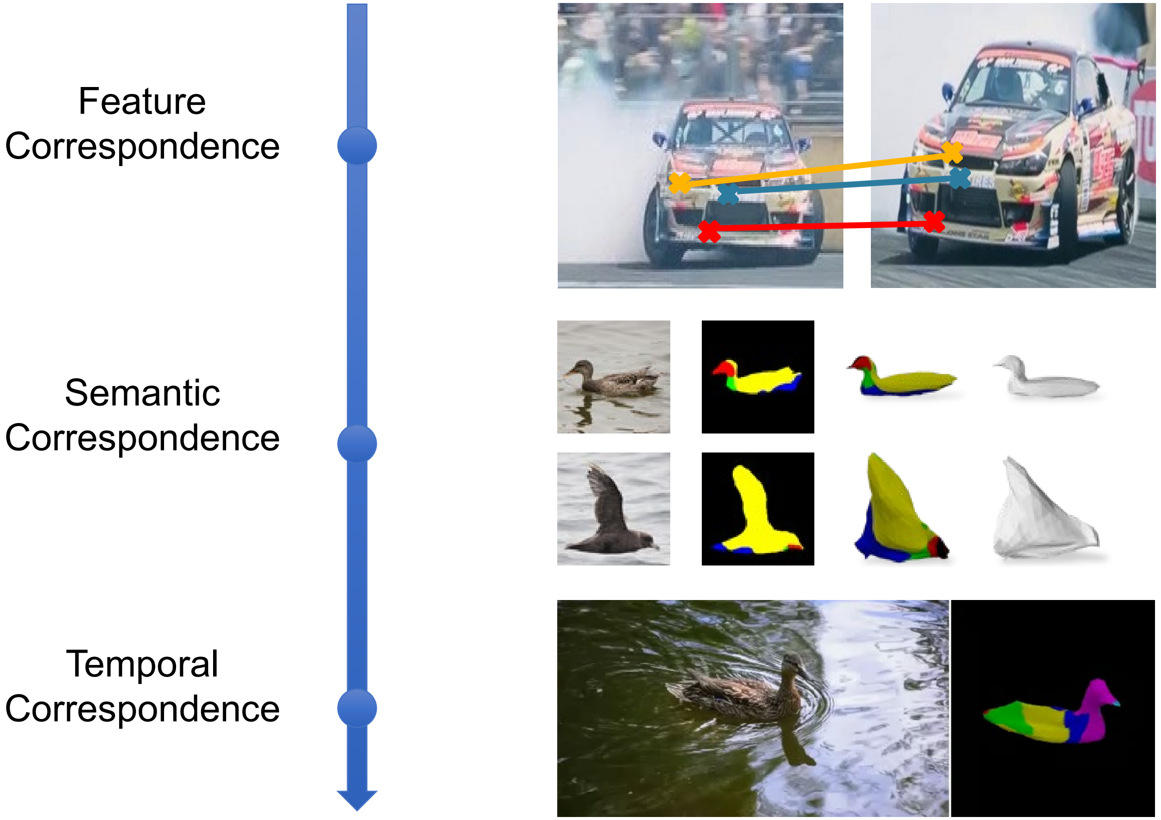
Recent Results on Learning with Diffusion Models

Ming-Hsuan Yang

UC Merced Google

Visual Correspondence



Feature Correspondence

Semantic Correspondence

Temporal Correspondence

Li et al. Joint-task Self-supervised Learning for Temporal Correspondence, NuerIPS 19

Xiao et al. Learning Contrastive Representation for Semantic Correspondence, IJCV 22

Tsai et al. Learning to Adapt Structured Output Space for Semantic Segmentation, CVPR 18

Hung et al. SCOPS: Self-Supervised Co-Part Segmentation, CVPR 19

Li et al. Online Adaptation for Consistent Mesh Reconstruction in the Wild, NuerIPS 20

Diffusion Models

- Text-to-image/video synthesis
    - Imagen, stable diffusion, DreamBooth, diffusion transformer, SORA
- Text-to-3D
    - Dreambooth3D, DreamFusion, Magic3D, ProlificDreamer, Score Jacobian Chaining
- 3D shape
    - 3DiM, latent diffusion model, SparseFusion, GeNVS, BANMo, Zero-1-2-3
- Semantic correspondence
    - Diffusion features (DIFT)
- Segmentation
    - Open vocabular panoptic segmentation (ODISE)
- L. Yang et al. Diffusion models: A Comprehensive Survey of Methods and Applications. ACM Computing Surveys, 2024.

Learning with Diffusion Models

- Exploiting diffusion features for correspondence (NeurIPS 2023)
- Geometric-aware semantic correspondence (CVPR 2024)
- Dense prediction with diffusion prior (CVPR 2024)
- DreaMo: Using diffusion prior for 3D reconstruction (arxiv 2024)

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

NuerIPS 2023

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

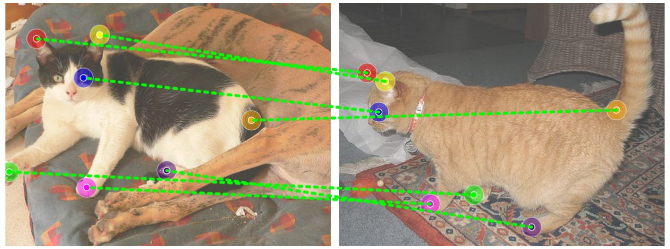
[1] Shanghai Jiao Tong University [2] Google Research [3] UC Merced

https://sd-complements-dino.github.io

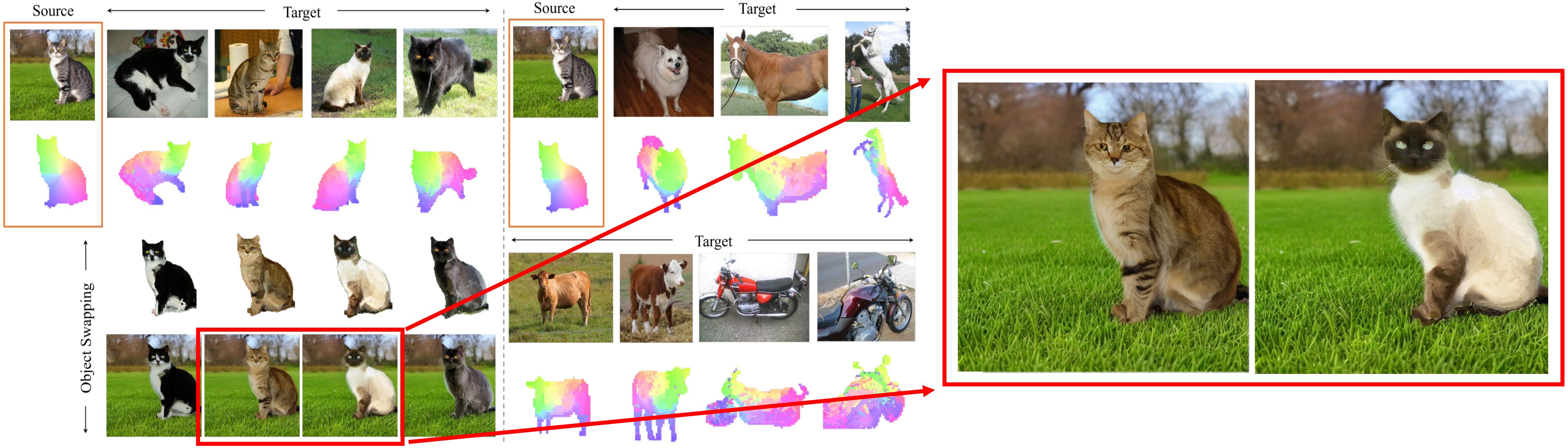
All about Features

- … -> SIFT -> CNN -> Vision Transformers -> Diffusion Models
- CNN features
  - Layers, visualization
  - Exploit features from different layers
- Response maps
  - Classification activation map (CAM)
  - ViT features
  - DINO ViT
- Properties of CNNs and vision transformers
  - Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, ICLR 2018
  - Shape and texture bias for visual recognition based on CNN features, ICLR 2021
  - Intriguing properties of vision transformers. NeurIPS 2021

Overview

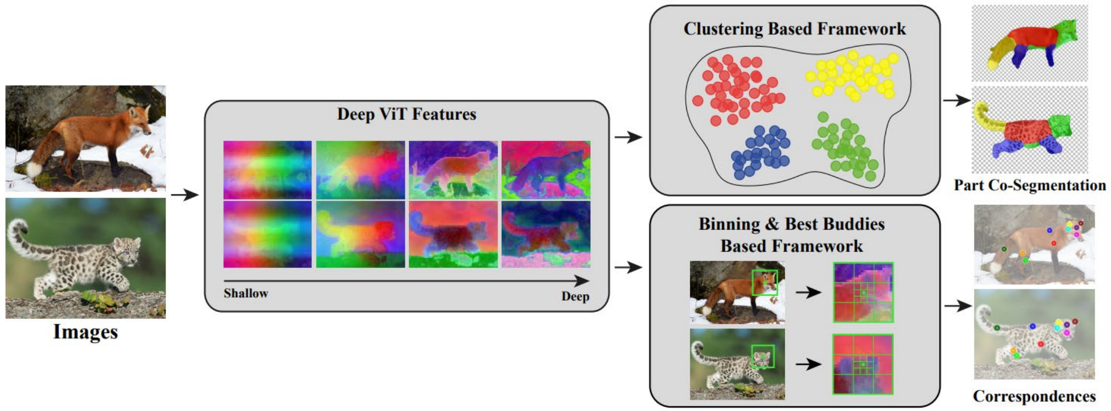


- Semantic correspondence: pixel-level semantic matching
- Empirical study of two recent representations (Stable Diffusion, DINOv2) for zero-shot semantic correspondence
- Explore complementary features and achieve surprising performance on benchmark datasets (70.5% relative improvement over SOTA on SPair-71k)

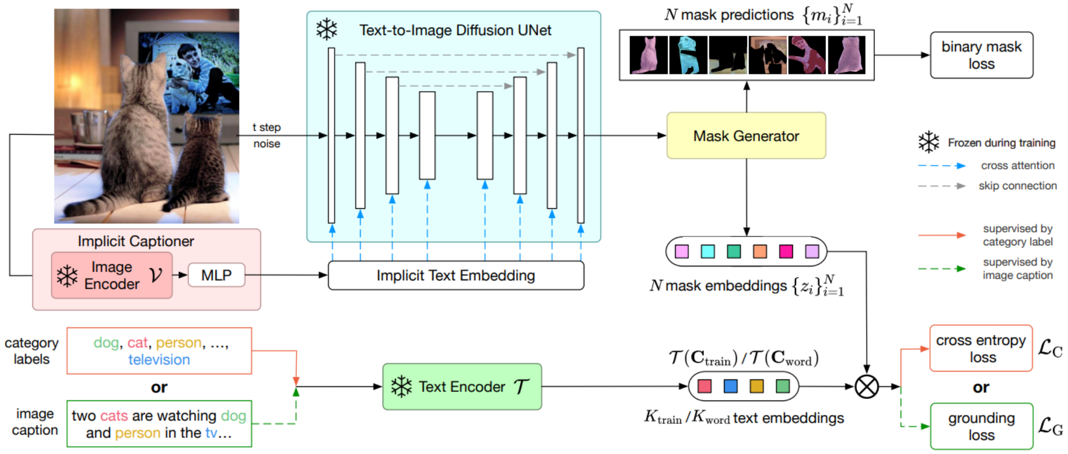


Motivation

- Two lines of work



DINO features for zeo-shot semantic correspondence [1]


Text-to-Image Diffusion UNet
t step noise
N mask predictions $\{m_i\}_{i=1}^N$
binary mask loss
Mask Generator
Frozen during training
cross attention
skip connection
supervised by category label
supervised by image caption
Implicit Captioner
Image Encoder $\mathcal{V}$
MLP
Implicit Text Embedding
N mask embeddings $\{z_i\}_{i=1}^N$
category labels
dog, cat, person, ..., television
or
image caption
two cats are watching dog and person in the tv...
Text Encoder $\mathcal{T}$
$\mathcal{T}(\mathbf{C}_{\text{train}})/\mathcal{T}(\mathbf{C}_{\text{word}})$
$K_{\text{train}}/K_{\text{word}}$ text embeddings
cross entropy loss
$\mathcal{L}_C$
or
grounding loss
$\mathcal{L}_G$

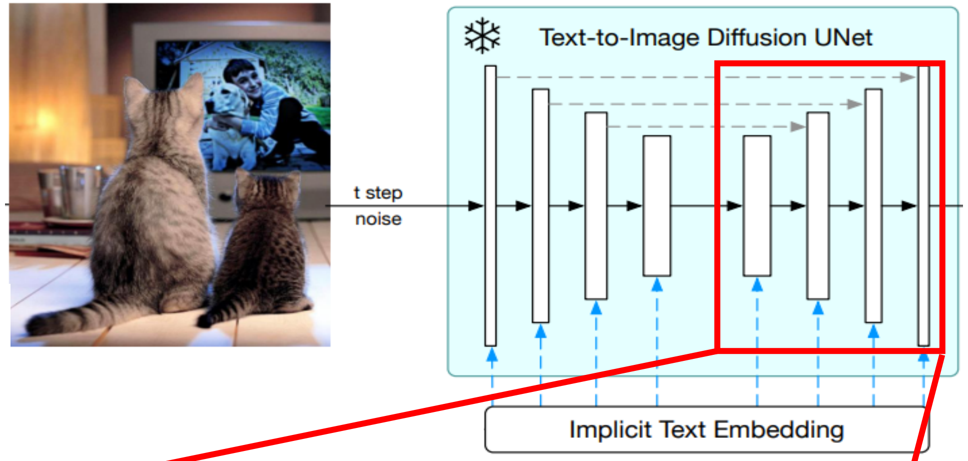
Diffusion features for panoptic segmentation (ODISE) [2]

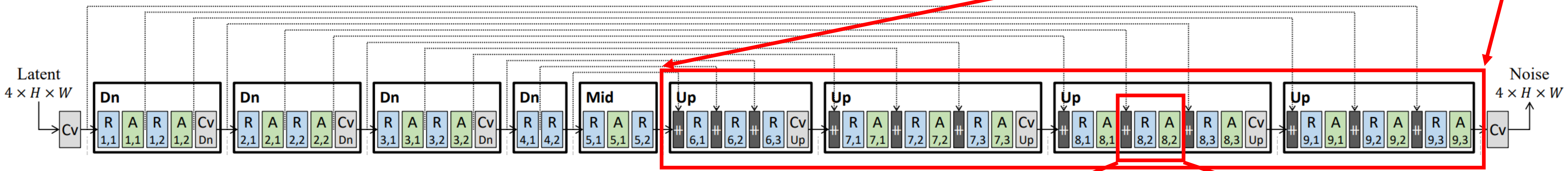
- Motivation:
  - Text-to-image generative models can generate instances of varying poses, appearances, and at different scenes → suggesting its potential of fine-grained, cross-image semantic understanding (not only instance-level information)
  - Properties of different layers, timesteps; comparison with other representations

[1] Amir et al. Deep vit features as dense visual descriptors. arXiv, 2021
[2] Xu et al. Open-vocabulary panoptic segmentation with text-to-image diffusion models. CVPR 2023

Extracting SD Features

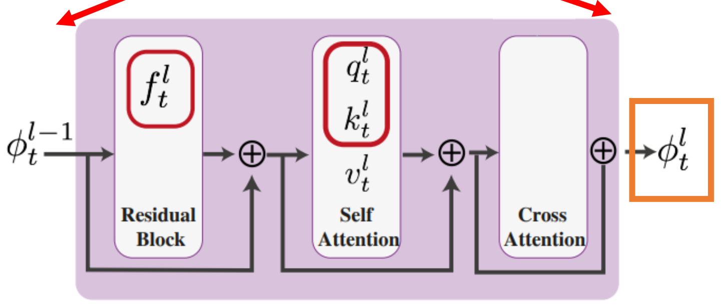
- Add noise to the given image, run a single denoising step using the latent code
- Extract features from UNet Decoder (similar to ODISE) $z_0 = \mathcal{E}(x_0), \quad z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \mathcal{F}_{\text{SD}} = \mathcal{U}(z_t, t, C),$
- 4 blocks x 3 layers/block = 12 layers



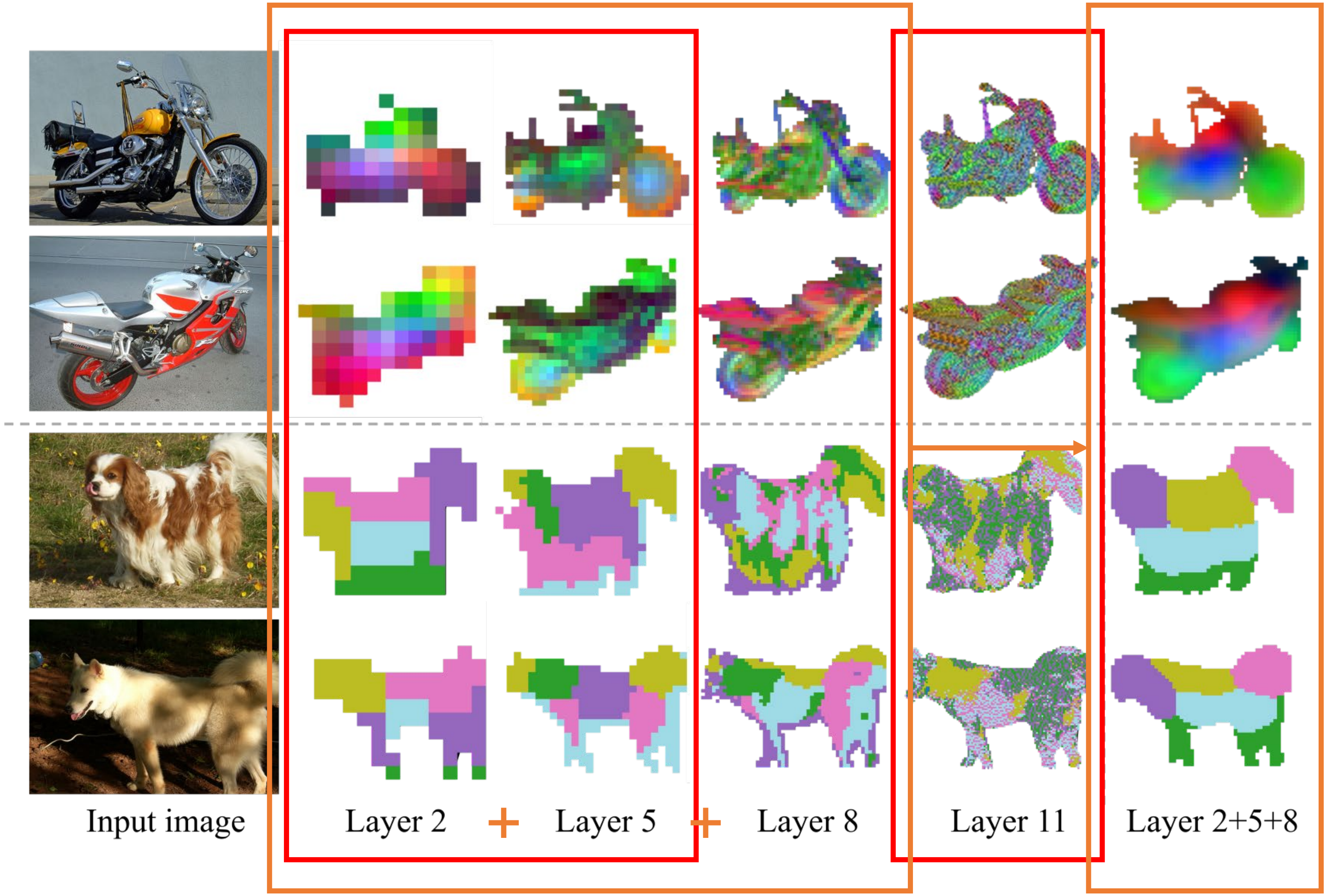


Latent $4 \times H \times W$
Cv
Dn R 1,1 A 1,1 R 1,2 A 1,2 Cv Dn
Dn R 2,1 A 2,1 R 2,2 A 2,2 Cv Dn
Dn R 3,1 A 3,1 R 3,2 A 3,2 Cv Dn
Dn R 4,1 R 4,2
Mid R 5,1 A 5,1 R 5,2
Up R 6,1 R 6,2 R 6,3 Cv Up
Up R 7,1 A 7,1 R 7,2 A 7,2 R 7,3 A 7,3 Cv Up
Up R 8,1 A 8,1 R 8,2 A 8,2 R 8,3 A 8,3 Cv Up
Up R 9,1 A 9,1 R 9,2 A 9,2 R 9,3 A 9,3
Cv
Noise $4 \times H \times W$

