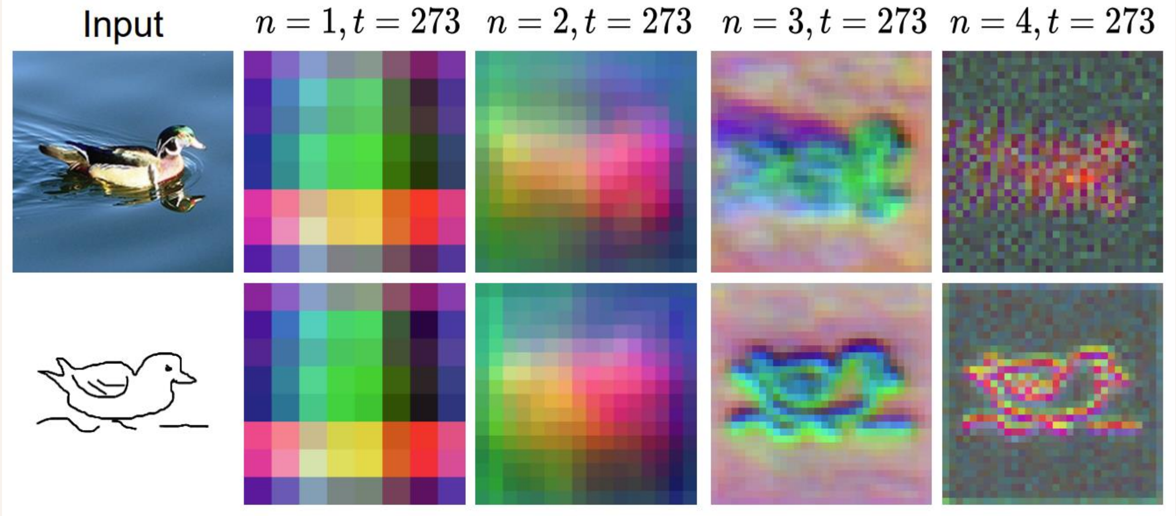
Got 2 Minutes? Start here ⬇



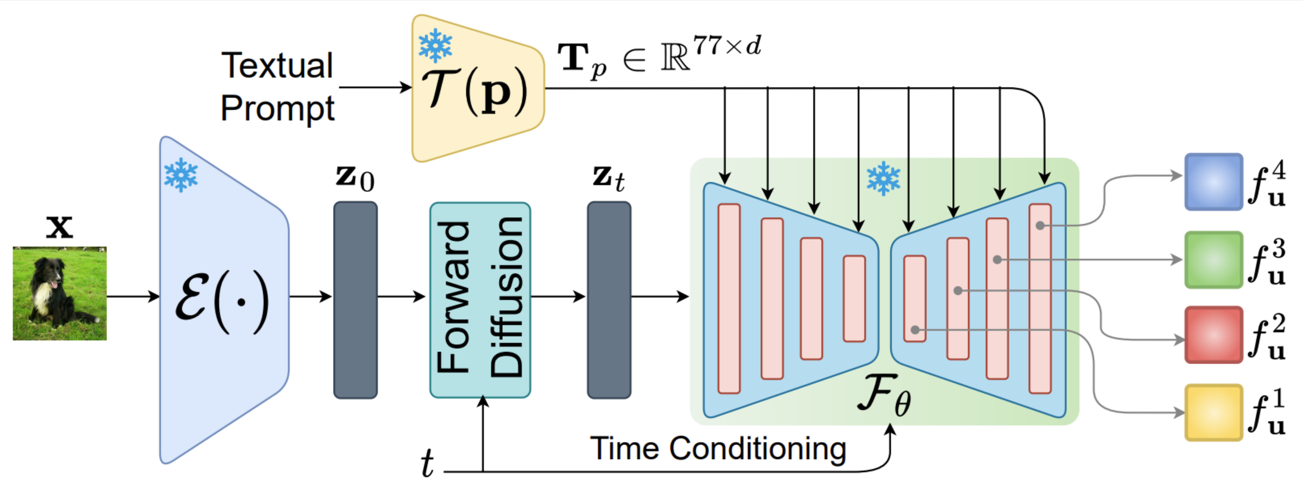
Existing Sketch-Based Image Retrieval backbones **Proposed**

Summary

- This paper unveils the latent potential of **diffusion models** as **backbone feature extractors for ZS-SBIR**.
- **Intermediate UNet features** from different upsampling blocks of pre-trained diffusion model depict **significant semantic similarity** across modalities (*e.g.*, sketch and photo).



