SketchFusion: Learning Universal Sketch Features through Fusing Foundation Models

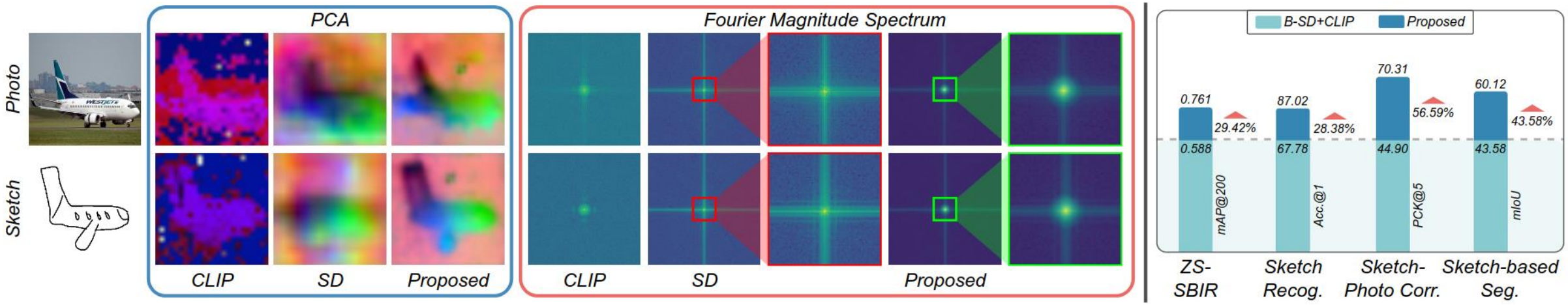
Subhadeep Koley, Tapas Kumar Dutta, Aneeshan Sain, Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Yi-Zhe Song

CVPR 2025

CVSSP
Centre for Vision, Speech and Signal Processing
People-Centred AI
UNIVERSITY OF SURREY
SketchX
UNIVERSITY OF SURREY
CVPR Nashville JUNE 11-15, 2025

Overview

- Introduces the first foundation model tailored for abstract, sparse hand-drawn sketches.
- Analyzes why diffusion models struggle with sketches.
- Proposes a hybrid Stable Diffusion + CLIP method leveraging complementary biases to improve sketch based downstream task performances.