- Sub-layers within a decoder layer
- Feature map: output of decoder layer
- Decoder features are better than encoder features & sub-layer within decoder layer (res/attn)



Qualitative Analysis of Stale Diffusion Features

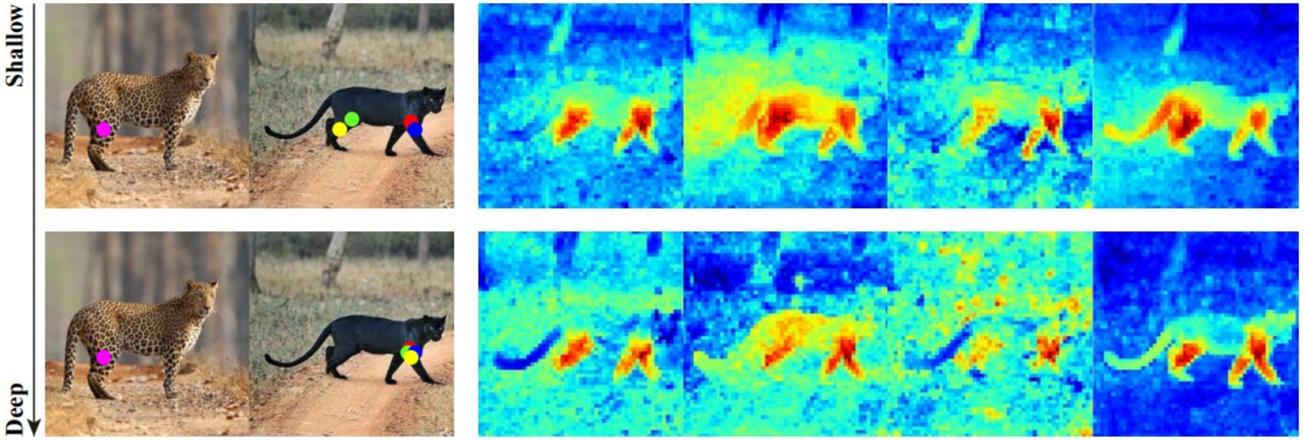


(Top) Visualization of first 3 channel of PCA features
(Bottom) Visualization of cluster & matching results

- Promising results
- Earlier layer (layer 2, 5) lower resolution and contains more semantic information;
- Last layer (layer 11) resolution is higher but focuses more on the appearance
- We ensemble features from early and intermediate layers (2, 5, 8) to trade-off between semantics and resolution, and apply co-PCA to reduce the dimension of features

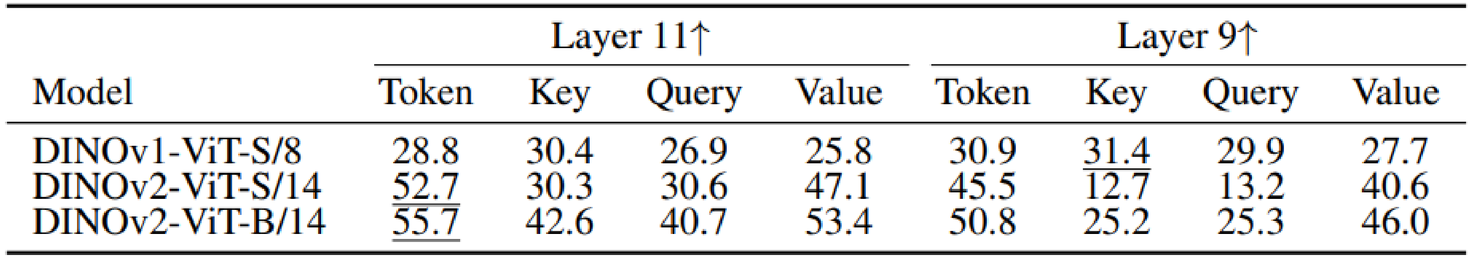
Extracting DINOv2 Features

- Different layers/facets of DINO ViT affects correspondence



(a) Source (b) Target (c) Token 🔴 (d) Query 🟢 (e) Value 🔵 (f) Key 🟡

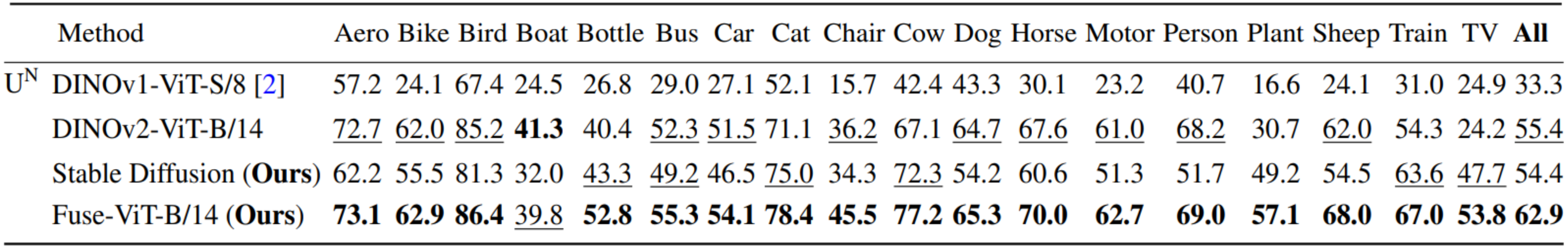
- For DINOv2, best performance is achieved by the “token” facet from the last (11th) layer, different from DINO



	Layer 11↑				Layer 9↑			
Model	Token	Key	Query	Value	Token	Key	Query	Value
DINOv1-ViT-S/8	28.8	30.4	26.9	25.8	30.9	31.4	29.9	27.7
DINOv2-ViT-S/14	52.7	30.3	30.6	47.1	45.5	12.7	13.2	40.6
DINOv2-ViT-B/14	55.7	42.6	40.7	53.4	50.8	25.2	25.3	46.0

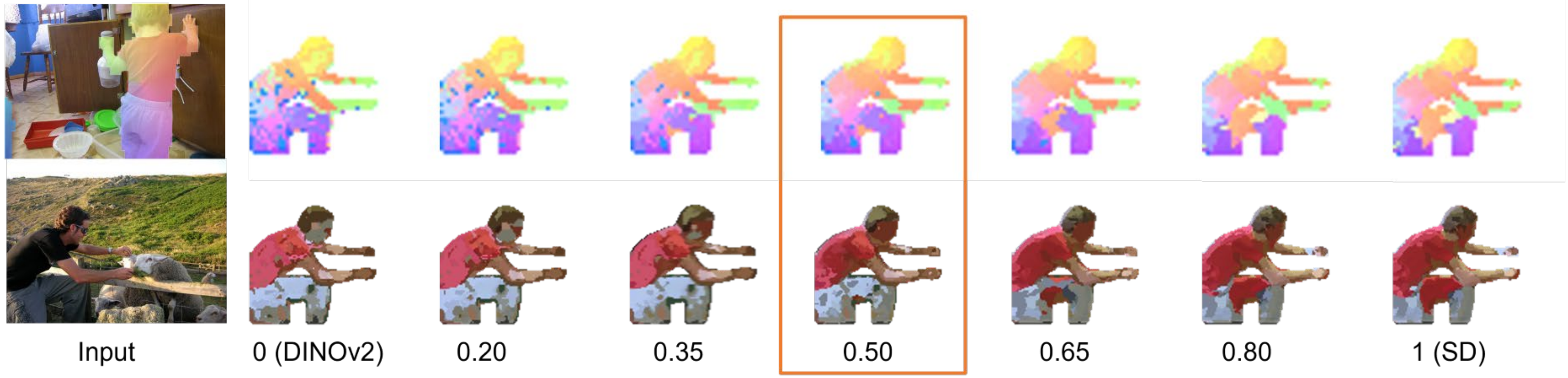
Fusing DINOv2 and SD Features

- Quantitative analysis of SD and DINOv2 features

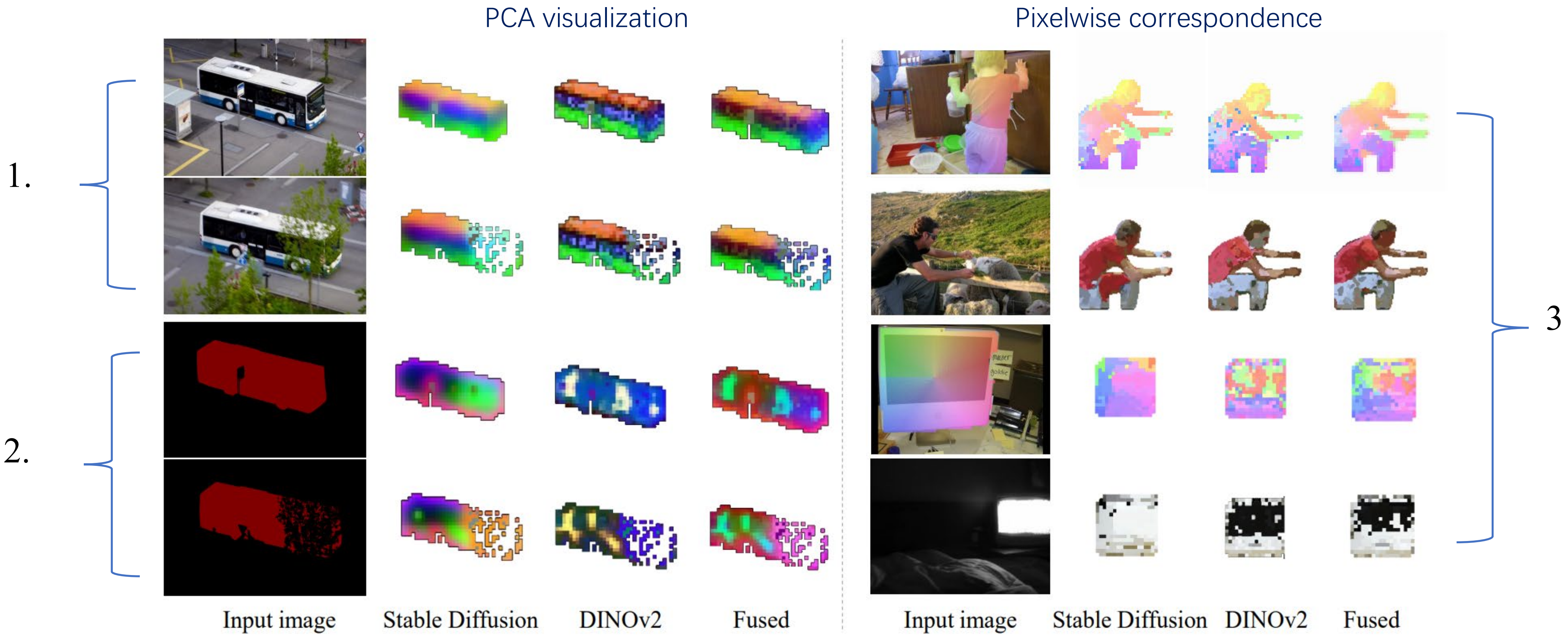


	Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dog	Horse	Motor	Person	Plant	Sheep	Train	TV	**All**
U^N	DINOv1-ViT-S/8 [2]	57.2	24.1	67.4	24.5	26.8	29.0	27.1	52.1	15.7	42.4	43.3	30.1	23.2	40.7	16.6	24.1	31.0	24.9	33.3
	DINOv2-ViT-B/14	72.7	62.0	85.2	**41.3**	40.4	52.3	51.5	71.1	36.2	67.1	64.7	67.6	61.0	68.2	30.7	62.0	54.3	24.2	55.4
	Stable Diffusion (**Ours**)	62.2	55.5	81.3	32.0	43.3	49.2	46.5	75.0	34.3	72.3	54.2	60.6	51.3	51.7	49.2	54.5	63.6	47.7	54.4
	Fuse-ViT-B/14 (**Ours**)	**73.1**	**62.9**	**86.4**	39.8	**52.8**	**55.3**	**54.1**	**78.4**	**45.5**	**77.2**	**65.3**	**70.0**	**62.7**	**69.0**	**57.1**	**68.0**	**67.0**	**53.8**	**62.9**

- Are these two features complementary?
  - $F_{Fuse} = (\alpha||F_{SD}||_2, (1-\alpha)||F_{DINOv2}||_2)$; with different α:



Fusion and Complement of Two Features



1. For easy case, both two features can find plausible correspondence

2. When textual signal is absent, DINOv2 fails while SD still capture shape prior

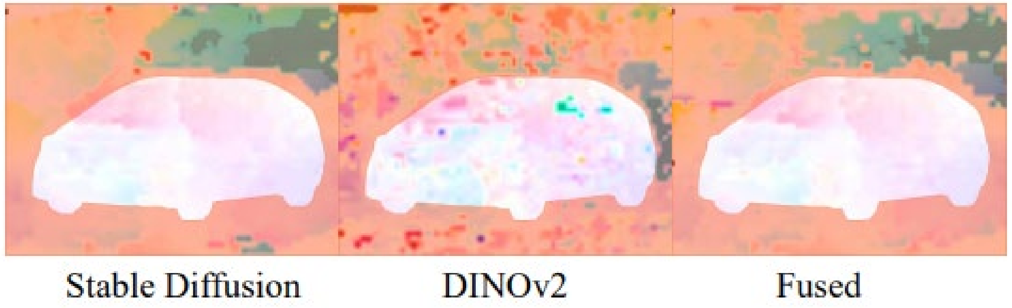
3. For challenging cases

- SD features generate *smooth* correspondences and have strong sense of *spatial layout*, but obtain inaccurate pixel level matching whereas

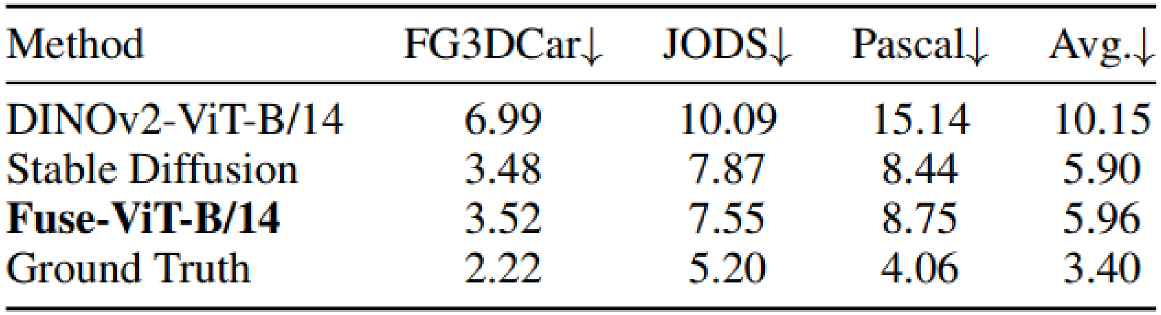
DINOv2 generates sparse but accurate matches.

Quantitative Analysis on Fusion

- Smoothness of semantic flow fields on TSS dataset



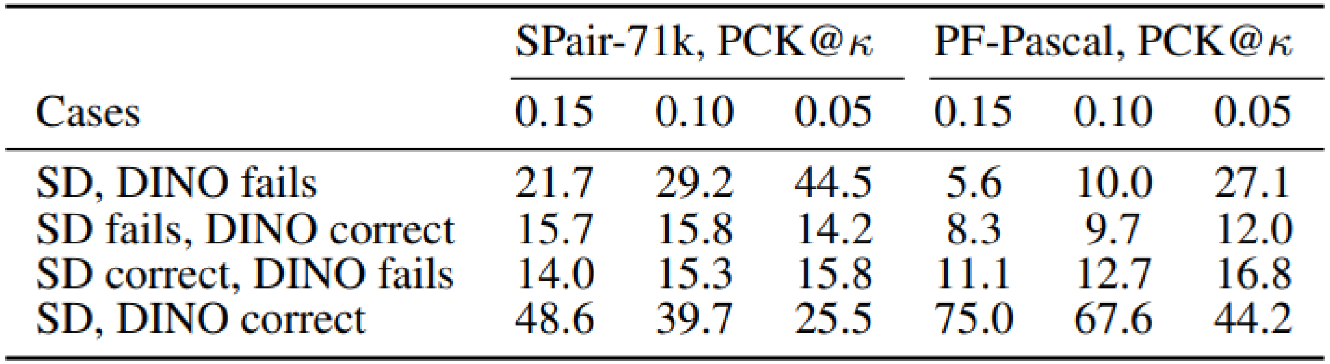
I. Sample visualization of the flow fields



Method	FG3DCar↓	JODS↓	Pascal↓	Avg.↓
DINOv2-ViT-B/14	6.99	10.09	15.14	10.15
Stable Diffusion	3.48	7.87	8.44	5.90
Fuse-ViT-B/14	3.52	7.55	8.75	5.96
Ground Truth	2.22	5.20	4.06	3.40

II. First-order difference (lower indicates smoother)

- Non-redundancy of SD and DINOv2 features



	SPair-71k, PCK@κ			PF-Pascal, PCK@κ		
Cases	0.15	0.10	0.05	0.15	0.10	0.05
SD, DINO fails	21.7	29.2	44.5	5.6	10.0	27.1
SD fails, DINO correct	15.7	15.8	14.2	8.3	9.7	12.0
SD correct, DINO fails	14.0	15.3	15.8	11.1	12.7	16.8
SD, DINO correct	48.6	39.7	25.5	75.0	67.6	44.2

III. Distribution of different outcomes (under 2 datasets and 3 PCK levels)

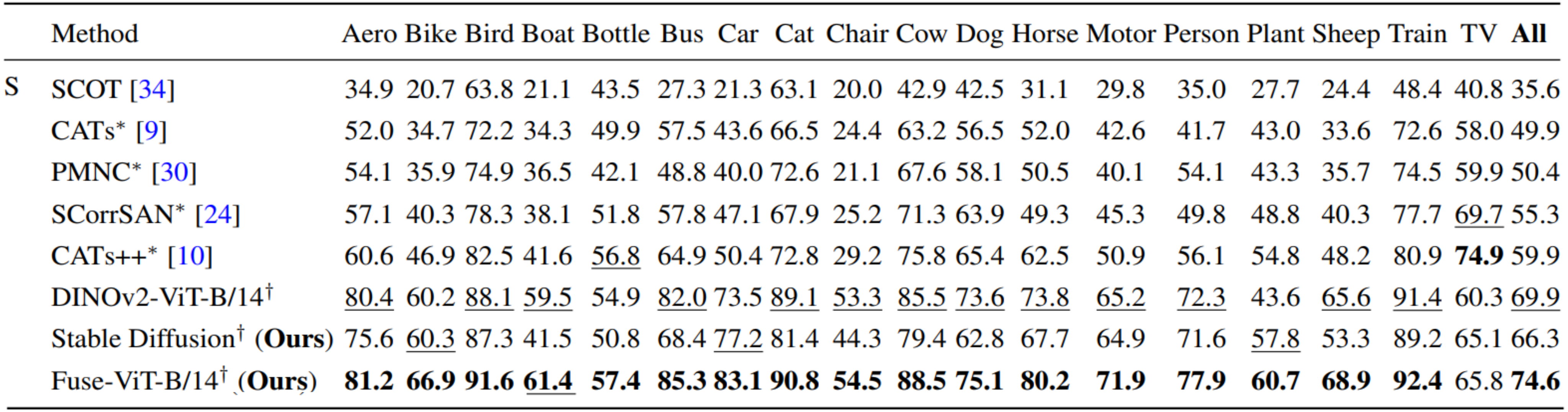
- In most settings, one feature succeeds while the other fails in 20~30% of total cases (row 2 and 3)
- SD and DINOv2 have a substantial amount of non-redundant information.