- We also introduce innovative design strategies, including **soft prompt learning** and **visual prompting**, for task-specific (*i.e.*, ZS-SBIR) adaptation of pre-trained diffusion model.
- Given an image-prompt pair (x, p), and a time-step t, we first generate the latent image $z_0 = \mathcal{E}(x)$.



- We then add noise from time-step t to transform z_0 to its t^{th}-step noisy latent image z_t.
- Now we feed, – *(i)* the noisy latent z_t, *(ii)* scalar time-step value t, and *(iii)* textual embedding $T_p = \mathcal{T}(p)$ into $\mathcal{F}_\theta(\cdot)$ to extract corresponding intermediate features from upsampling layers.
- With this **diffusion-based backbone feature extractor**, we demonstrate marked improvements in all forms of ZS-SBIR (*i.e.*, **category-level** and **fine-grained**).

Text-to-Image Diffusion Models are Great Sketch-Photo Matchmakers

Subhadeep Koley[1,2], Ayan Kumar Bhunia[1], Aneeshan Sain[1], Pinaki Nath Chowdhury[1], Tao Xiang[1,2], Yi-Zhe Song[1,2]

[1]*SketchX, CVSSP, University of Surrey*

[2]*iFlyTek-Surrey Joint Research Centre on Artificial Intelligence*

CVPR JUNE 17-21, 2024 SEATTLE, WA

SketchX

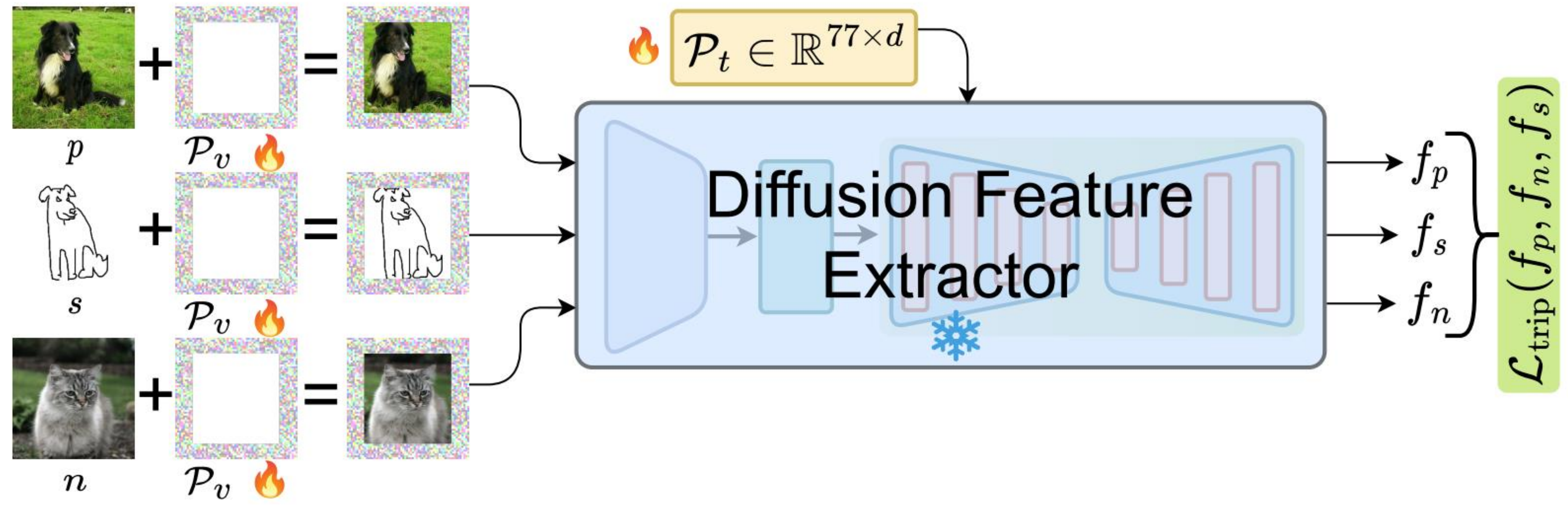
UNIVERSITY OF SURREY

CVSSP Centre for Vision, Speech and Signal Processing

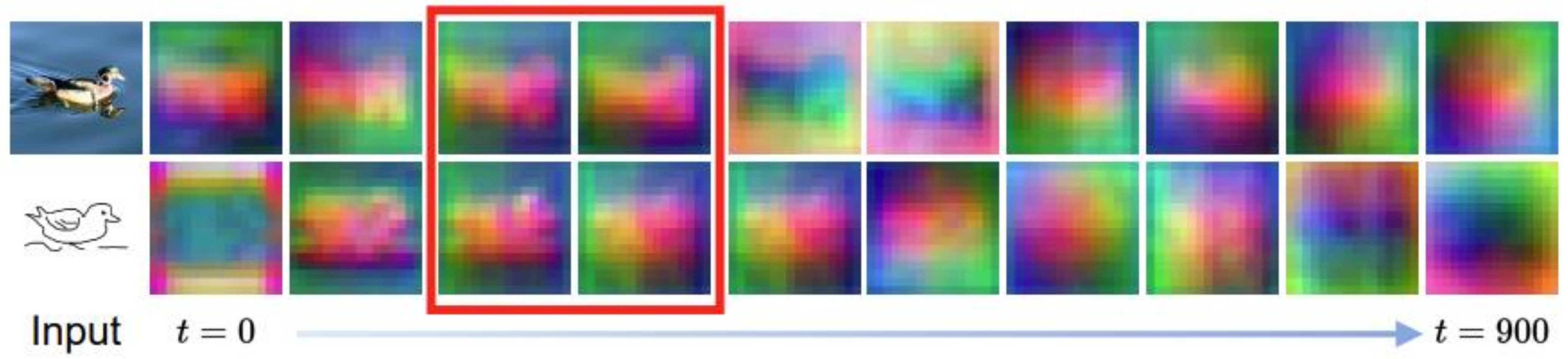
Proposed Model

- **Salient Components**
  1. Stable Diffusion (SD) model as backbone feature extractor.
  2. Learnable task-specific visual prompt for task-specific adaptation.
  3. Learnable textual prompt to harness the visio-linguistic prior of Stable Diffusion.

- **Visual prompting** learns a soft image perturbation in the pixel space to adapt SD model to our problem setup of ZS-SBIR.



- Text-to-image SD model being trained on **text-to-image generative objective**, works best with explicit textual prompts.
