NEURAL INFORMATION PROCESSING SYSTEMS

A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence

Junyi Zhang[1], Charles Herrmann[2], Junhwa Hur[2], Luisa F. Polanía[2], Varun Jampani[2], Deqing Sun[2], Ming-Hsuan Yang[2,3]

① Shanghai Jiao Tong University ② Google Research ③ UC Merced

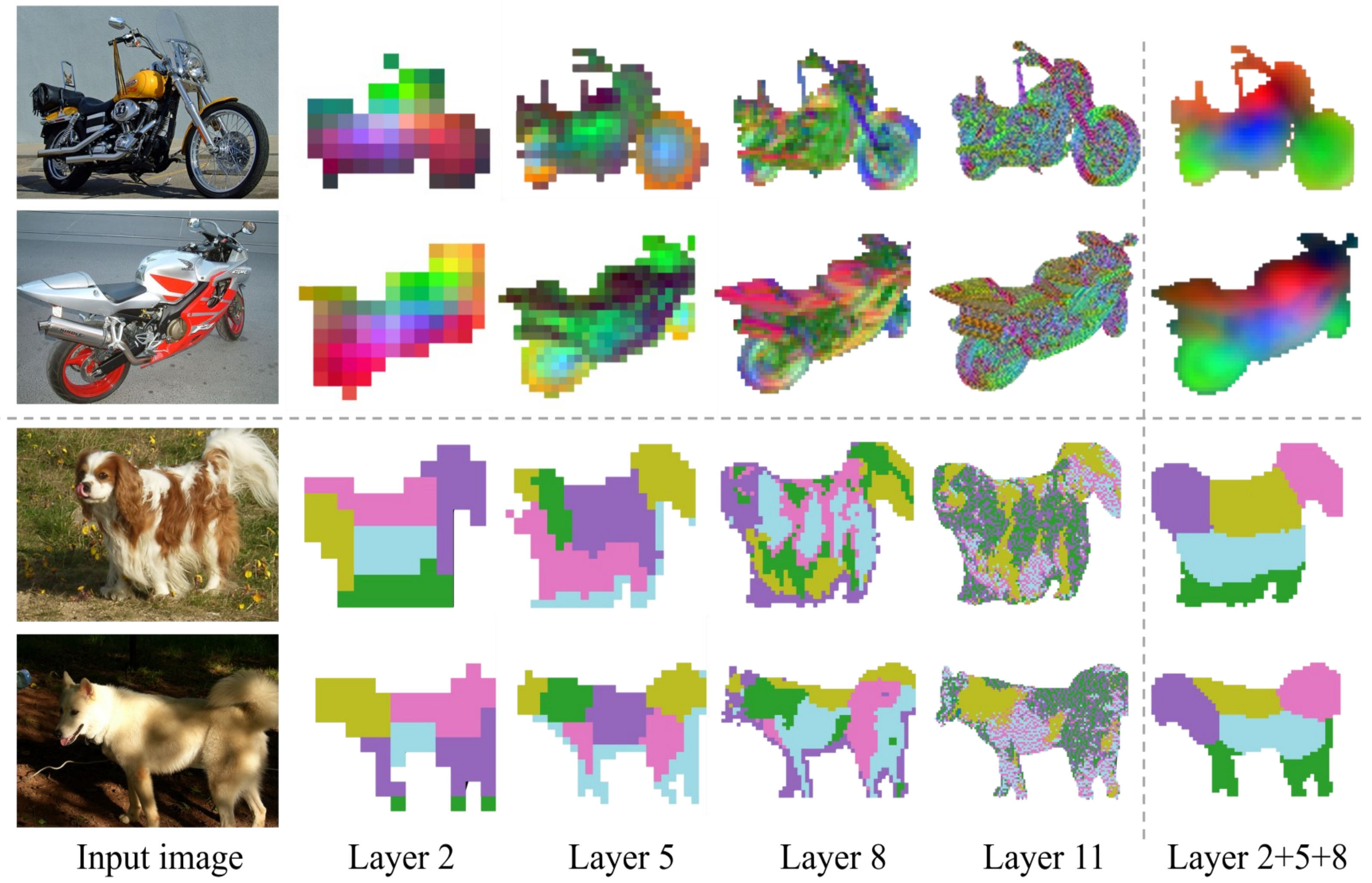
Code & Visual Results

Key Takeaways

- Stable Diffusion **(SD) features** shows great potential for **semantic and dense correspondence**
- **SD features** have very different properties compared to the **DINOv2 features** and naturally **forms a complementary**
- A **simple fusion strategy** can improve both single features
- The **fused features** with only a zero-shot evaluation can largely **outperform many SOTA methods**
- **Instance swapping** with high-quality correspondence