Discussions on Feature Behavior

- Causes for distinct behaviors of SD and DINOv2 feature maps:
  - Training paradigm:
    - Text-to-image synthesis -- Self-supervised learning
  - Training data:
    - Text-image pairs -- Image only
  - Architecture:
    - Conv-based UNet -- ViT
- More discussions in our paper: different SD and DINO variants
- *What else can we do with the observation?*
- Offset the limitations of DINOv2 features with
  - Bilateral filter -> spatial coherence
  - Ensemble early layer -> spatial awareness

Supervised Setting

- Train a *bottleneck layer* on top of the extracted features,
- Guided by the CLIP-style *symmetric cross entropy loss* with respect to *corresponding keypoints* [1] on Spair-71k



	Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dog	Horse	Motor	Person	Plant	Sheep	Train	TV	**All**
S	SCOT [34]	34.9	20.7	63.8	21.1	43.5	27.3	21.3	63.1	20.0	42.9	42.5	31.1	29.8	35.0	27.7	24.4	48.4	40.8	35.6
	CATs* [9]	52.0	34.7	72.2	34.3	49.9	57.5	43.6	66.5	24.4	63.2	56.5	52.0	42.6	41.7	43.0	33.6	72.6	58.0	49.9
	PMNC* [30]	54.1	35.9	74.9	36.5	42.1	48.8	40.0	72.6	21.1	67.6	58.1	50.5	40.1	54.1	43.3	35.7	74.5	59.9	50.4
	SCorrSAN* [24]	57.1	40.3	78.3	38.1	51.8	57.8	47.1	67.9	25.2	71.3	63.9	49.3	45.3	49.8	48.8	40.3	77.7	69.7	55.3
	CATs++* [10]	60.6	46.9	82.5	41.6	56.8	64.9	50.4	72.8	29.2	75.8	65.4	62.5	50.9	56.1	54.8	48.2	80.9	**74.9**	59.9
	DINOv2-ViT-B/14†	80.4	60.2	88.1	59.5	54.9	82.0	73.5	89.1	53.3	85.5	73.6	73.8	65.2	72.3	43.6	65.6	91.4	60.3	69.9
	Stable Diffusion† **(Ours)**	75.6	60.3	87.3	41.5	50.8	68.4	77.2	81.4	44.3	79.4	62.8	67.7	64.9	71.6	57.8	53.3	89.2	65.1	66.3
	Fuse-ViT-B/14† **(Ours)**	**81.2**	**66.9**	**91.6**	**61.4**	**57.4**	**85.3**	**83.1**	**90.8**	**54.5**	**88.5**	**75.1**	**80.2**	**71.9**	**77.9**	**60.7**	**68.9**	**92.4**	65.8	**74.6**

- Improvement over previous methods (*: fine-tuned backbone)
- Effectiveness of fusion also applies to supervised setting

[1] Luo et al. Diffusion hyperfeatures: searching through time and space for semantic correspondence. NeurIPS, 2023.

Application: Instance Swapping

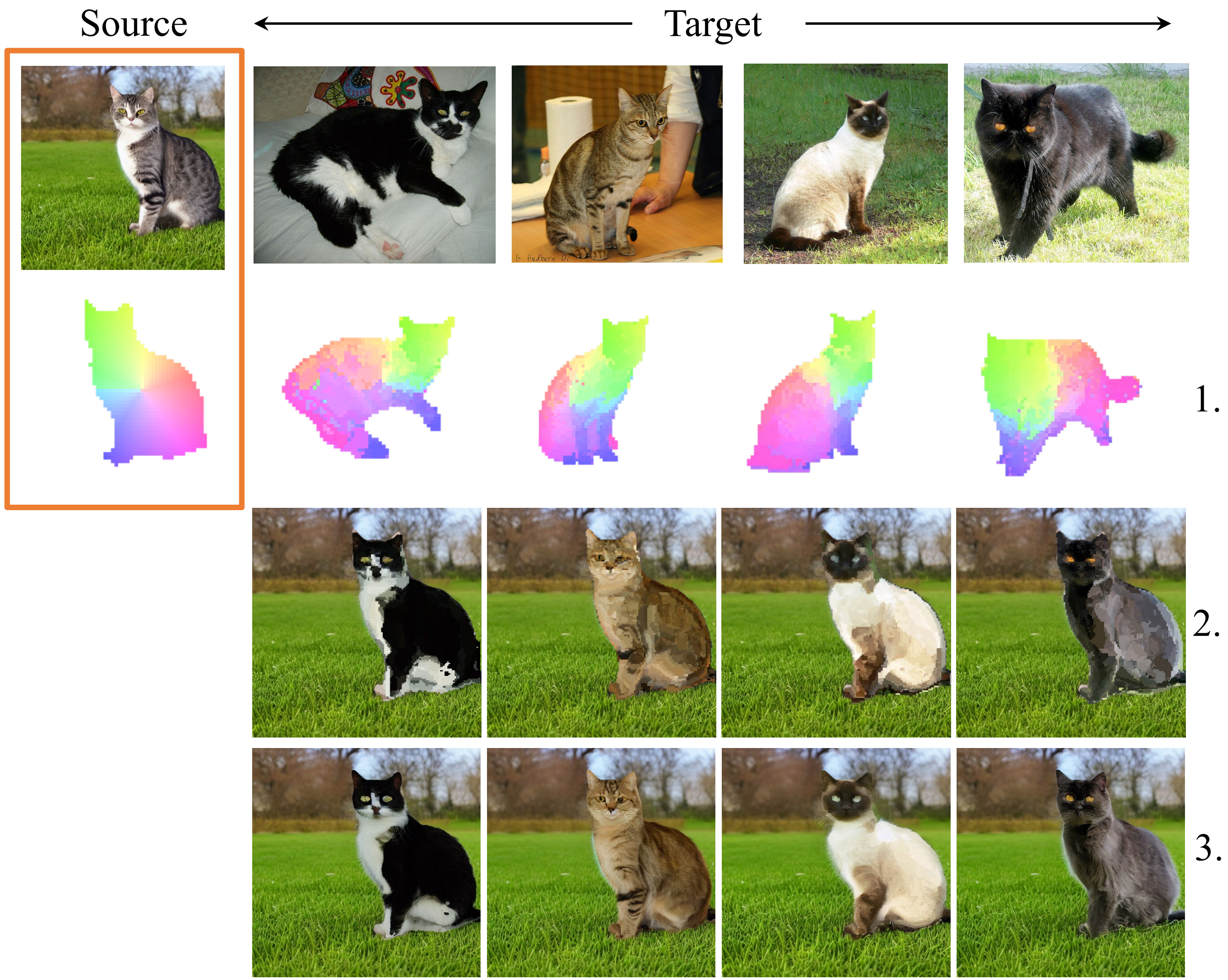
Given the source and target images

1. Build dense correspondence map with the extracted features (same color indicates matching)

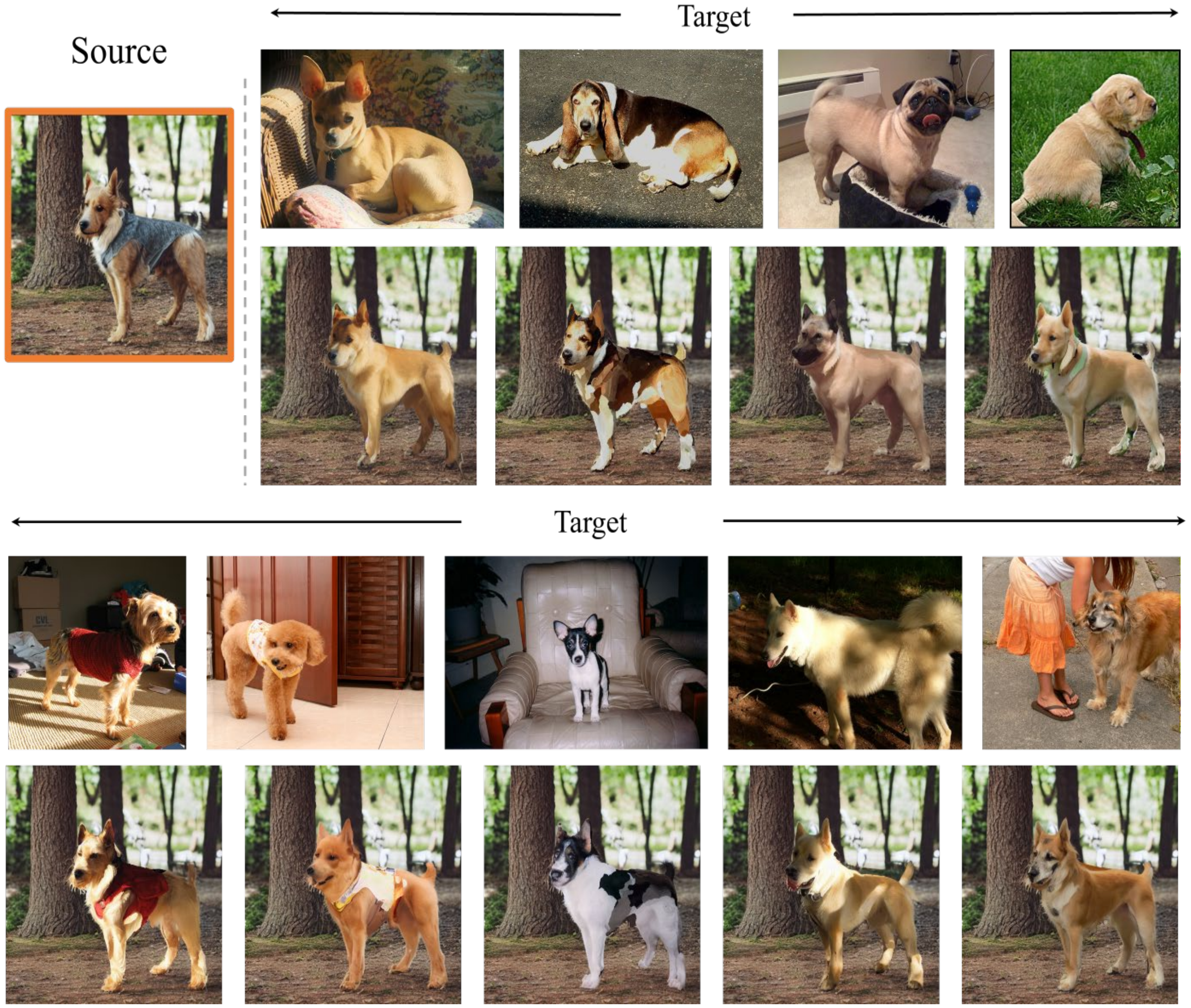
2. Initial swapping result with the correspondence map

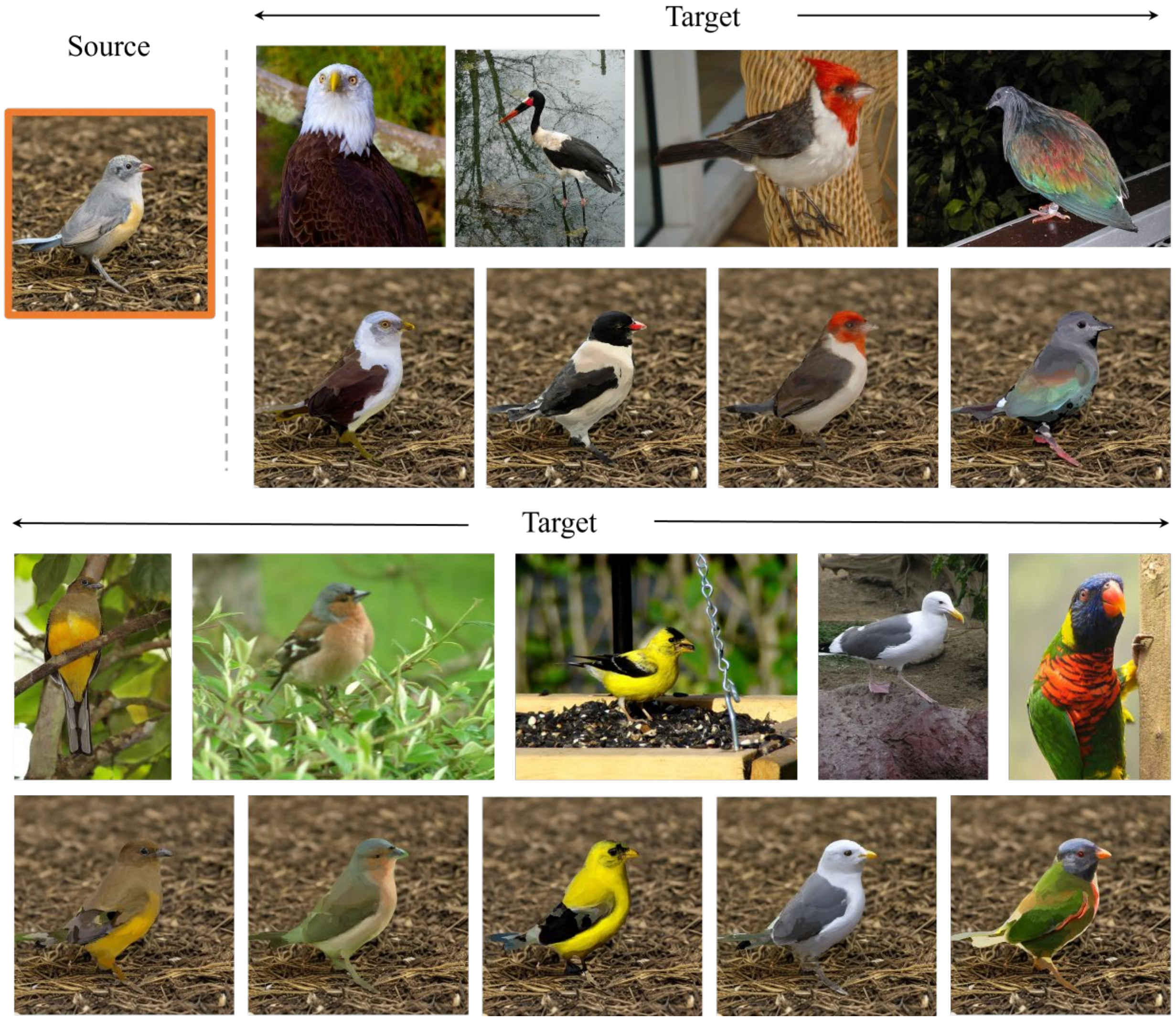
3. A diffusion-based refinement process yields more plausible result

    - Invert the initial image with DDIM inversion
    - Run DDIM denoising sampling to extract the spatial features of each timestep
    - Generate the refined image with the prompt “A high quality image of [CAT]” where we also inject the extracted spatial features

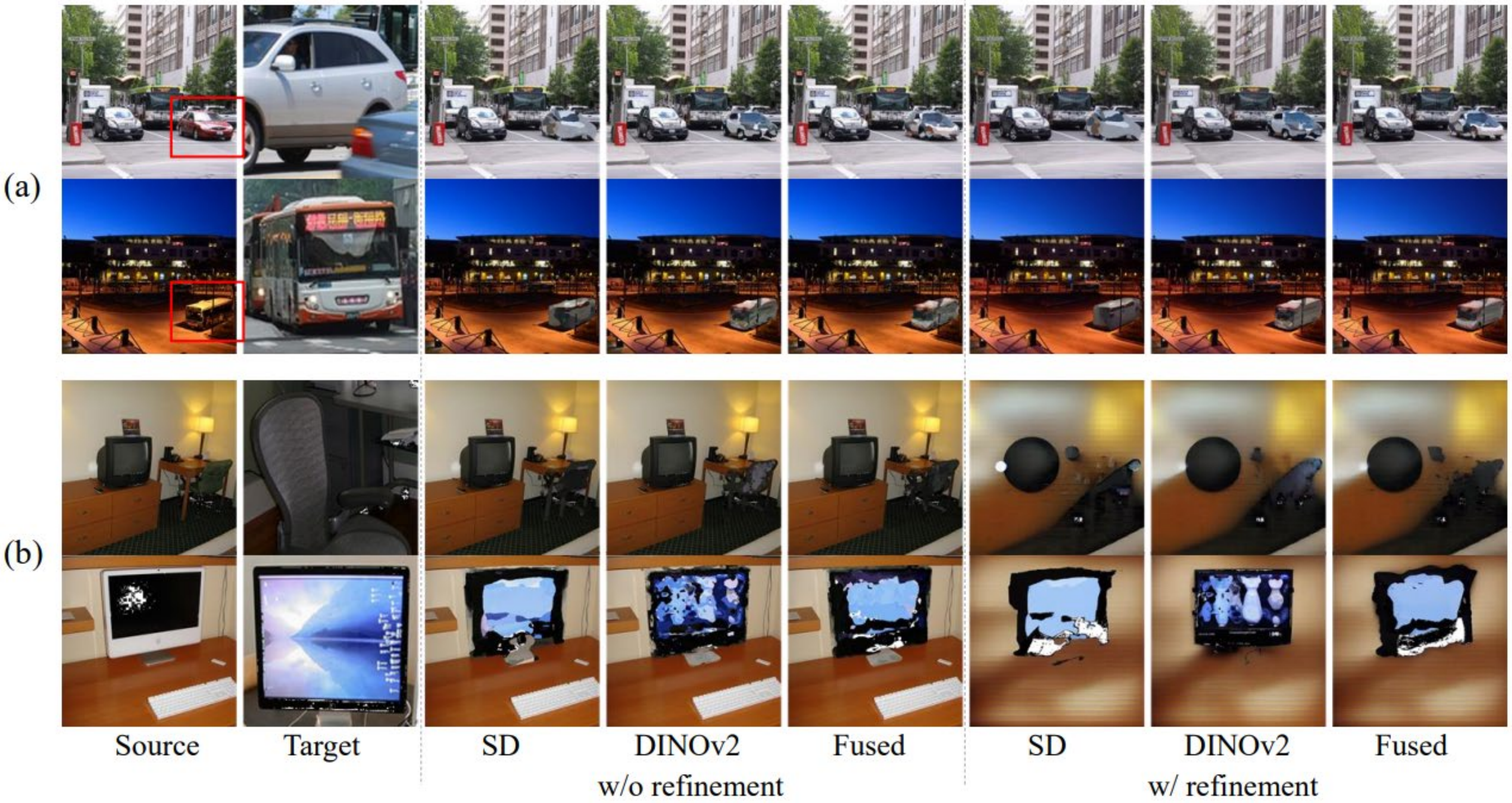


Instance Swapping





Instance Swapping (Failure Cases)

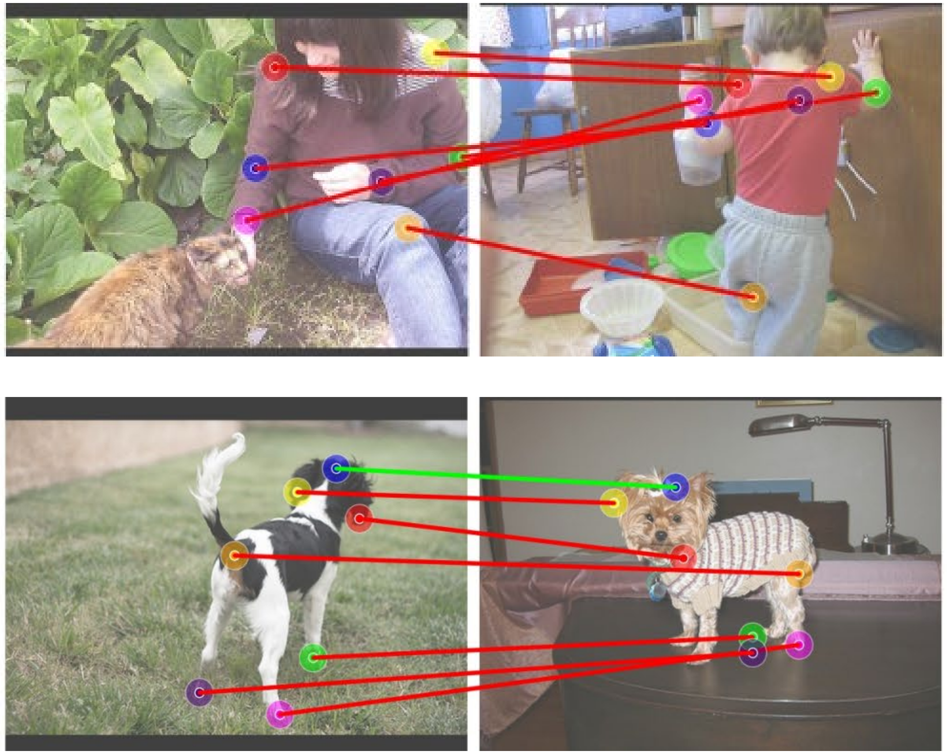


(a) The relative size of instance of interest is tiny in the image (limited by the resolution of the extracted features)