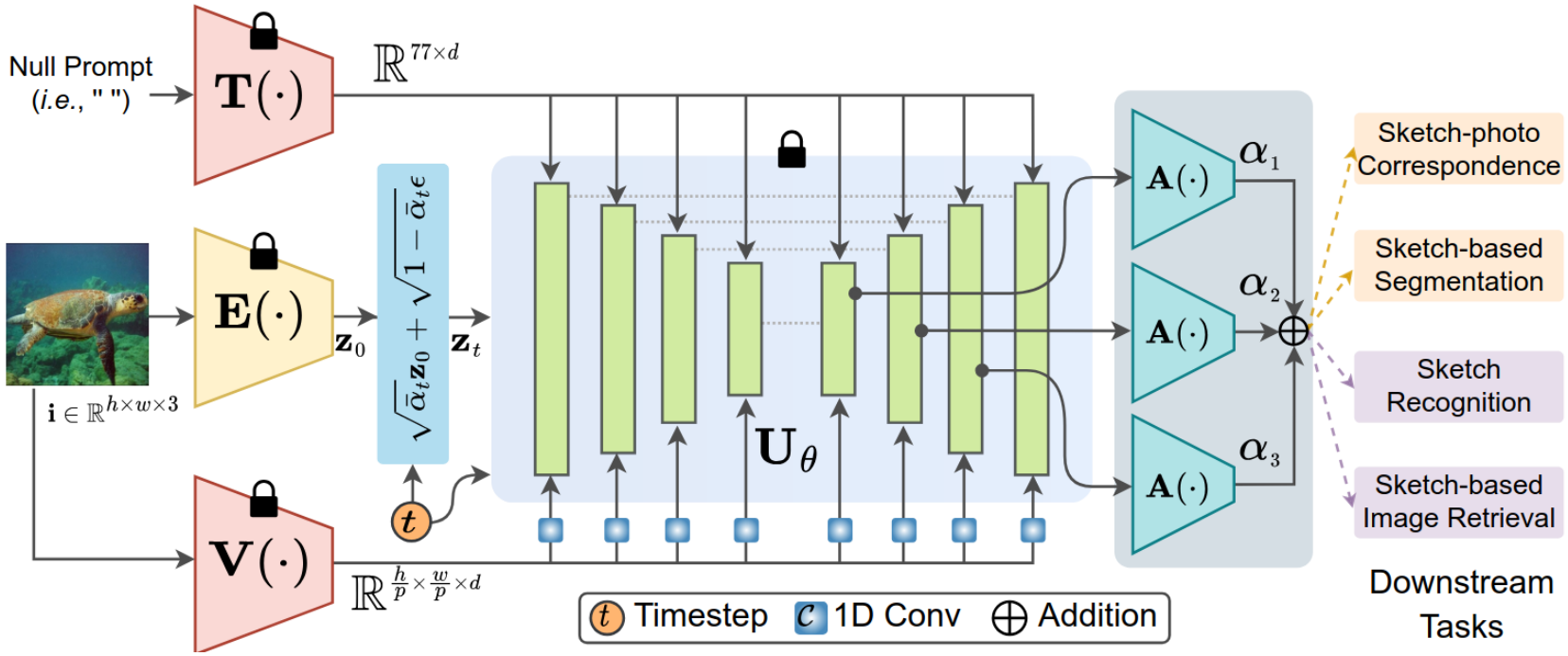
Problems of Existing Methods

1. No single model handles multiple sketch-based downstream tasks.
2. Diffusion models favor high-frequency details absent in sketches.
3. Manual feature aggregation needed for each task.

Our Solutions

1. Use CLIP’s low-frequency bias via adapters to enhance Diffusion models without retraining.
2. Experiments show strong gains across sketch tasks, setting a new universal feature learning standard.
3. Introduce adaptive feature aggregation to optimize for each task automatically.

Model Architecture



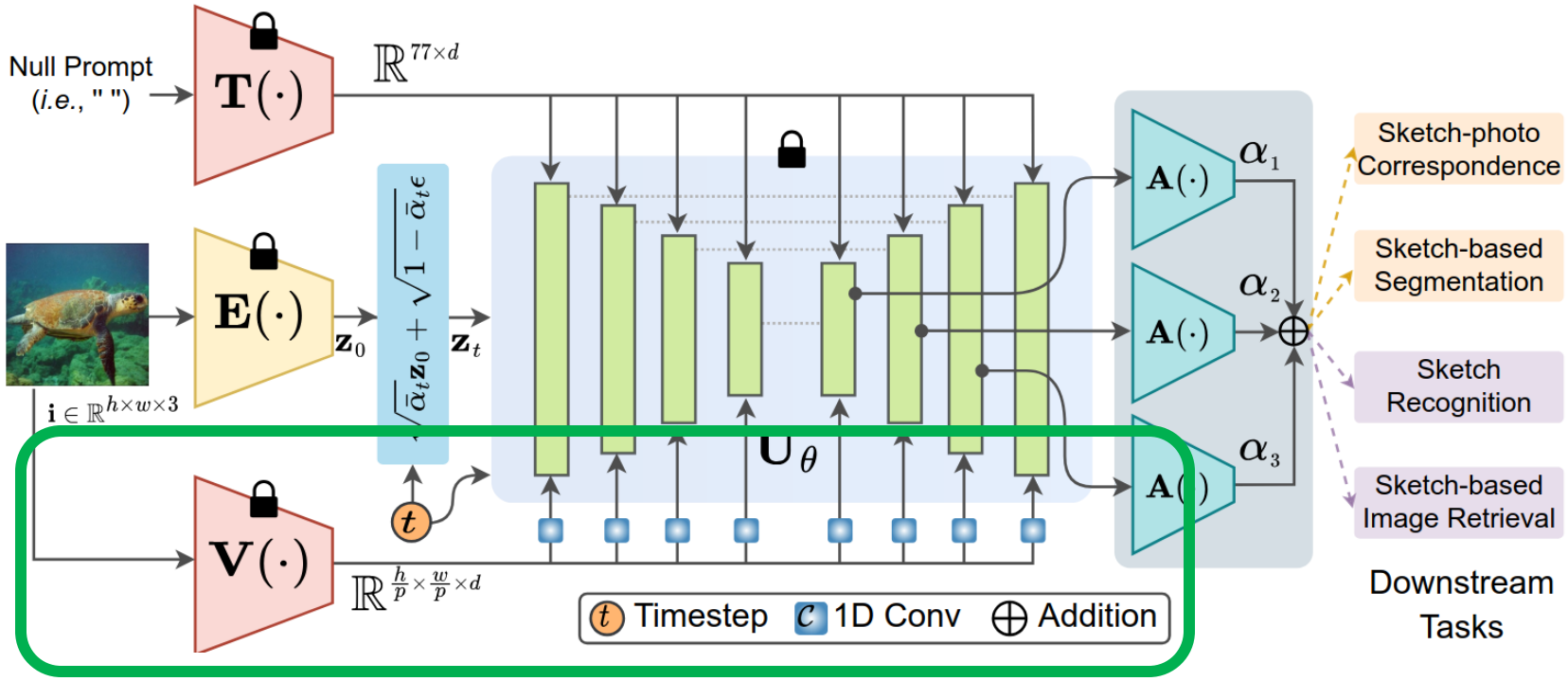
Salient Design Components:

1. **Multi-Layer Feature Injection:** CLIP features are transformed via 1D convolution layers, prior to being injected with UNet of SD.

2. **Adaptive Feature Aggregation:** Features from different UNet layers are fused dynamically using a ResNet-based aggregation with learnable weights.

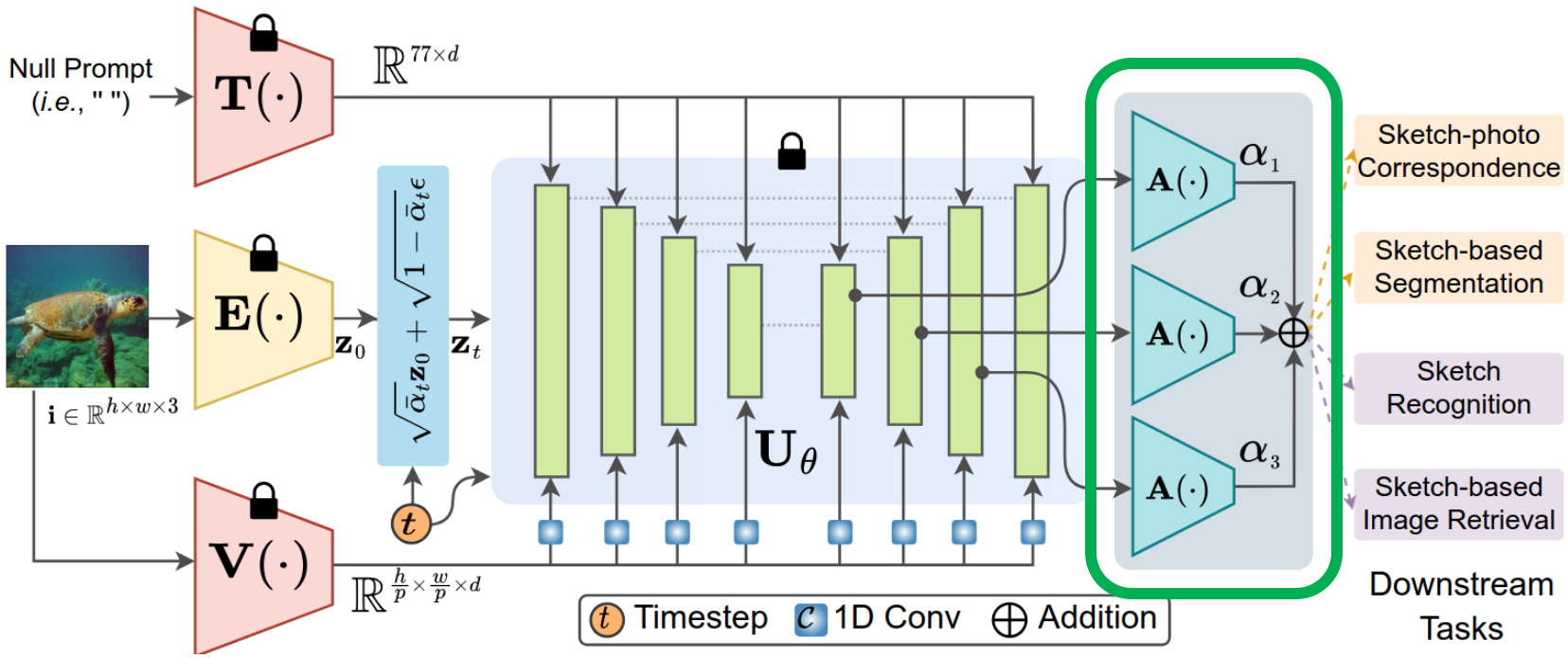
3. **Task-Agnostic Architecture:** The framework is designed to support diverse sketch-based vision tasks.