SD Features for Semantic Correspondence

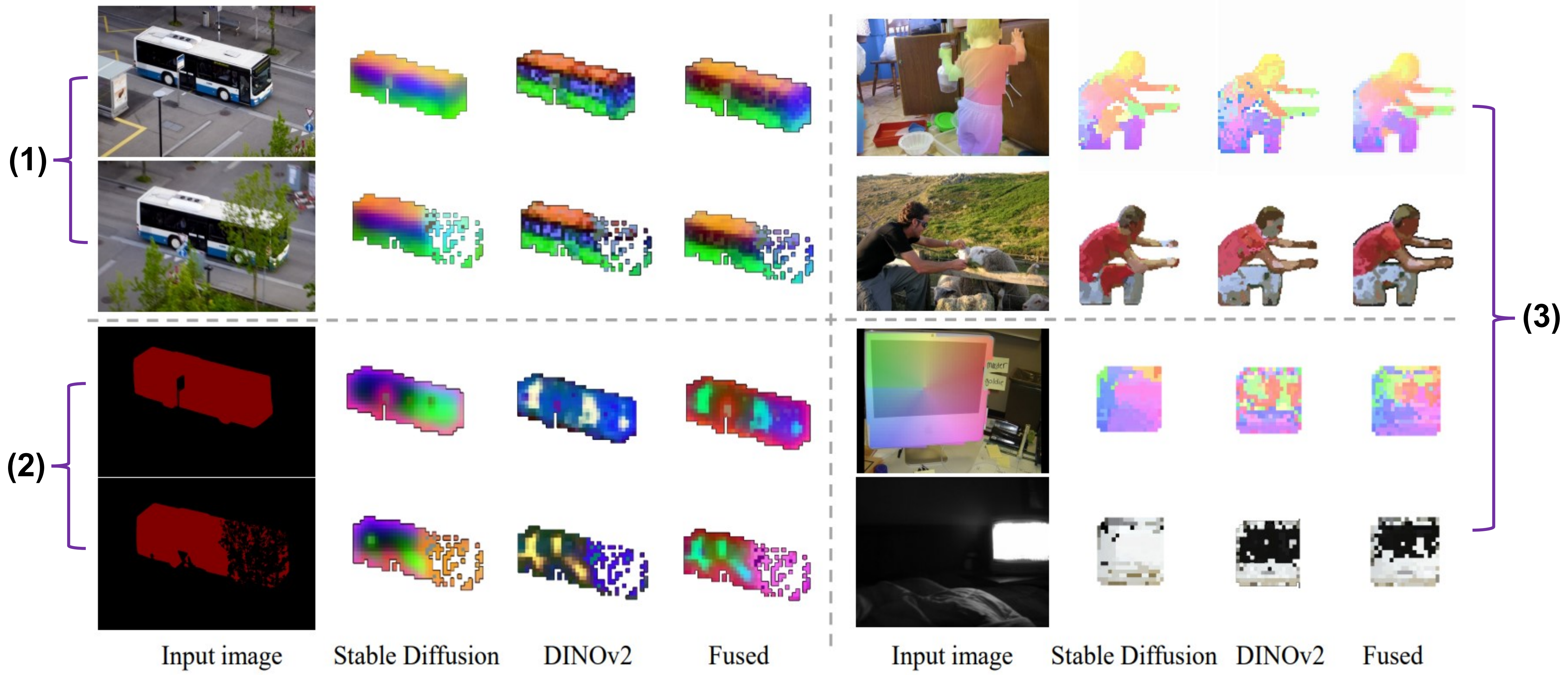
(Top) Visualization of first 3 channel of PCA-ed features
(Bottom) Visualization of cluster & match results