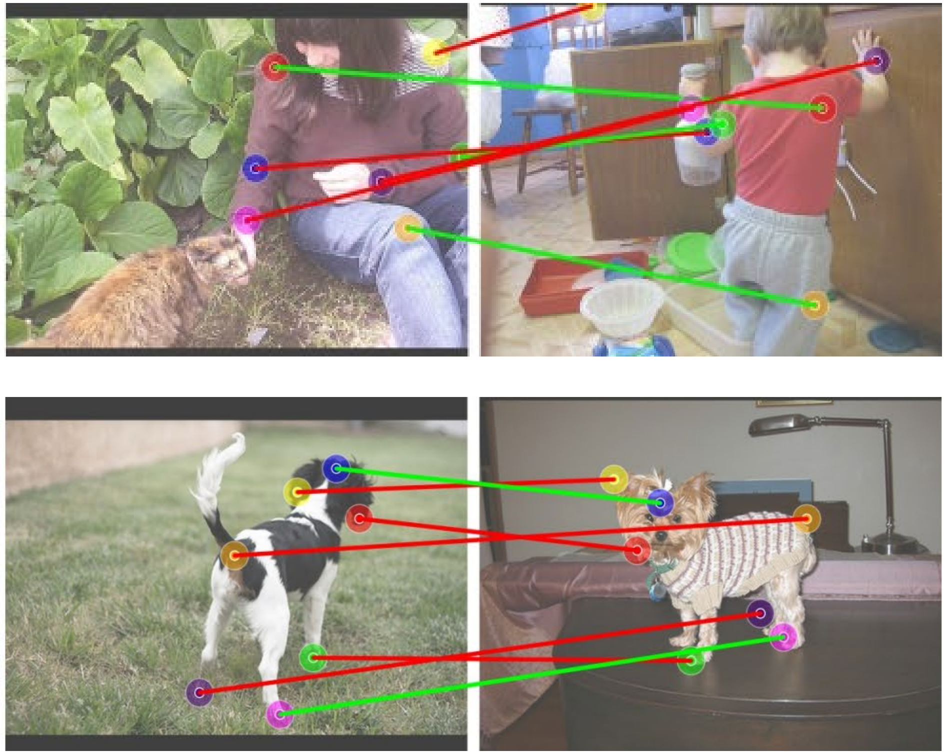
(b) Artifacts introduced by DDIM inversion

Limitations

- Speed (1s per image on single RTX 3090 GPU)
- Resolution (1/16 of the input resolution, not flexible)
- Sometimes struggle with semantically similar points (especially under large viewpoint variation)



SD+DINO



SD+DINO (Supervised)

Summary

- Stable Diffusion (SD) features shows great potential for semantic and dense correspondence, on par with SOTA self-supervised learning representations (DINOv2)

- Two features have different properties and naturally complement each other, and a simple fusion strategy can achieve the best of both worlds

- Significant improvement of semantic correspondence over previous SOTA on both zero-shot and supervised settings

- Instance swapping with high-quality correspondence

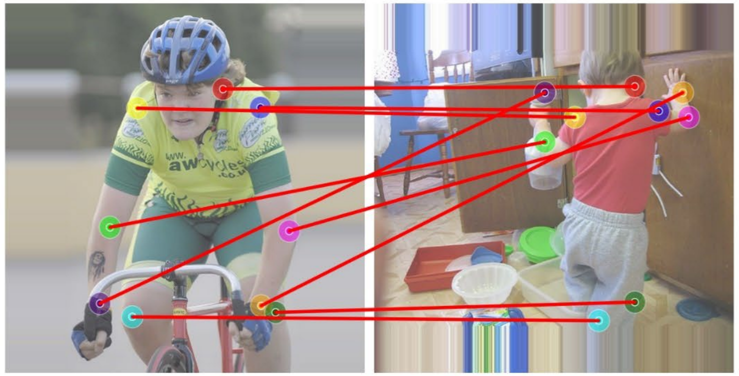
Telling Left from Right: Identifying Geometry-Aware Semantic Correspondence

arxiv 2023

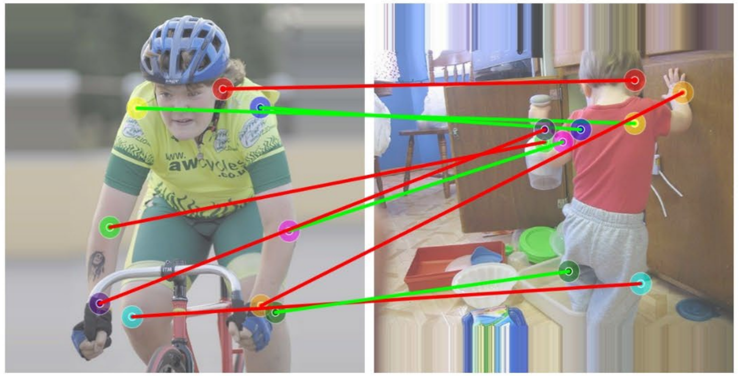
Junyi Zhang[1] **Charles Herrmann**[2] **Junhwa Hur**[2] **Eric Chen**[3]
Varun Jampani[4] **Deqing Sun**[2] **Ming-Hsuan Yang**[2,5]

[1] Shanghai Jiao Tong University [2] Google Research [3] UIUC [4] Stability AI [5] UC Merced

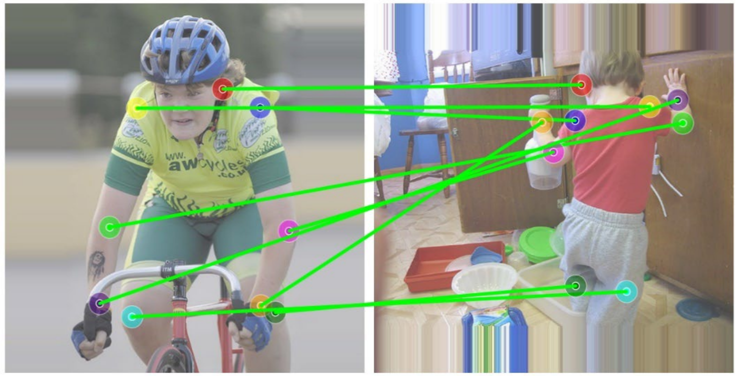
https://telling-left-from-right.github.io



SD+DINO



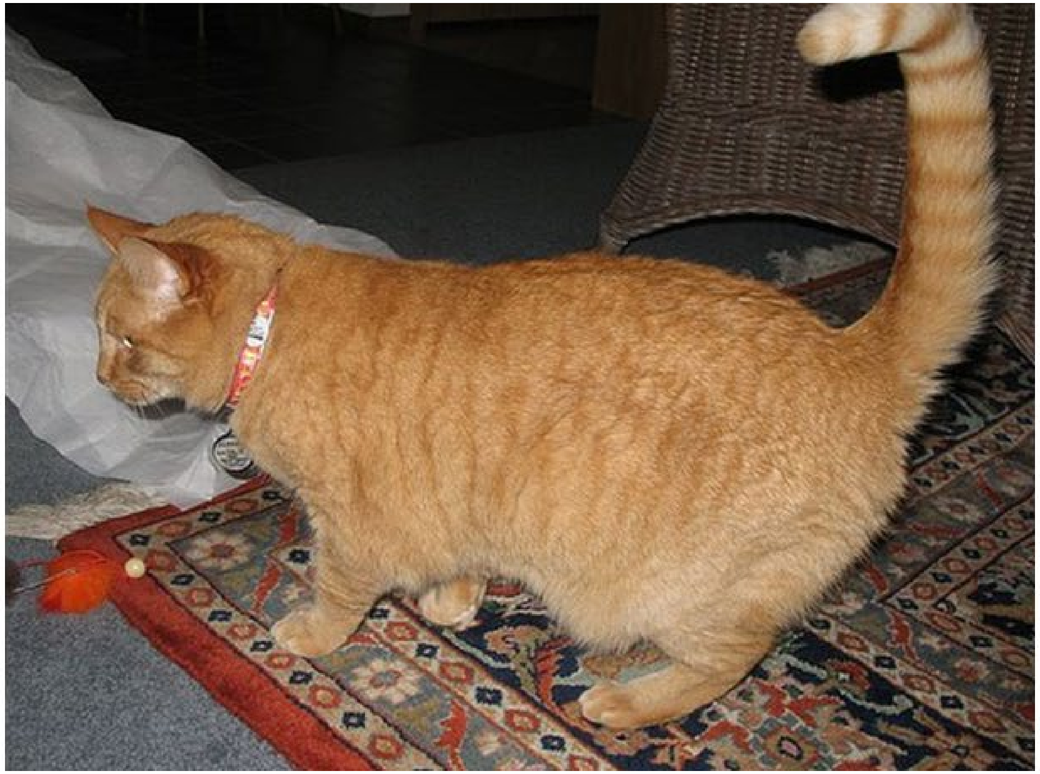
SD+DINO (Supervised)



Ours

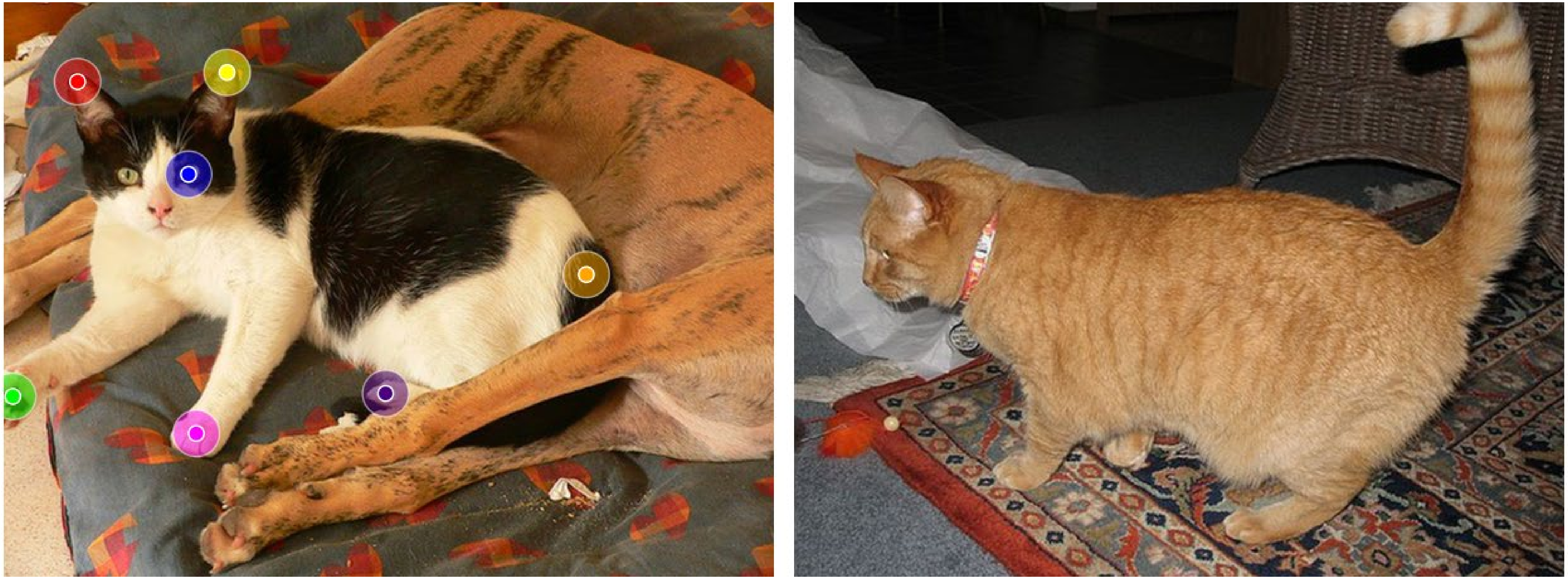
Background

- Semantic correspondence (definition)



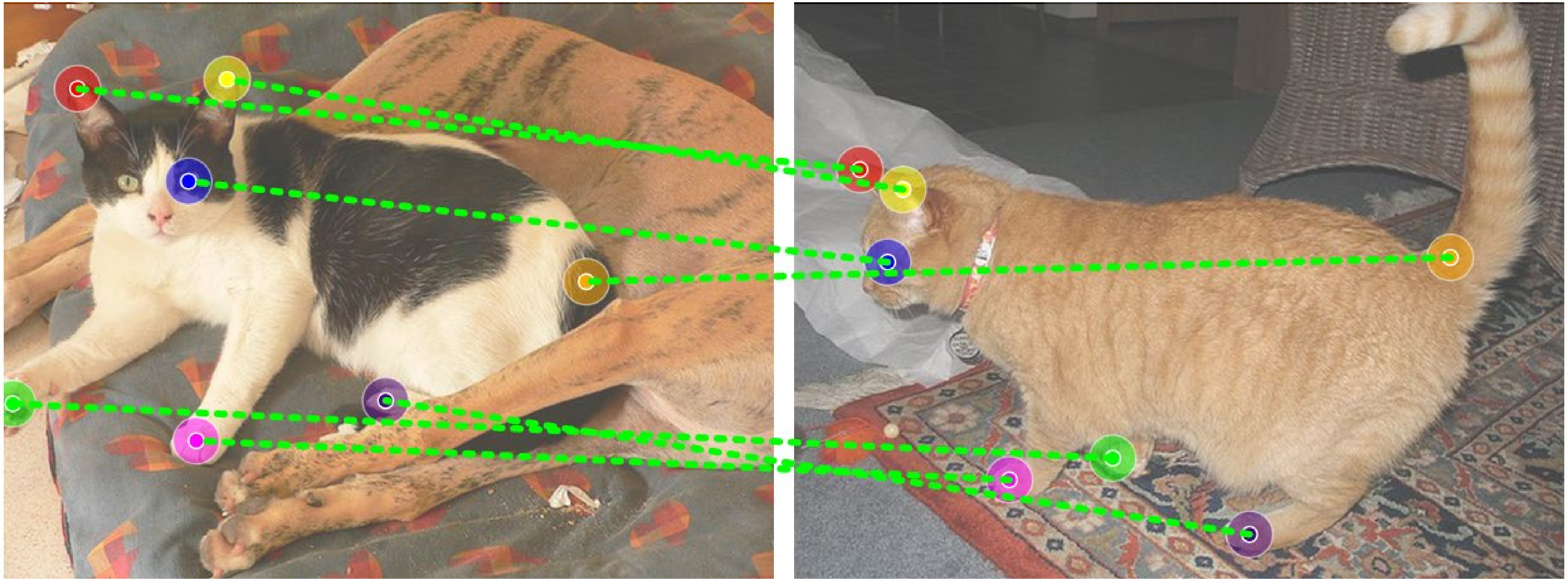
Background

- Semantic correspondence (definition)



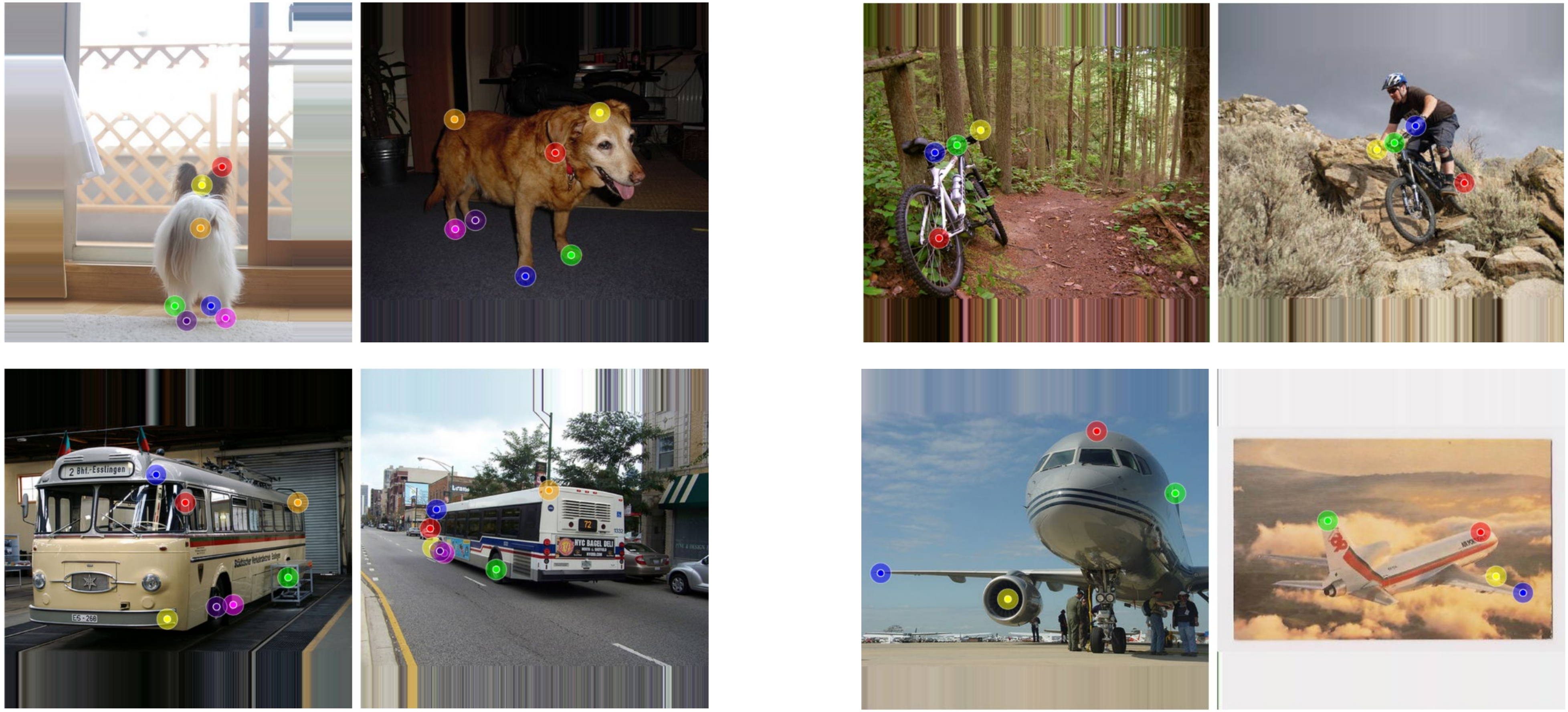
Background

- Semantic correspondence



Background

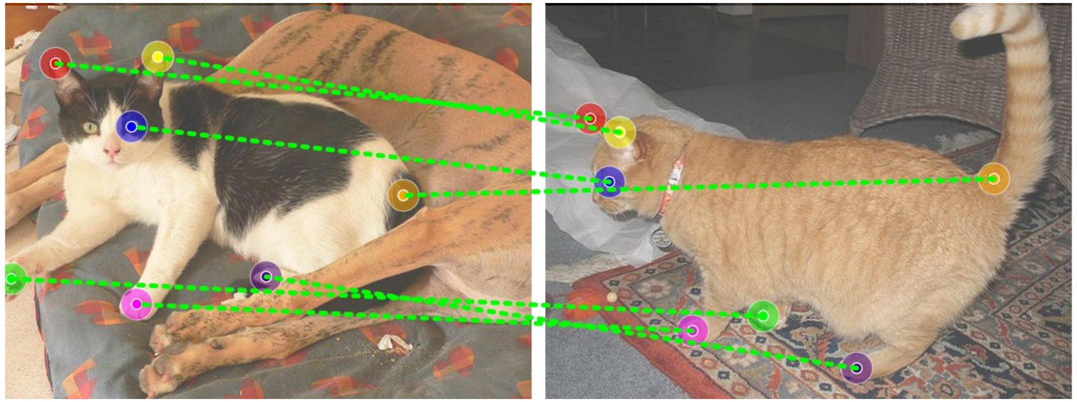
- Challenges in semantic correspondence

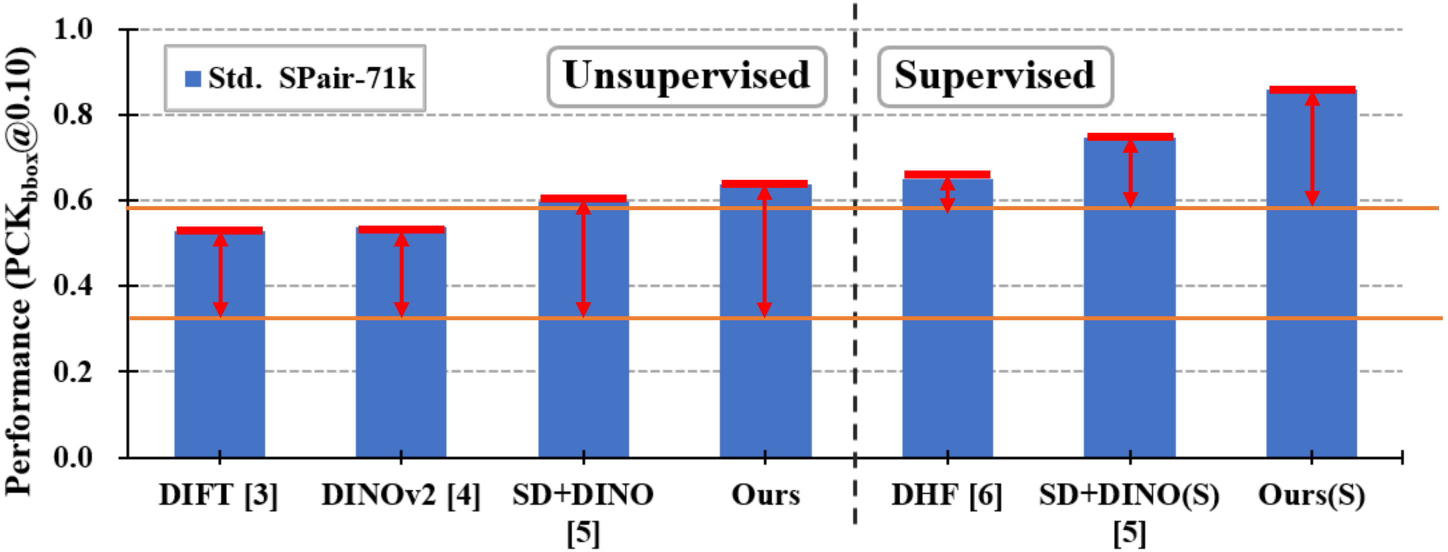


from large intra-class variations to different backgrounds, lighting, or viewpoints.