Multi-layer Feature Injection



- We encode sketches/photos as feature vectors via *pre-trained* CLIP vision encoder.
- This feature vector is transformed via adapters and combined to each UNet's feature representation via vector addition.

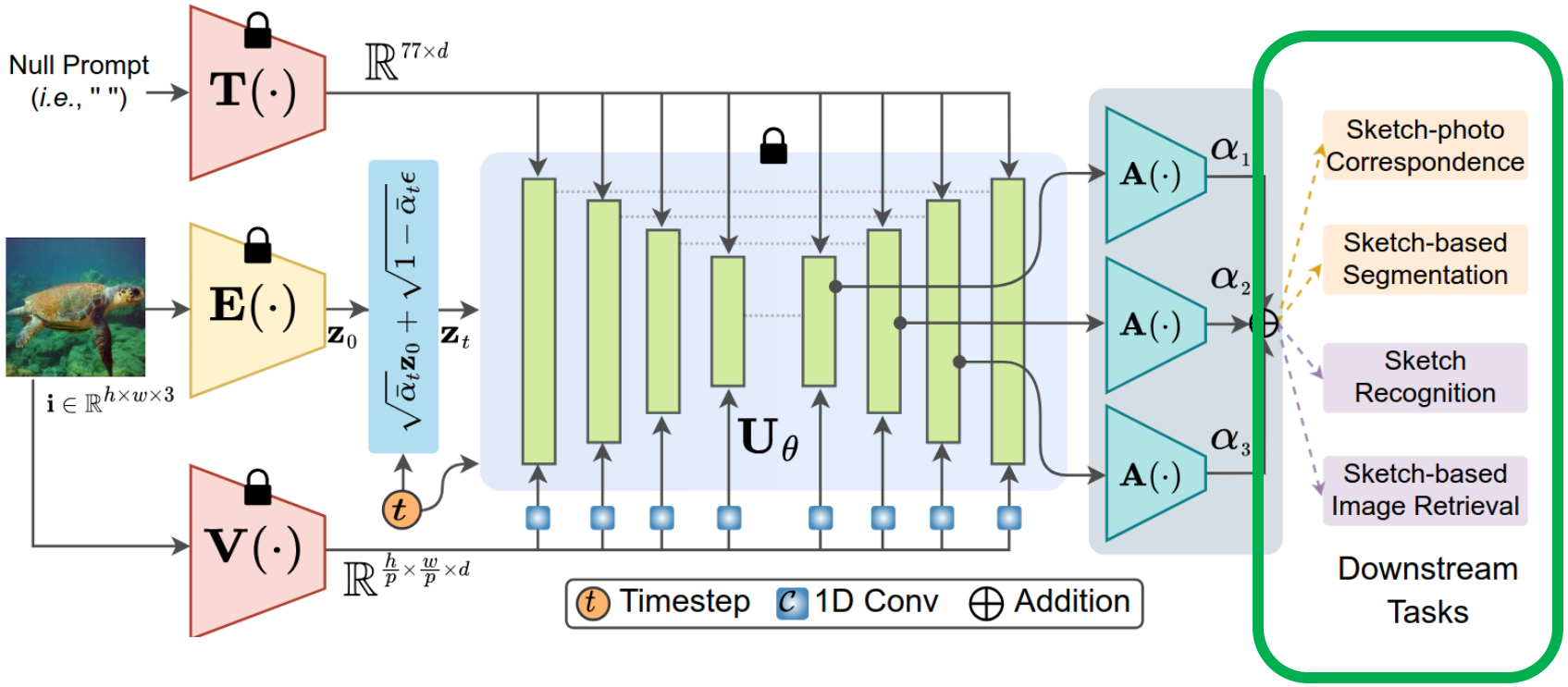
Adaptive Feature Aggregation



- We posit that dynamic combination of multiple feature representations works better than manual ensembling for each downstream tasks.

- Features from different UNet layers are fused *dynamically* using a ResNet-based aggregation module with learnable weights for each feature branch.