- Thus, instead of actual textual prompt embedding $T_p = \mathcal{T}(p) \in \mathbb{R}^{77\times d}$, we use a **learnable continuous textual prompt embedding** matrix $\mathcal{P}_t \in \mathbb{R}^{77\times d}$, influencing the SD feature extraction process via cross-attention.
- **Forward diffusion invokes stochasticity** due to the random noise sampling, which deteriorates the quality of extracted features. To tackle this, we extract SD features for each image/sketch six times each from different noise samples, and ensemble them by averaging to obtain the final feature.
- Empirically, we observe that timestep $t = 273$, and the decoder level of $n = 1, 2$ and $n = 3, 4$ works best for ZS-SBIR and ZS-FG-SBIR respectively.



Experiments & Results

Table 1. Results for category-level ZS-SBIR.

Methods	Sketchy [81]		TU-Berlin [21]		Quick, Draw! [27]	
	mAP@200	P@200	mAP@all	P@100	mAP@all	P@200
ZS-CAAE [103]	0.156	0.260	0.005	0.003	–	–
ZS-CVAE [103]	0.225	0.333	0.005	0.001	0.003	0.003
ZS-CCGAN [20]	–	–	0.297	0.426	–	–
ZS-GRL [16]	0.369	0.370	0.110	0.121	0.075	0.068
ZS-SAKE [52]	0.497	0.598	0.475	0.599	–	–
ZS-IIAE [35]	0.373	0.485	0.412	0.503	–	–
ZS-Sketch3T [77]	0.579	0.648	0.507	0.671	–	–
ZS-LVM [78]	0.723	0.725	0.651	0.732	0.202	0.388
B-Fine-Tuning	0.115	0.174	0.010	0.006	0.002	0.003
B-Linear-Probing	0.441	0.535	0.410	0.582	0.092	0.099
B-Triplet+VP (VGG)	0.651	0.682	0.582	0.673	0.134	0.310
B-Triplet+VP (ResNet)	0.326	0.342	0.354	0.512	0.105	0.275
B-Triplet+VP (ViT)	0.681	0.697	0.601	0.694	0.185	0.321
Ours	**0.746**	**0.747**	**0.680**	**0.744**	**0.231**	**0.397**

Table 2. Results for cross-category ZS-FG-SBIR on Sketchy

Methods	Acc.@1	Acc.@5	Methods	Acc.@1	Acc.@5
CC-Gen [62]	22.60	49.00	B-Triplet+VP (VGG)	24.20	43.61
CC-Grad [85]	13.40	34.90	B-Triplet+VP (ResNet)	15.61	27.64
CC-LVM [78]	28.68	62.34	B-Triplet+VP (ViT)	26.11	46.81
B-Fine-Tuning	1.85	6.01	*Ours*	**31.94**	**65.81**
B-Linear-Probing	17.32	41.23