Background

- Semantic correspondence
- Pre-trained vision models





CATs++ [2], supervised SOTA

ASIC [1], unsupervised SOTA

[1] ASIC: Aligning Sparse in-the-wild Image Collections. *ICCV*, 2023.

[2] CATs++: Boosting Cost Aggregation With Convolutions and Transformers. *TPAMI*, 2022.

[3] Emergent Correspondance from Image Diffusion. *NeurIPS*, 2023.

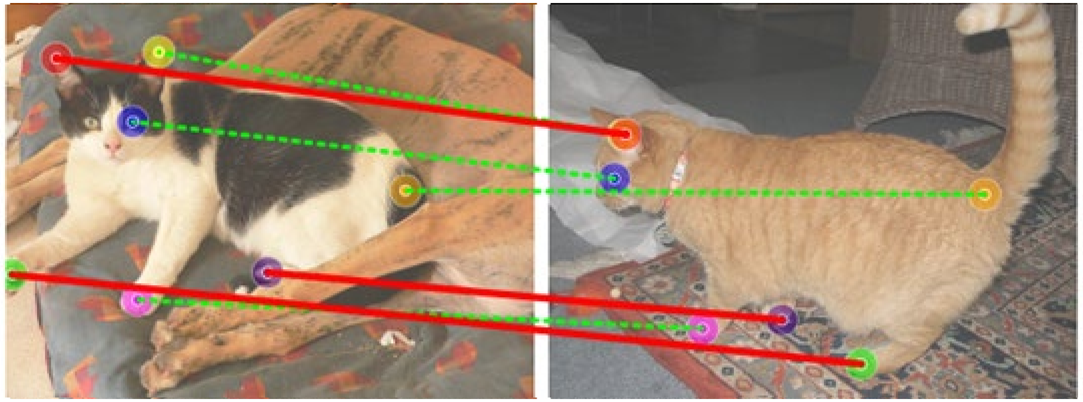
[4] DINOv2: Learning Robust Visual Features without Supervision. *Arxiv*, 2023.

[5] A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence. *NeurIPS*, 2023.

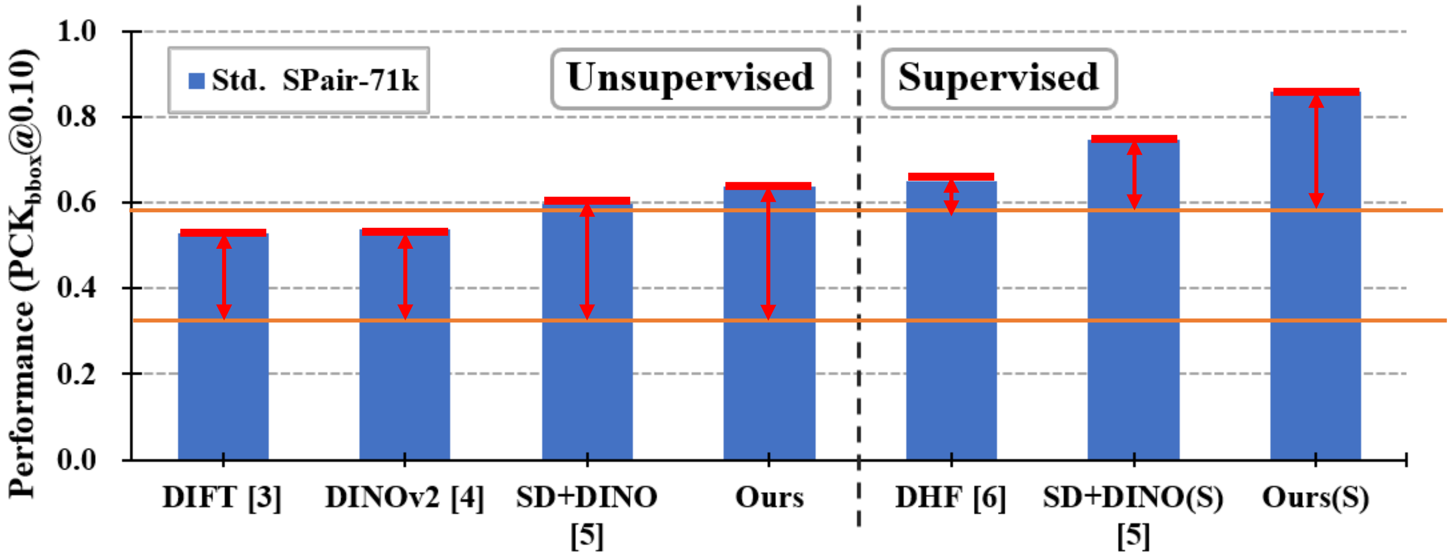
[6] Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence. *NeurIPS*, 2023.

Background

- Semantic correspondence
- Pre-trained vision models



(a) SD+DINO [5] struggles at “telling left from right” (red solid lines).



CATs++ [2], previous supervised

ASIC [1], previous unsupervised

[1] ASIC: Aligning Sparse in-the-wild Image Collections. *ICCV*, 2023.

[2] CATs++: Boosting Cost Aggregation With Convolutions and Transformers. *TPAMI*, 2022.

[3] Emergent Correspondance from Image Diffusion. *NeurIPS*, 2023.

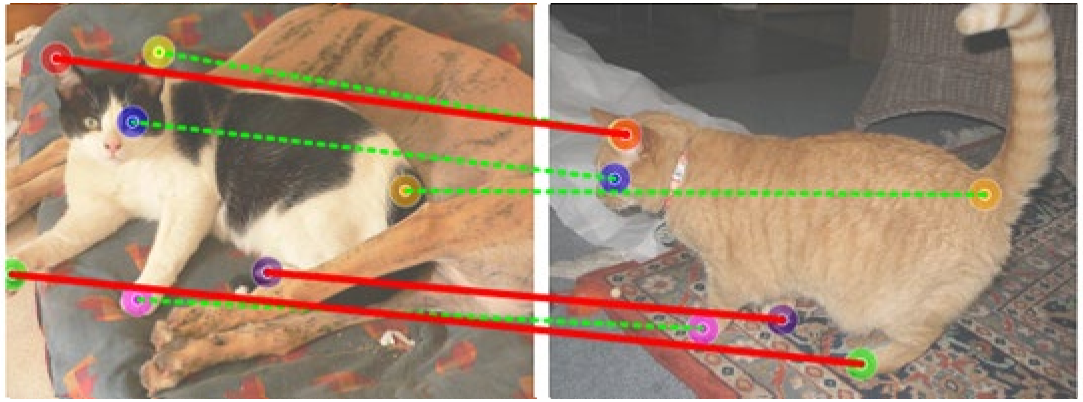
[4] DINOv2: Learning Robust Visual Features without Supervision. *Arxiv*, 2023.

[5] A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence. *NeurIPS*, 2023.

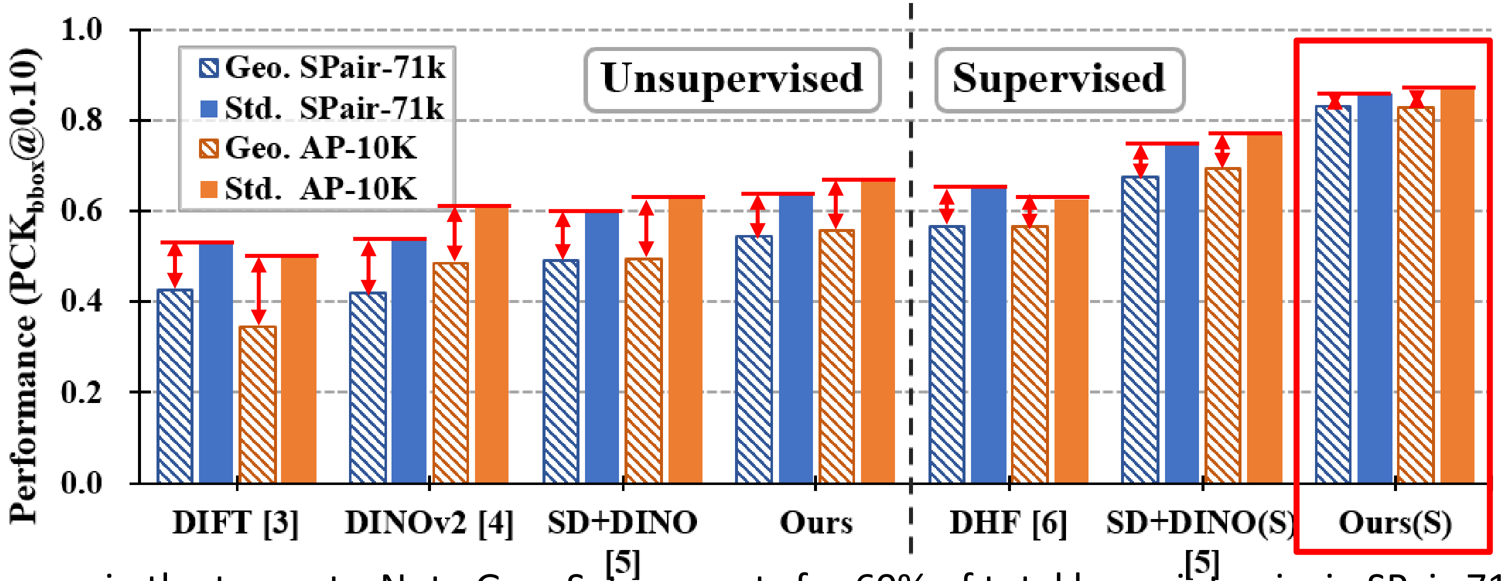
[6] Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence. *NeurIPS*, 2023.

Background

- Semantic correspondence
- Pre-trained vision models



(a) SD+DINO [5] struggles at “telling left from right” (red solid lines).

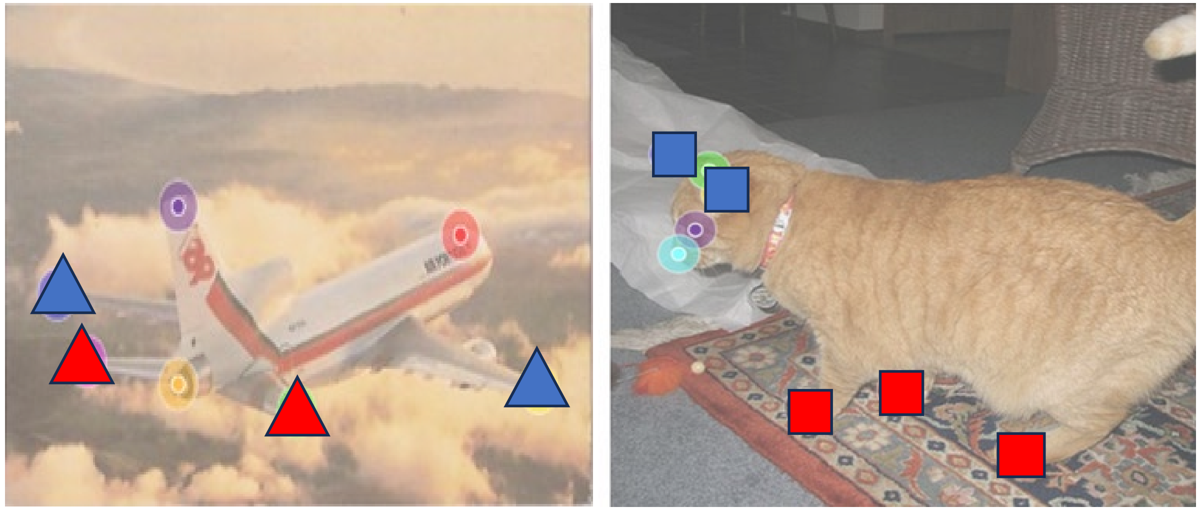


(b) Performance gap in the two sets. Note Geo. Set accounts for 60% of total keypoint pairs in SPair-71k and 45% in AP-10K.

[1] ASIC: Aligning Sparse in-the-wild Image Collections. *ICCV*, 2023.
[2] CATs++: Boosting Cost Aggregation With Convolutions and Transformers. *TPAMI*, 2022.
[3] Emergent Correspondance from Image Diffusion. *NeurIPS*, 2023.
[4] DINOv2: Learning Robust Visual Features without Supervision. *Arxiv*, 2023.
[5] A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence. *NeurIPS*, 2023.
[6] Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence. *NeurIPS*, 2023.

Geometric Aware SD/DINO Features

- Geo-aware semantic correspondence



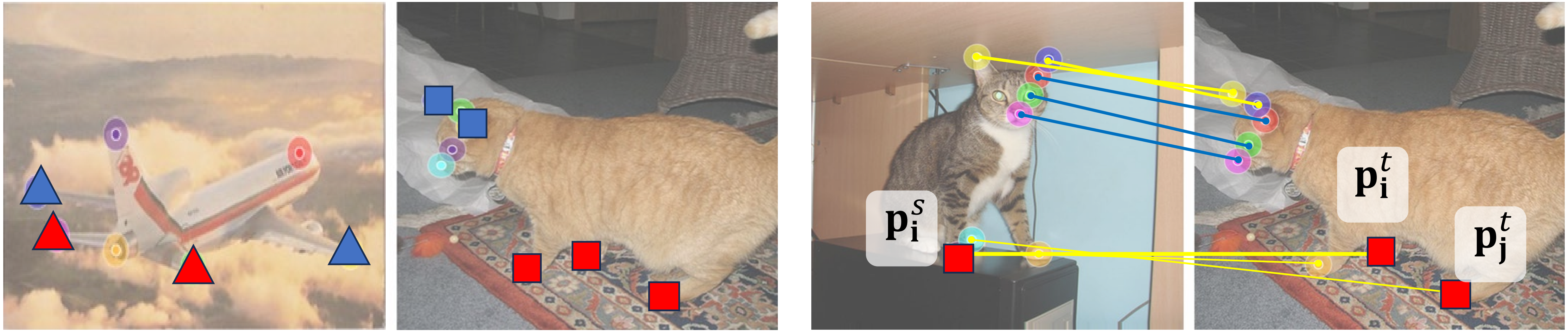
(a) Semantically-similar keypoint subgroups in images.

e.g., a group in cat category could be:

$\mathcal{G}_{\text{paws}} = \{\mathbf{p}_{(\text{paws, front left})}, \ \mathbf{p}_{(\text{paws, front right})}, \ \mathbf{p}_{(\text{paws, rear left})}, \ \mathbf{p}_{(\text{paws, rear right})}\}$

Geometric Aware SD/DINO Features

- Geo-aware semantic correspondence to identify ambiguities


$\mathbf{p}_{\mathbf{i}}^{s}$
$\mathbf{p}_{\mathbf{i}}^{t}$
$\mathbf{p}_{\mathbf{j}}^{t}$

(a) Semantically-similar keypoint subgroups in images.

(b) Annotations of geo-aware semantic correspondence (yellow lines).

e.g., a group in cat category could be:

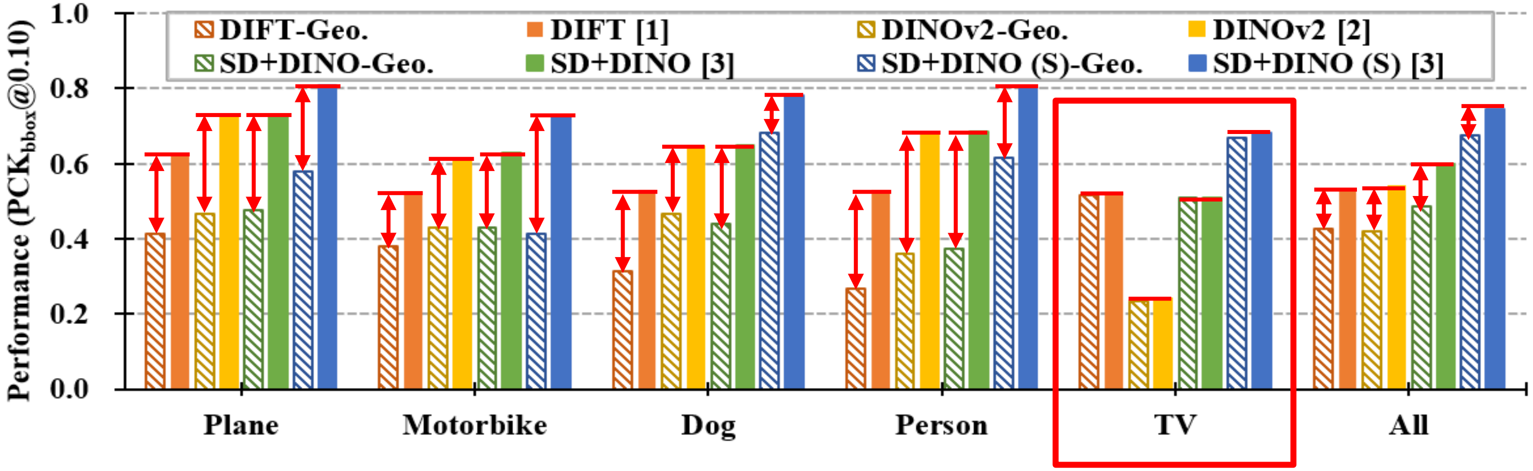
$\mathcal{G}_{\text{paws}} = \{\mathbf{p}_{(\text{paws,front left})}, \mathbf{p}_{(\text{paws,front right})}, \mathbf{p}_{(\text{paws,rear left})}, \mathbf{p}_{(\text{paws,rear right})}\}$

$\langle \mathbf{p}_{\mathbf{i}}^{s}, \mathbf{p}_{\mathbf{i}}^{t} \rangle$ is considered as a "geometry-aware" correspondence if:

1. $\mathbf{p}_{\mathbf{i}}^{s} \in \mathcal{G}_{part}^{s}, \mathbf{p}_{\mathbf{i}}^{t} \in \mathcal{G}_{part}^{t},$
2. $\exists\, \mathbf{j} \neq \mathbf{i}$ s.t. $\mathrm{p}_{\mathbf{j}}^{t} \in \mathcal{G}_{part}^{t}$

Geometric Aware SD/DINO Features

- Evaluation on the geometry-aware subset



(a) Per-category performance on geo-aware set.

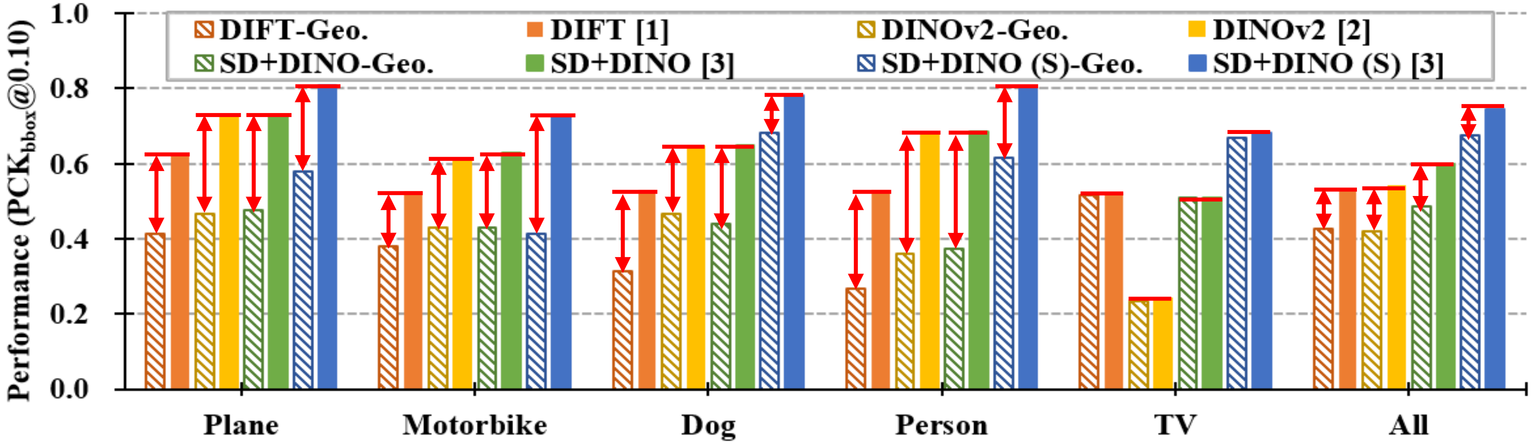
[1] Emergent Correspondance from Image Diffusion. *NeurIPS*, 2023.