Task-Agnostic Architecture



- The framework is designed to support diverse sketch-based vision tasks using a common task-agnostic encoder.

- Only the injection and aggregation modules are trained using sketch-photo pairs (keeping the SD and CLIP models frozen), enabling multiple downstream tasks *without* text prompts.

Sketch-based Image Retrieval

Aim:

1. *Category-level* SBIR retrieves a photo from the same class as the input sketch, using a gallery with multiple classes and images per class.

2. *Fine-grained SBIR* extends this to instance-level matching across categories.

3. *Zero-shot SBIR* tests generalization by evaluating on unseen categories, where training and testing class sets are mutually exclusive.

Methods	Sketchy [77]		TU-Berlin [25]		Quick, Draw! [28]	
	mAP@200	P@200	mAP@all	P@100	mAP@all	P@200
B-CLIP	0.250	0.261	0.228	0.257	0.080	0.141
B-DINO	0.493	0.481	0.450	0.492	0.167	0.249
B-DINOv2	0.527	0.533	0.481	0.532	0.170	0.268
B-SD	0.558	0.571	0.510	0.561	0.179	0.287
B-SD+CLIP	0.588	0.592	0.537	0.589	0.179	0.311
B-Finetuning	0.120	0.172	0.011	0.010	0.003	0.006
SAKE [53]	0.497	0.598	0.475	0.599	–	–
IIAE [36]	0.373	0.485	0.412	0.503	–	–
CAAE [101]	0.156	0.260	0.005	0.003	–	–
CCGAN [24]	–	–	0.297	0.426	–	–
CVAE [101]	0.225	0.333	0.005	0.001	0.003	0.003
GRL [20]	0.369	0.370	0.110	0.121	0.075	0.068
LVM [75]	0.723	0.725	0.651	0.732	0.202	0.388
SD-PL [45]	0.746	0.747	0.680	0.744	0.231	0.397
Proposed	**0.761**	**0.763**	**0.695**	**0.753**	**0.242**	**0.399**

Results for *category-level ZS-SBIR*.

Methods	Acc.@1	Acc.@5	Methods	Acc.@1	Acc.@5
B-CLIP	11.84	21.66	B-Finetuning	5.67	9.17
B-DINO	22.49	46.97	Gen-VAE [62]	22.60	49.00
B-DINOv2	21.19	44.31	LVM [75]	28.68	62.34
B-SD	23.98	49.42	SD-PL [45]	31.94	65.81
B-SD+CLIP	24.16	52.01	***Proposed***	**33.01**	**67.92**

Results on Sketchy [77] for *cross-category ZS-FG-SBIR*.

Sketch Recognition

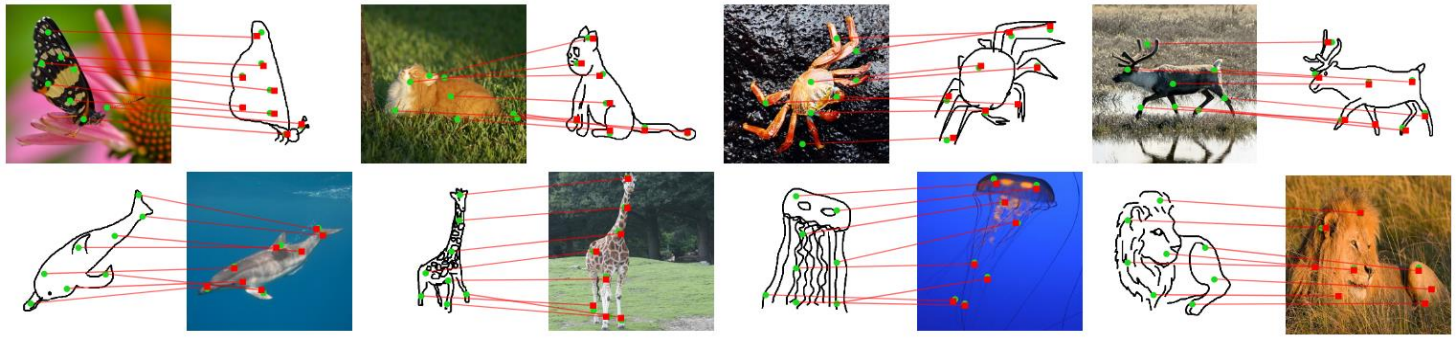
Aim: Classify freehand sketches into predefined classes using labeled training data.