- Early layers (2,5): lower resolution, more semantic information
- Last layer (11): higher resolution, focuses on appearance
- **Our approach**: ensemble early and intermediate layers (2,5,8) to trade-off between resolution and semantics, apply co-PCA to reduce the dimension

Diverse Properties of SD Features and DINOv2 Features

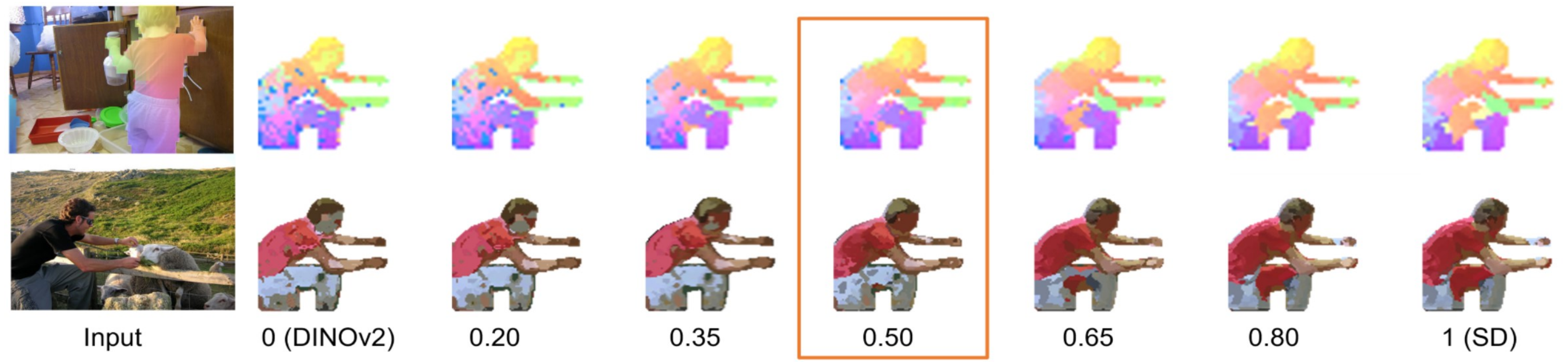
Analysis of different features for correspondence with
(Left) PCA visualization (Right) dense correspondence



- **(1)** Easier case: both two features can build plausible correspondence
- **(2)** Absent textual signal: DINOv2 fails while **SD** still **provide shape prior**
- **(3)** Challenging cases: **SD features** generate **smooth** correspondences and have a strong sense of **spatial layout**, but **inaccurate pixel level matching**.
- **DINOv2** generates **sparse** but **accurate matches**.

Simple Fusion Strategy to Leverage the Complementary

- $\mathcal{F}_{\text{FUSE}} = (\alpha||\mathcal{F}_{\text{SD}}||_2,\ (1-\alpha)||\mathcal{F}_{\text{DINO}}||_2)$; visualization with different α:

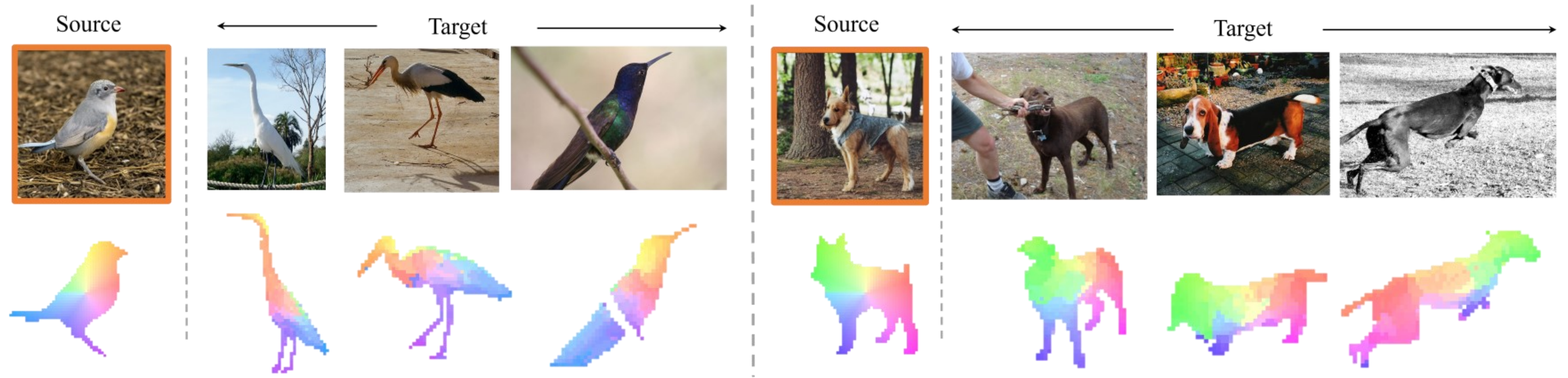


- The fused representation can utilize the strengths of both feature via simply **normalizing and concatenating** the two; α is set to 0.5 for optimal balance

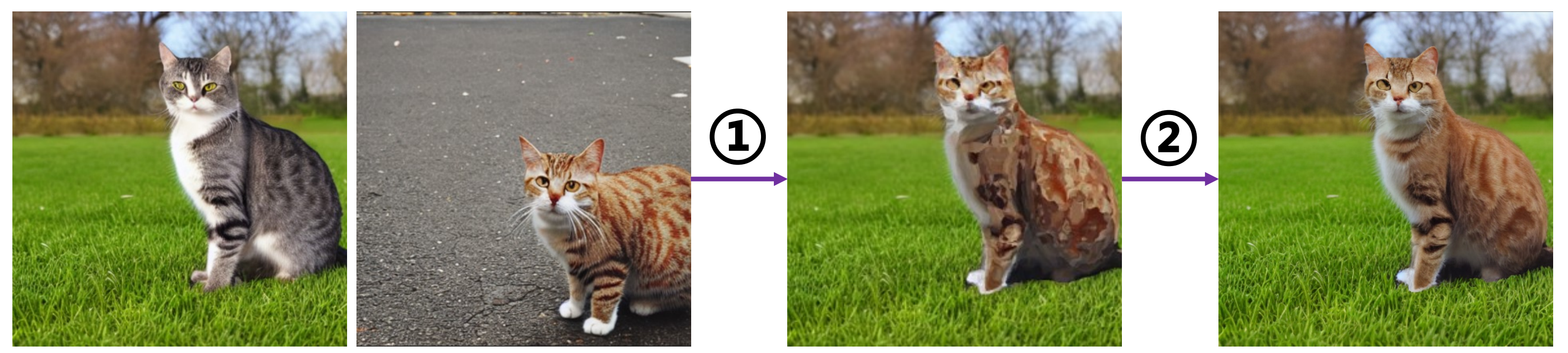
Quantitative Result – PCK@0.10 on Spair-71k

	Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dog	Horse	Motor	Person	Plant	Sheep	Train	TV	**All**
U^N	DINOv1-ViT-S/8	57.2	24.1	67.4	24.5	26.8	29.0	27.1	52.1	15.7	42.4	43.3	30.1	23.2	40.7	16.6	24.1	31.0	24.9	33.3
	DINOv2-ViT-B/14	72.7	62.0	85.2	**41.3**	40.4	52.3	51.5	71.1	36.2	67.1	64.6	67.6	61.0	68.2	30.7	62.0	54.3	24.2	55.6
	Stable Diffusion (**Ours**)	63.1	55.6	80.2	33.8	44.9	49.3	47.8	74.4	38.4	70.8	53.7	61.1	54.4	55.0	54.8	53.5	65.0	53.3	57.2
	Fuse-ViT-B/14 (**Ours**)	**73.0**	**64.1**	**86.4**	40.7	**52.9**	**55.0**	**53.8**	**78.6**	**45.5**	**77.3**	**64.7**	**69.7**	**63.3**	**69.2**	**58.4**	**67.6**	**66.2**	**53.5**	**64.0**

Qualitative Result - Dense Correspondence



Qualitative Result - Instance Swapping



①: initial swapping based on dense correspondence
②: inversion-based refinement process