[2] DINOv2: Learning Robust Visual Features without Supervision. *Arxiv*, 2023.

[3] A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence. *NeurIPS*, 2023.

Geometric Aware SD/DINO Features

- Evaluation on the geometry-aware subset



(a) Per-category performance on geo-aware set.

- Sensitivity to pose variation

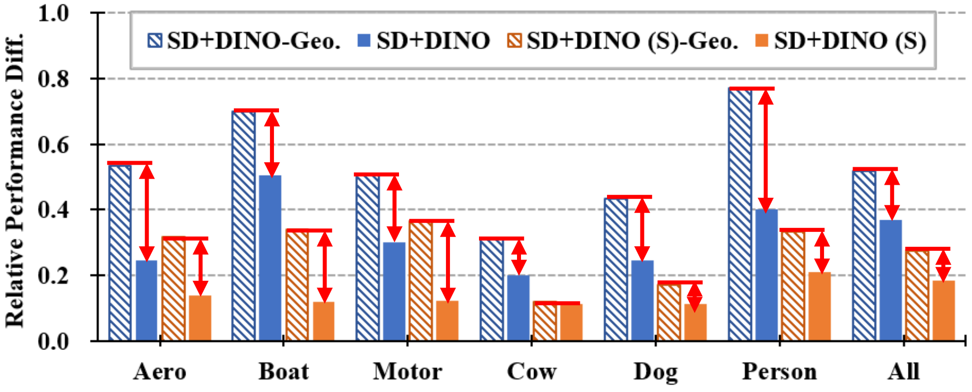
Divide 5 sets by azimuth difference

Performance on the five subsets:

$$\mathcal{A} = \{\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_4\}$$

Normalized relative difference:

$$\mathbf{d} = \frac{\max(\mathcal{A}) - \min(\mathcal{A})}{\max(\mathcal{A})}$$



(b) Sensitivity to pose variation (higher value = more sensitivity).

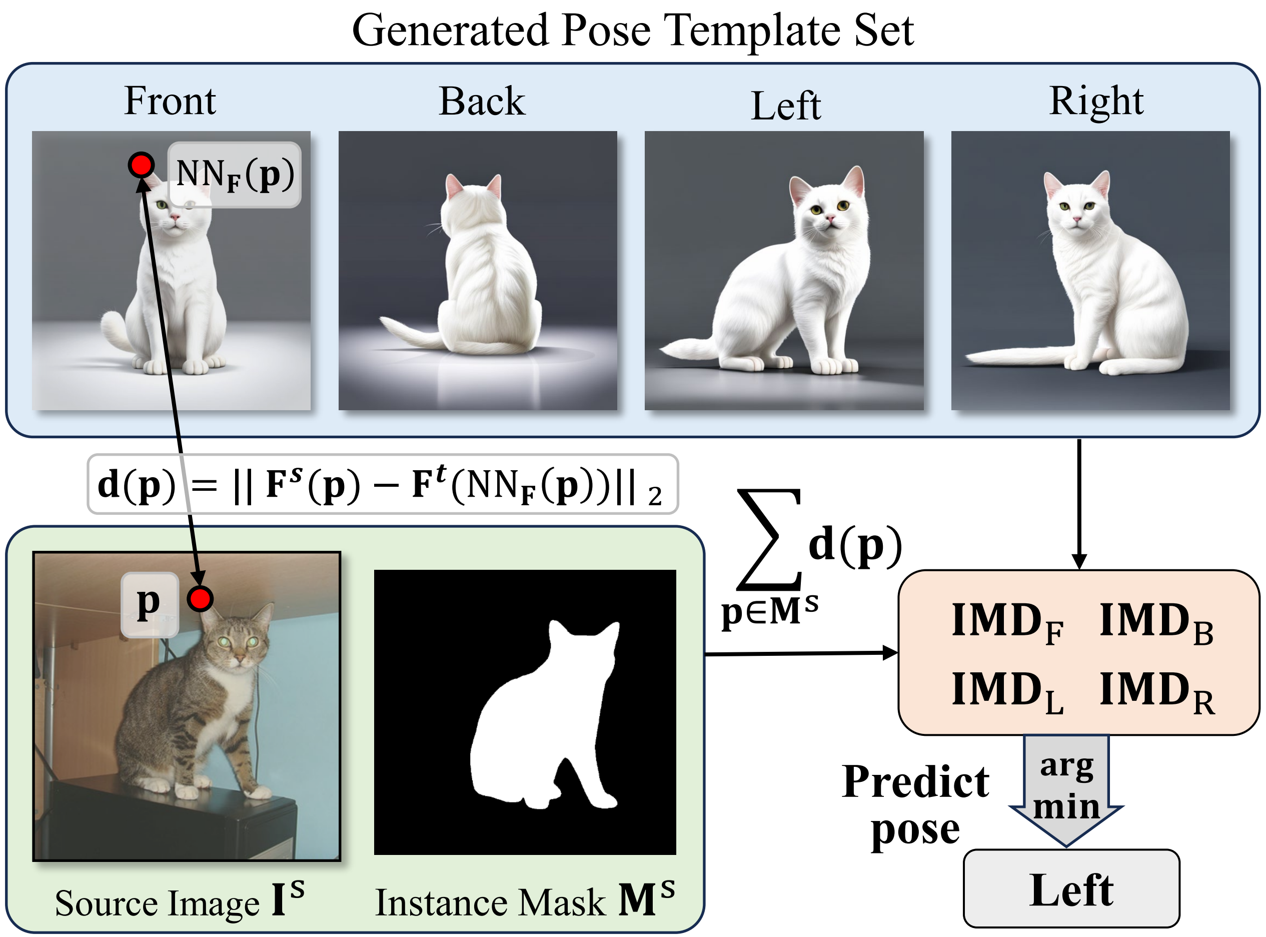
[1] Emergent Correspondance from Image Diffusion. *NeurIPS*, 2023.

[2] DINOv2: Learning Robust Visual Features without Supervision. *Arxiv*, 2023.

[3] A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence. *NeurIPS*, 2023.

Geometric Aware SD/DINO Features

- Global pose awareness of deep features



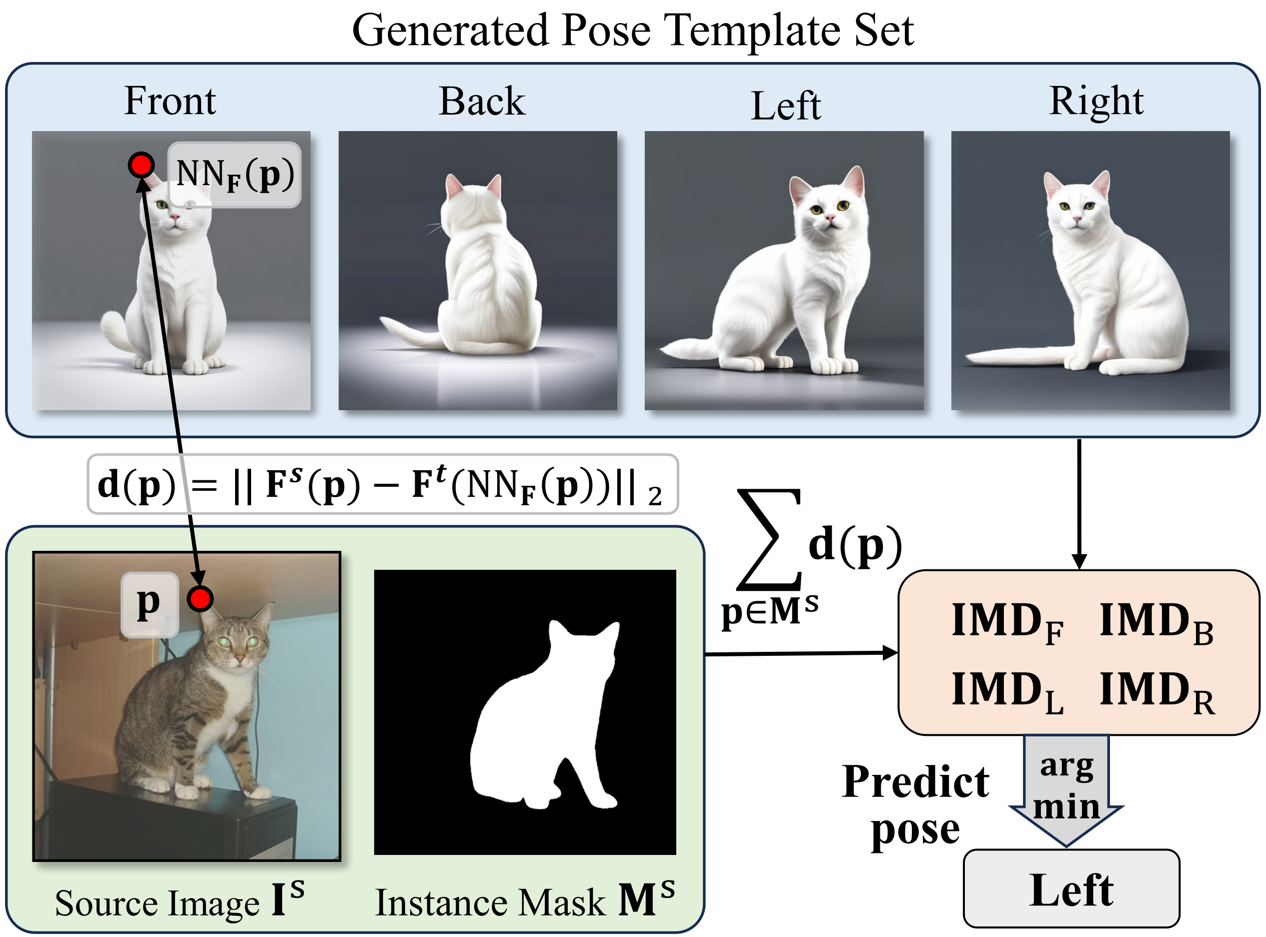
Generated Pose Template Set
Front
Back
Left
Right
$\mathrm{NN_F(p)}$
$\mathbf{d(p)} = ||\,\mathbf{F}^s(\mathbf{p}) - \mathbf{F}^t(\mathrm{NN}_\mathbf{F}(\mathbf{p}))||_2$
$\sum_{\mathbf{p} \in \mathbf{M}^S} \mathbf{d(p)}$
p
Source Image $\mathbf{I}^S$
Instance Mask $\mathbf{M}^S$
$\mathrm{IMD_F}$ $\mathrm{IMD_B}$
$\mathrm{IMD_L}$ $\mathrm{IMD_R}$
Predict pose
arg min
Left

(a) Rough pose prediction with feature distance.

Manually annotated 100 cat images from SPair-71k with pose labels {left, right, front, and back}

Geometric Aware SD/DINO Features

- Global pose awareness of deep features



Generated Pose Template Set
Front
Back
Left
Right
$\mathbf{NN_F(p)}$
$\mathbf{d(p)} = \|\mathbf{F}^s(\mathbf{p}) - \mathbf{F}^t(\mathrm{NN}_\mathbf{F}(\mathbf{p}))\|_2$
$\sum_{\mathbf{p} \in \mathbf{M}^S} \mathbf{d(p)}$
$\mathbf{IMD_F}$ $\mathbf{IMD_B}$
$\mathbf{IMD_L}$ $\mathbf{IMD_R}$
Predict pose
arg min
Left
p
Source Image $\mathbf{I}^S$
Instance Mask $\mathbf{M}^S$

(a) Rough pose prediction with feature distance.

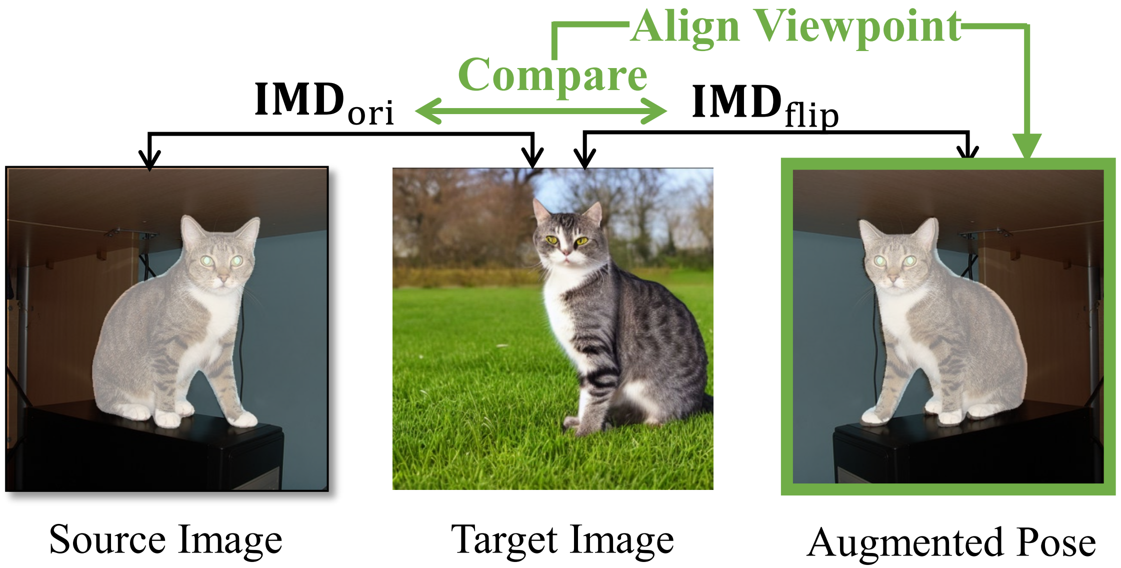
Feature	L/R	F/B	L/R or F/B	L/R/F/B
DINOv2	63.8	100.0	75.0	51.0
SD	95.7	96.8	96.0	78.0
SD+DINO	98.6	100.0	99.0	84.0

(b) Zero-shot rough pose prediction result with instance matching distance (IDM). We manually annotated 100 cat images from SPair-71k with rough pose labels {left, right, front, and back} and report the accuracy of predicting left or right (L/R), front or back (F/B), either of the two cases (L/R or F/B), and one of the four directions (L/R/F/B).

Deep features are aware of global pose
DINO v2 performs on F/B but not L/R
SD performs well on F/B and L/R
SD+DINO performs best

Improving Geo-Aware Semantic Correspondence

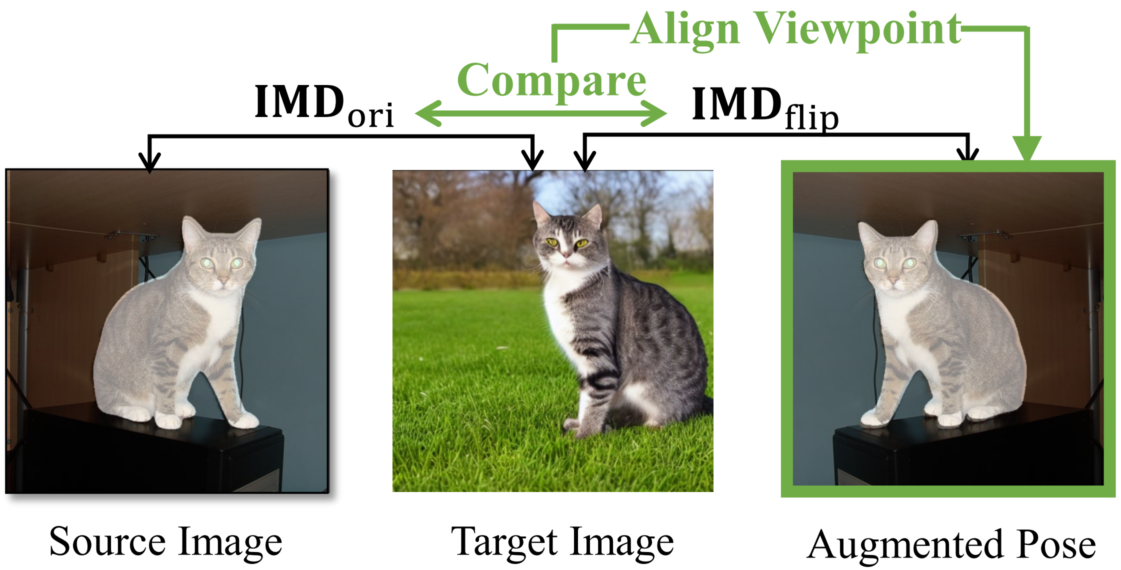
- Zero-shot setting: test-time adaptive pose alignment



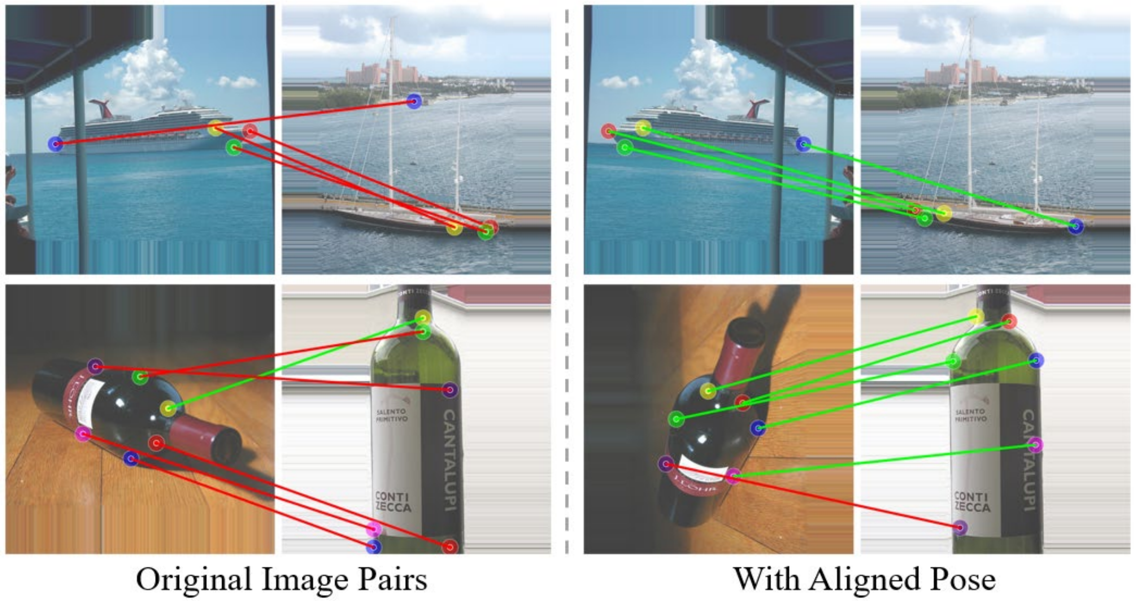
(a) Adaptive pose alignment with feature space distance.

Improving Geo-Aware Semantic Correspondence

- Test-time adaptive pose alignment (using a set of pose-variant augmentations, e.g., flip, rotations)



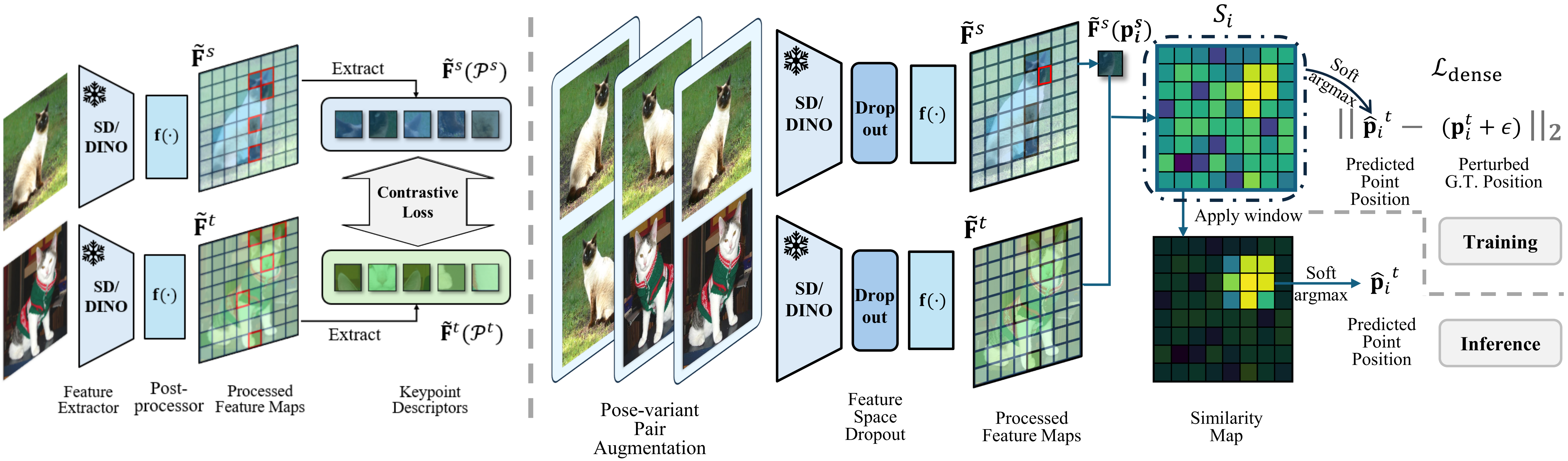
(a) Adaptive pose alignment with feature space distance.