Methods	TU-Berlin	Quick, Draw!	Methods	TU-Berlin	Quick, Draw!
B-CLIP	30.09	30.87	B-Finetuning	10.29	12.37
B-DINO	51.79	53.28	SketchMate [95]	77.96	79.44
B-DINOv2	58.01	59.12	SketchGNN [100]	76.43	77.31
B-SD	61.37	63.57	SketchXAI [65]	–	86.10
B-SD+CLIP	65.47	67.78	***Proposed***	**84.96**	**87.02**

Acc.@1 results for *sketch recognition*.

Sketch-Photo Correspondence Learning

Aim: Learn how to map a sketch to its corresponding photo by identifying and aligning semantically meaningful keypoints between them. Given a pair consisting of a sketch and its matching photo, along with a set of keypoints marked on the sketch, the task is to predict the locations of the corresponding keypoints on the photo.

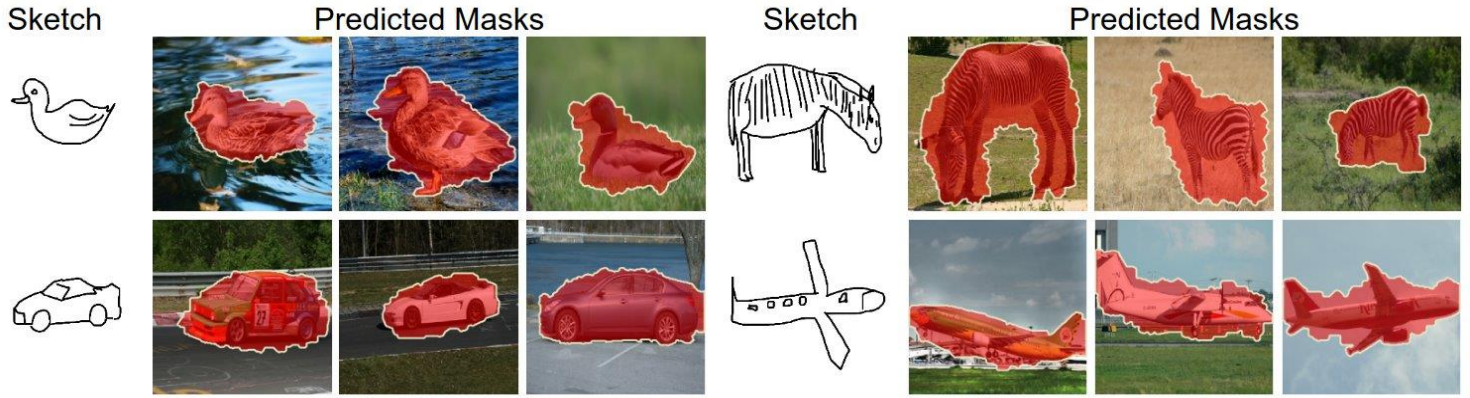


Methods	PCK@5	PCK@10	Methods	PCK@5	PCK@10
B-CLIP	22.29	30.41	B-Finetuning	11.58	17.69
B-DINO	34.21	48.59	WeakAlign [70]	43.55	78.60
B-DINOv2	39.91	56.46	WarpC [85]	56.78	79.70
B-SD	41.42	58.71	Self-Sup. [55]	58.00	84.93
B-SD+CLIP	44.90	62.02	***Proposed***	**70.31**	**89.86**

Results on PSC6K [55] for *sketch-photo correspondence*.

Sketch-based Image Segmentation

Aim: Predict a binary mask that highlights pixel locations where the sketched concept appears in a candidate image from that category. The predicted mask assigns 1 to pixels belonging to the concept and 0 to the rest.



Methods	mIoU	pAcc.	Methods	mIoU	pAcc.
B-CLIP	20.63	26.82	B-Finetuning	20.36	21.84
B-DINO	32.81	42.06	DeepLabv3+Sketch [35]	40.63	55.23
B-DINOv2	34.17	45.80	ZS-Seg [95]	44.73	60.97
B-SD	39.02	49.03	Sketch-a-Segmenter [35]	46.45	60.28
B-SD+CLIP	41.87	51.91	***Proposed***	**60.12**	**74.91**

Results for *sketch-based image segmentation.*

Thanks!