(b) Qualitative results of the adaptive alignment.

Improving Geo-Aware Semantic Correspondence

- Supervised framework

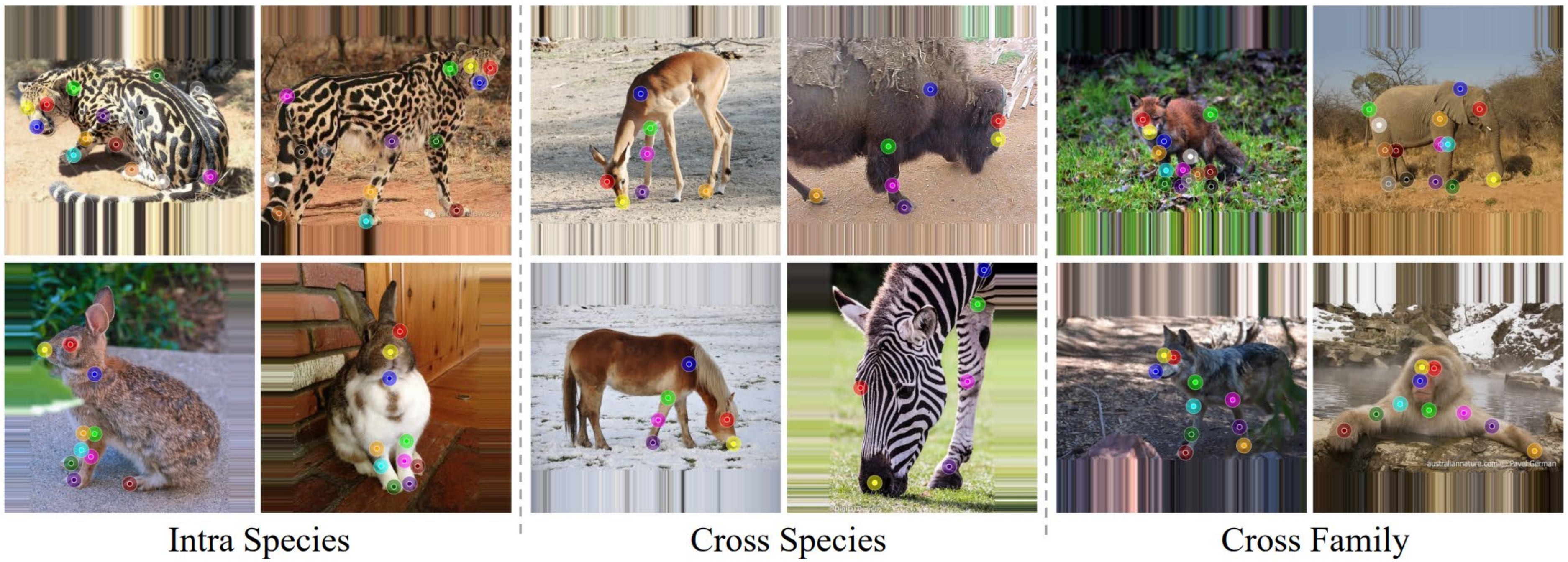


(a) (Left) previous supervised methods [1,2] with a sparse training objective. (Right) an overview of our supervised method

[1] Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence. *NeurIPS*, 2023.
[2] A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence. *NeurIPS*, 2023.

Semantic Correspondence on AP-10K

- AP-10K, in-the-wild animal pose estimation dataset, 10,015 images, across 23 families and 54 species
- Construct a benchmark: 261k training, 17k validation, 36k testing image pairs
- Testing setting: intra-species, cross-species, cross-family

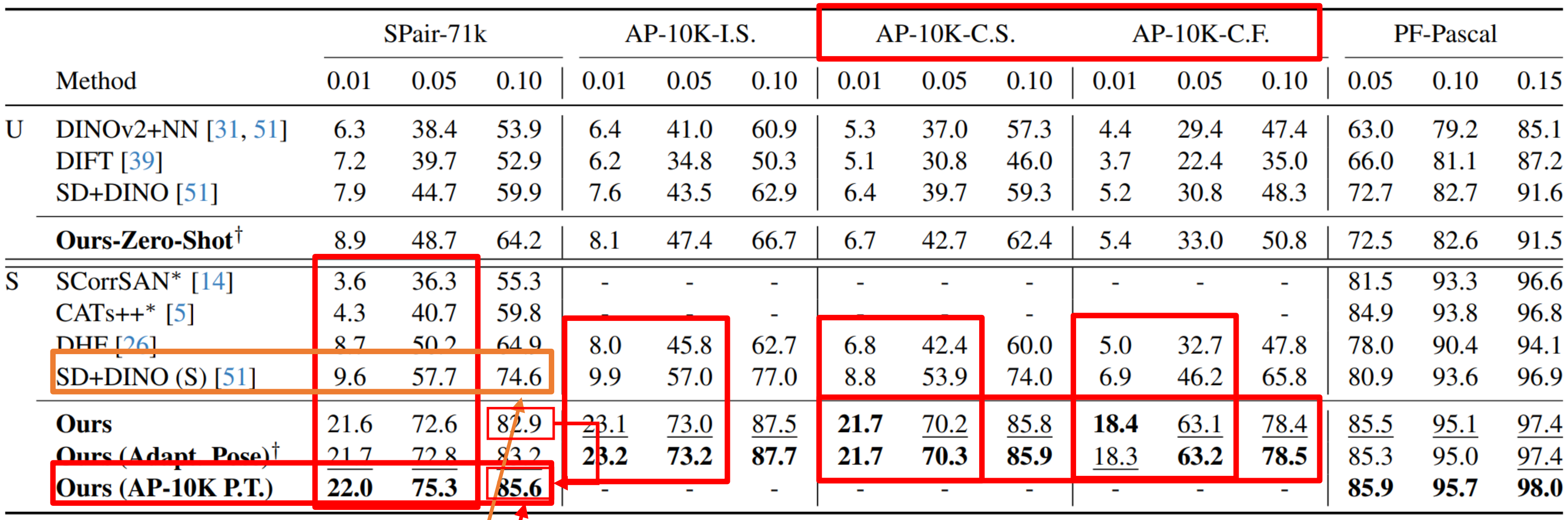


(a) Sample image pairs from AP-10K semantic correspondence benchmark.

AP-10K: A Benchmark for Animal Pose Estimation in the Wild. *NeurIPS D&B Track*, 2021.

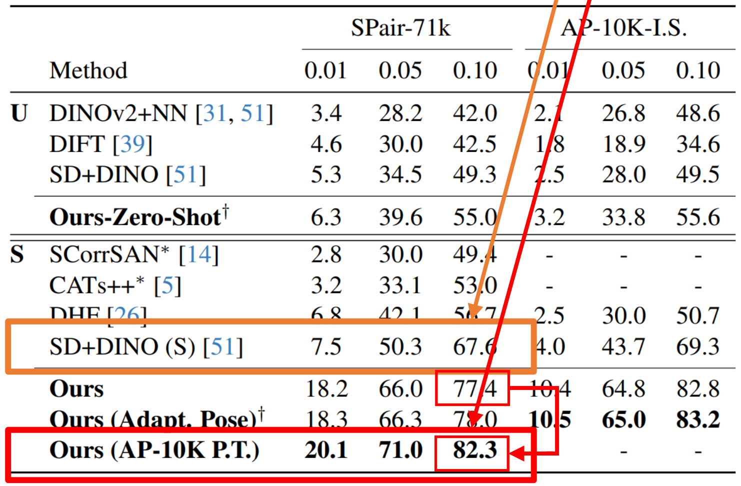
Quantitative Results

(a) Quantitative comparison across different datasets (standard) and PCK levels.



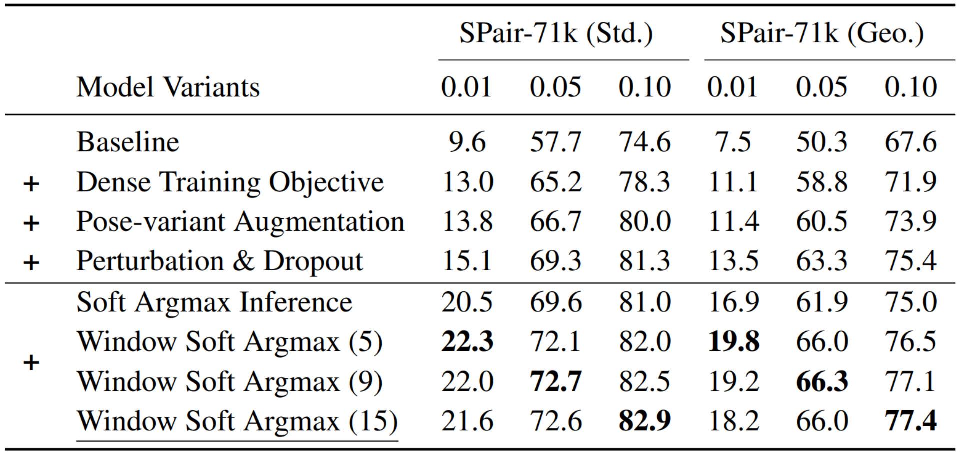
		SPair-71k			AP-10K-I.S.			AP-10K-C.S.			AP-10K-C.F.			PF-Pascal		
	Method	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10	0.05	0.10	0.15
U	DINOv2+NN [31, 51]	6.3	38.4	53.9	6.4	41.0	60.9	5.3	37.0	57.3	4.4	29.4	47.4	63.0	79.2	85.1
	DIFT [39]	7.2	39.7	52.9	6.2	34.8	50.3	5.1	30.8	46.0	3.7	22.4	35.0	66.0	81.1	87.2
	SD+DINO [51]	7.9	44.7	59.9	7.6	43.5	62.9	6.4	39.7	59.3	5.2	30.8	48.3	72.7	82.7	91.6
	Ours-Zero-Shot†	8.9	48.7	64.2	8.1	47.4	66.7	6.7	42.7	62.4	5.4	33.0	50.8	72.5	82.6	91.5
S	SCorrSAN* [14]	3.6	36.3	55.3	-	-	-	-	-	-	-	-	-	81.5	93.3	96.6
	CATs++* [5]	4.3	40.7	59.8	-	-	-	-	-	-	-	-	-	84.9	93.8	96.8
	DHF [26]	8.7	50.2	64.9	8.0	45.8	62.7	6.8	42.4	60.0	5.0	32.7	47.8	78.0	90.4	94.1
	SD+DINO (S) [51]	9.6	57.7	74.6	9.9	57.0	77.0	8.8	53.9	74.0	6.9	46.2	65.8	80.9	93.6	96.9
	Ours	21.6	72.6	82.9	23.1	73.0	87.5	**21.7**	70.2	85.8	**18.4**	63.1	78.4	85.5	95.1	97.4
	Ours (Adapt. Pose)†	21.7	72.8	83.2	**23.2**	**73.2**	**87.7**	**21.7**	**70.3**	**85.9**	18.3	**63.2**	**78.5**	85.3	95.0	97.4
	Ours (AP-10K P.T.)	**22.0**	**75.3**	**85.6**	-	-	-	-	-	-	-	-	-	**85.9**	**95.7**	**98.0**

(b) Quantitative comparison across different datasets (Geo.) and PCK levels.



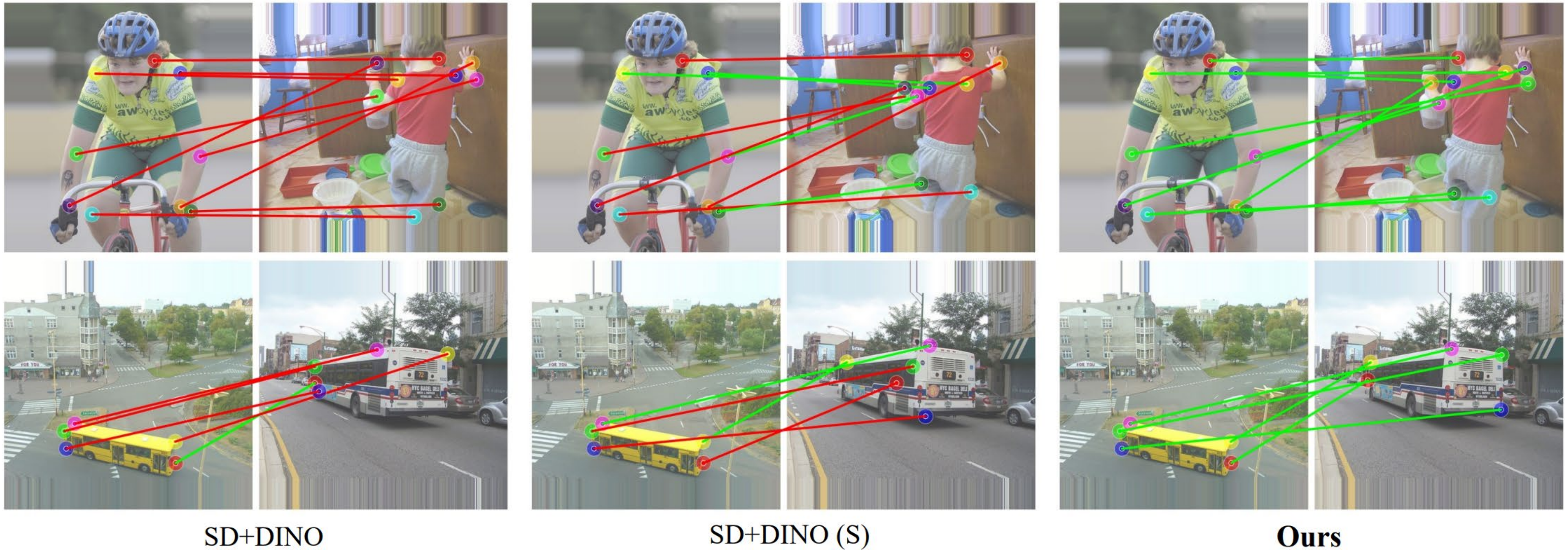
		SPair-71k			AP-10K-I.S.		
	Method	0.01	0.05	0.10	0.01	0.05	0.10
U	DINOv2+NN [31, 51]	3.4	28.2	42.0	2.1	26.8	48.6
	DIFT [39]	4.6	30.0	42.5	1.8	18.9	34.6
	SD+DINO [51]	5.3	34.5	49.3	2.5	28.0	49.5
	Ours-Zero-Shot†	6.3	39.6	55.0	3.2	33.8	55.6
S	SCorrSAN* [14]	2.8	30.0	49.4	-	-	-
	CATs++* [5]	3.2	33.1	53.0	-	-	-
	DHF [26]	6.8	42.1	56.7	2.5	30.0	50.7
	SD+DINO (S) [51]	7.5	50.3	67.6	4.0	43.7	69.3
	Ours	18.2	66.0	77.4	10.4	64.8	82.8
	Ours (Adapt. Pose)†	18.3	66.3	78.0	**10.5**	**65.0**	**83.2**
	Ours (AP-10K P.T.)	**20.1**	**71.0**	**82.3**	-	-	-

(c) Ablation study on the Spair-71k standard and the geometry-aware sets.

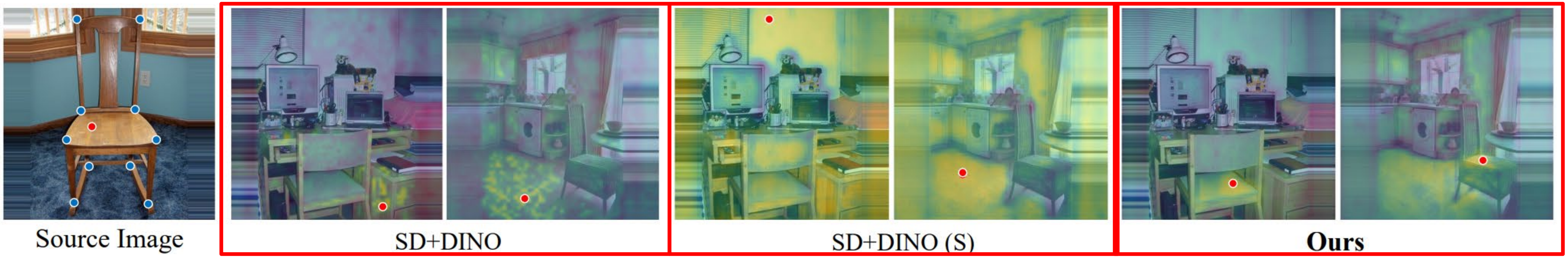


	Model Variants	SPair-71k (Std.) 0.01	SPair-71k (Std.) 0.05	SPair-71k (Std.) 0.10	SPair-71k (Geo.) 0.01	SPair-71k (Geo.) 0.05	SPair-71k (Geo.) 0.10
	Baseline	9.6	57.7	74.6	7.5	50.3	67.6
+	Dense Training Objective	13.0	65.2	78.3	11.1	58.8	71.9
+	Pose-variant Augmentation	13.8	66.7	80.0	11.4	60.5	73.9
+	Perturbation & Dropout	15.1	69.3	81.3	13.5	63.3	75.4
+	Soft Argmax Inference	20.5	69.6	81.0	16.9	61.9	75.0
	Window Soft Argmax (5)	**22.3**	72.1	82.0	**19.8**	66.0	76.5
	Window Soft Argmax (9)	22.0	**72.7**	82.5	19.2	**66.3**	77.1
	Window Soft Argmax (15)	21.6	72.6	**82.9**	18.2	66.0	**77.4**

Qualitative Results

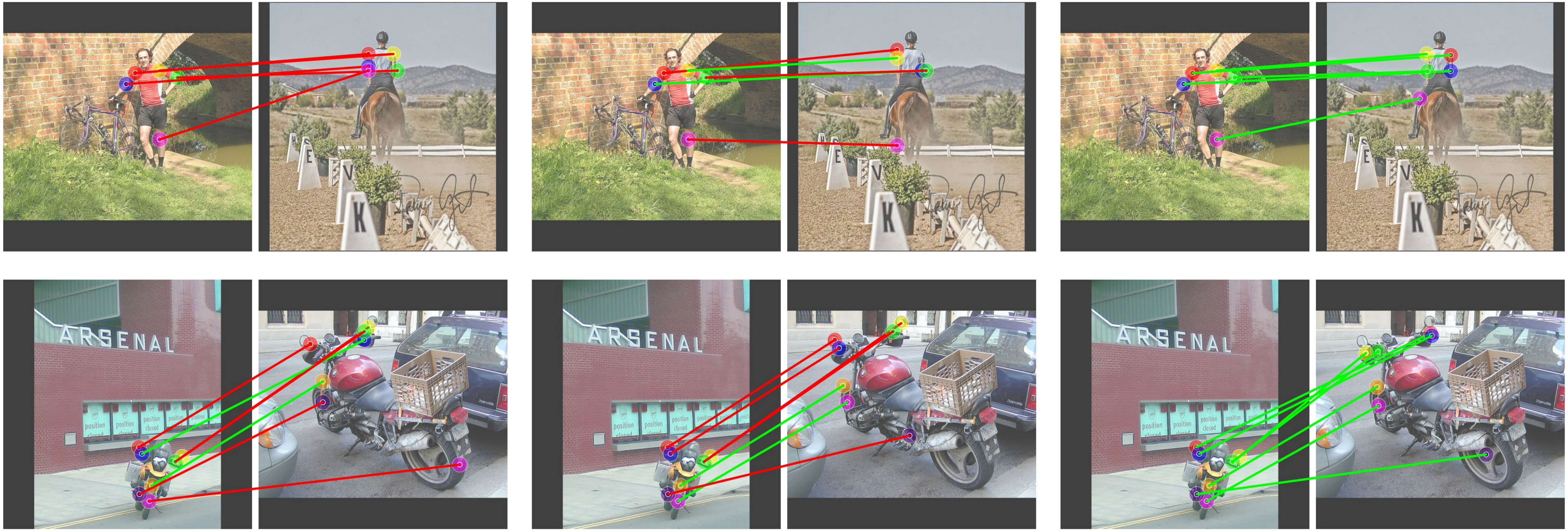


(a) Qualitative comparison with state-of-the-art methods in cases with extreme viewpoint variations.



(b) Visualization of similarity map. The query and predicted points are red, and the keypoint supervision of the “chair” category is blue.

Qualitative Results on SPair-71k

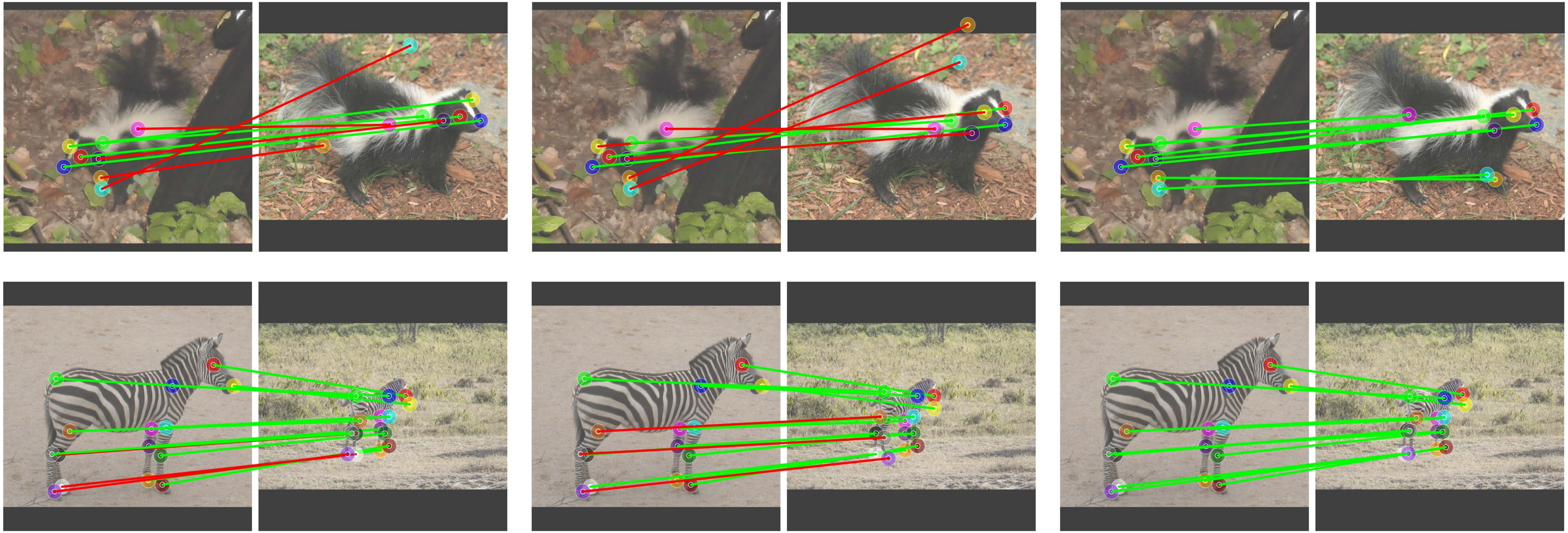


SD+DINO

SD+DINO (S)

Ours

Qualitative Results on AP-10K Intra-Species



SD+DINO

SD+DINO (S)

Ours

Summary

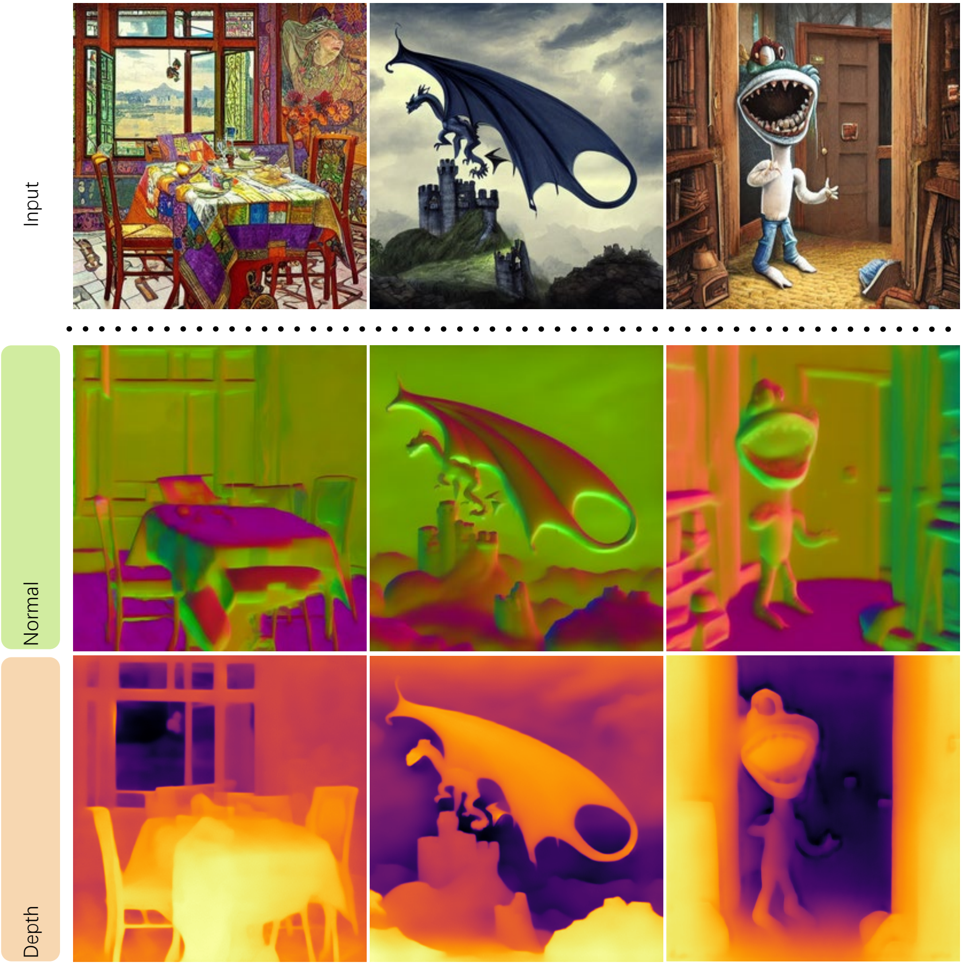
- Identify the problem of geometry-aware semantic correspondence
- Improve geometric awareness of the features in unsupervised and supervised setting
- Introduce a large-scale and challenging benchmark
- Boosts the performance on multiple benchmark datasets, especially on geometry-aware subset, achieves 85.6 PCK@0.10 on Spair-71k (15% gain over SOTA).

Exploiting Diffusion Prior for Generalizable Pixel-Level Semantic Prediction

CVPR 2024

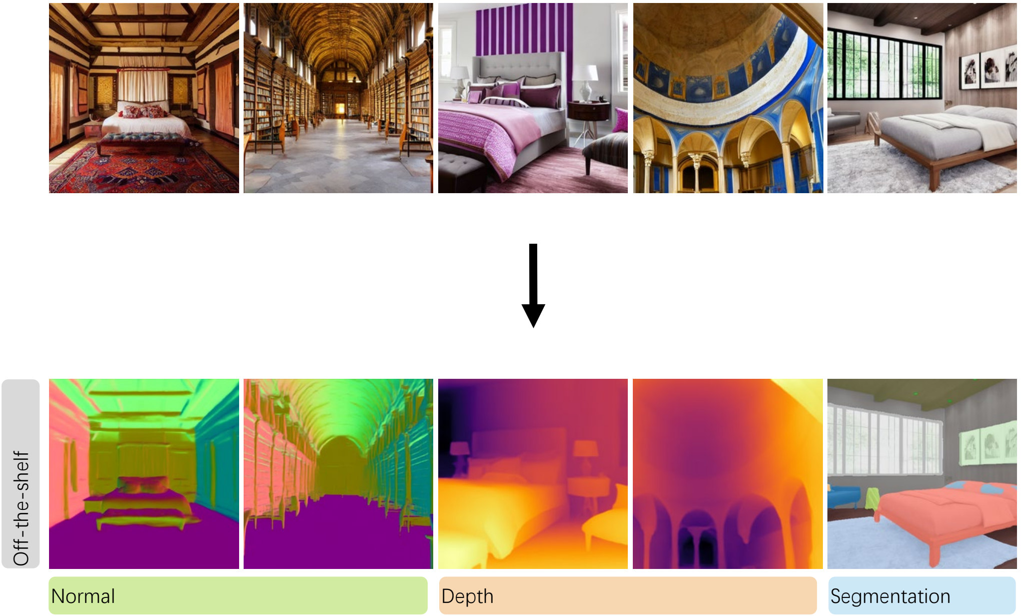
Hsin-Ying Lee[1] Hung-Yu Tseng[2] Hsin-Ying Lee[3] Ming-Hsuan Yang[1,4]

[1]UC Merced [2]Meta [3]Snap [4]Google Research



Pixel-Level/Dense Semantic Prediction

Real-world images



Generated images

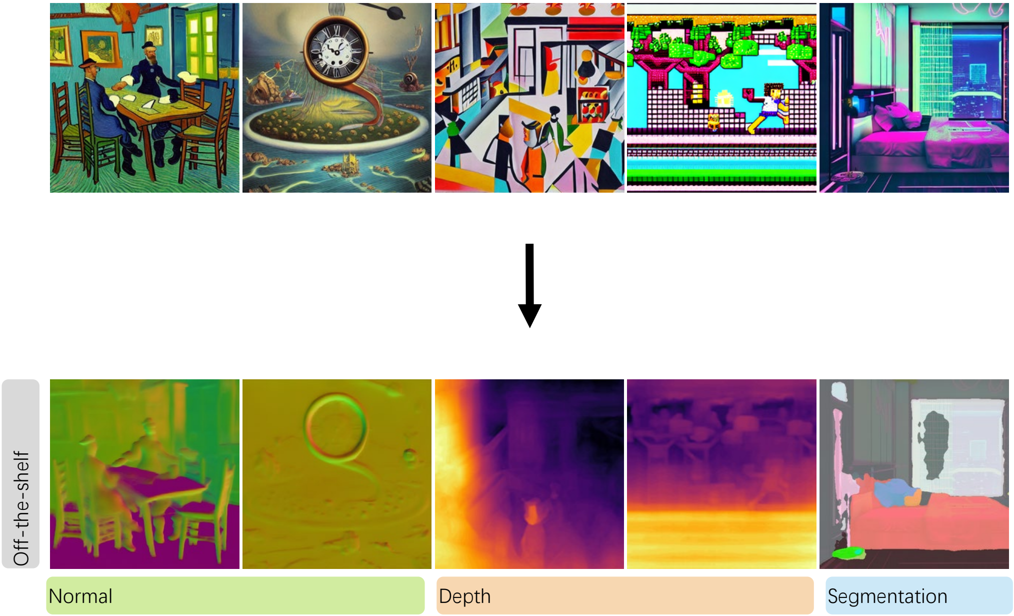


Image properties, e.g., normal, depth, segmentation, are important for image understanding
Existing generative models performs well on real-world images but not on generated images
How to adapt LTM for new task and maintain generalization?

- Address mismatch between deterministic prediction tasks and stochastic T2I model
- Preserve generalizability

Adapt Diffusion Models for Prediction?

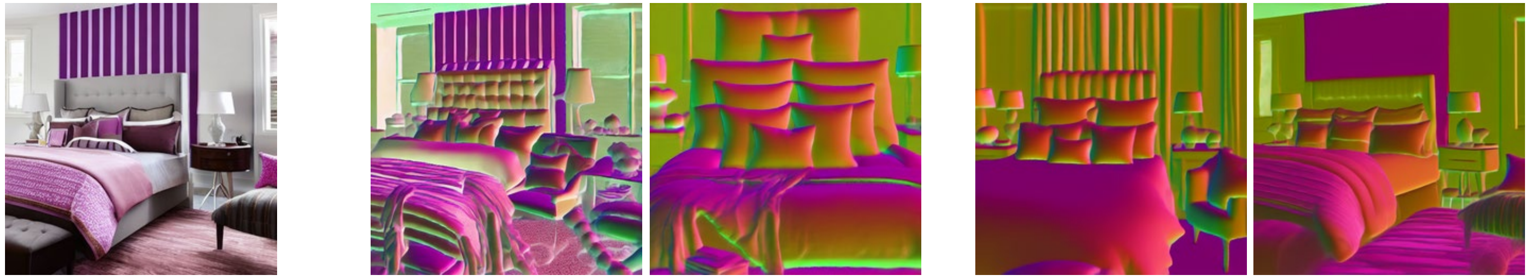
Naïve solution: consider dense prediction as an image-to-image (I2I) problem using diffusion models

Problems

- Diffusion models are inherently stochastic.
- Balance between generalizability and learning new tasks should be maintained.

Stochasticity: sampling initial noise and adding additional noise in the generation process
For deterministic prediction tasks, existing methods of I2I translation cannot be applied effectively

Inversion | Feature Injection | Concatenation



strong guidance | weak guidance | with injection

Not applicable due to extra layers

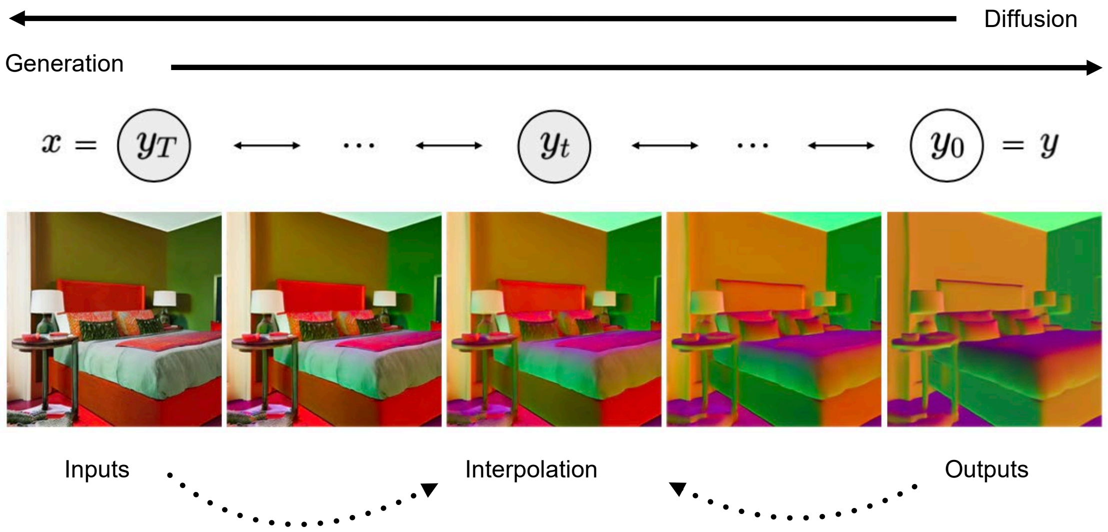
surface normal

Diffusion Models as Prior (DMP)

Solution

- Re-formulating the diffusion process with the blending perspective

Diffusion and generation process are entirely deterministic.



Diffusion
Generation
$x = y_T \longleftrightarrow \cdots \longleftrightarrow y_t \longleftrightarrow \cdots \longleftrightarrow y_0 = y$
Inputs
Interpolation
Outputs

Training with interpolations of inputs and outputs
Diffusion: morphing of output to input images
Generation: demorphing input to output images

x : input y: output

x and y are always paired

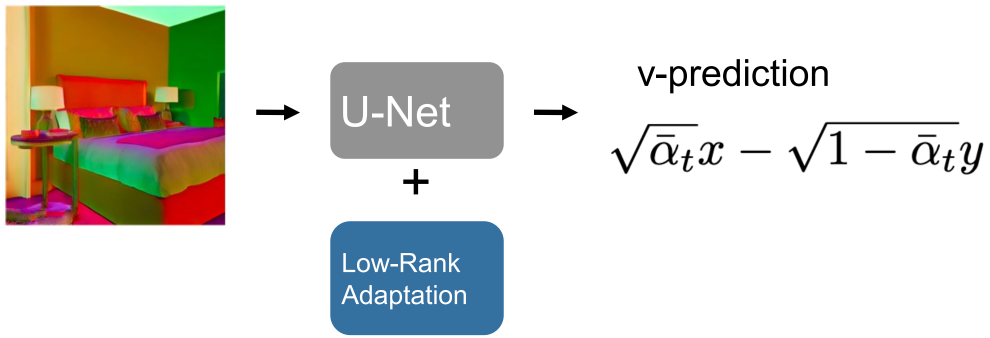
latent $y_t = \sqrt{\bar{\alpha}_t} y + \sqrt{1 - \bar{\alpha}_t} x \quad t = [1, \cdots, T].$

original $y_t = \sqrt{\bar{\alpha}_t} y + \sqrt{1 - \bar{\alpha}_t} \epsilon_t \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$

Diffusion Models as Prior

Solution

- Fine-tuning with low-rank adaptation to preserve generalizability
- Training to predict v-prediction



$$\sqrt{\bar\alpha_t}x - \sqrt{1-\bar\alpha_t}y$$

Training

$$L_{\text{DMP}} = \mathbb{E}_{(x,y),t}\left[\|(\sqrt{\bar\alpha_t}x - \sqrt{1-\bar\alpha_t}y) - v_\theta(y_t,t)\|_2^2\right]$$

Generation

$$\begin{aligned} y_{t-1} = {} & \sqrt{\bar\alpha_{t-1}}(\sqrt{\bar\alpha_t}y_t - \sqrt{1-\bar\alpha_t}v_\theta(y_t,t)) \\ & + \sqrt{1-\bar\alpha_{t-1}}x \qquad t = [T,\cdots,1], \end{aligned}$$

Fine-tune a pre-trained T2I model with LoRA and inject it for v prediction
Avoid extra noise in the generation process with deterministic sampling

Datasets and Setups

- Generate diverse text descriptions using a MTF model with filter by Parmar
- Using LDM to synthesize images with text descriptions
- Pseudo ground truth:
  - Omnidata v2 for normal
  - ZoeDepth for depth,
  - EVA2 for semantic segmentation
  - PIE-Net for image decomposition (albedo and shading)
- Using experimental setups as Bhattad et al.

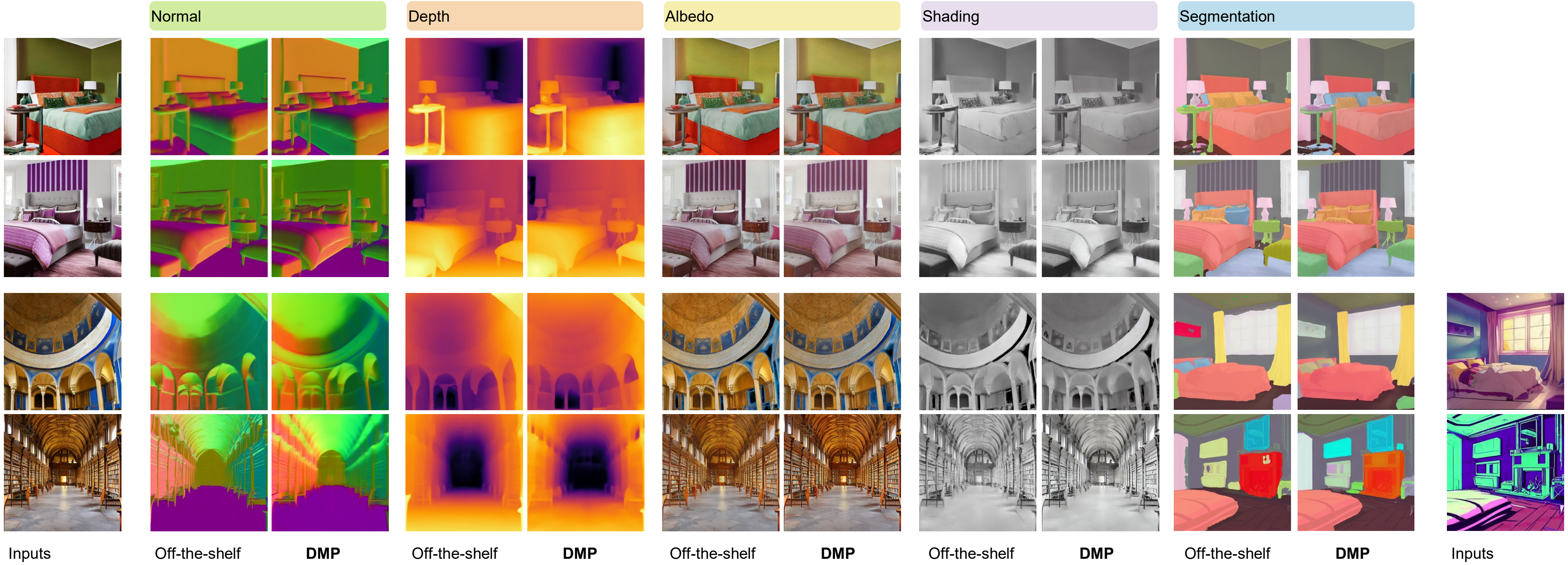
Evaluation

Training

Bedrooms
(10K)

• •

Test

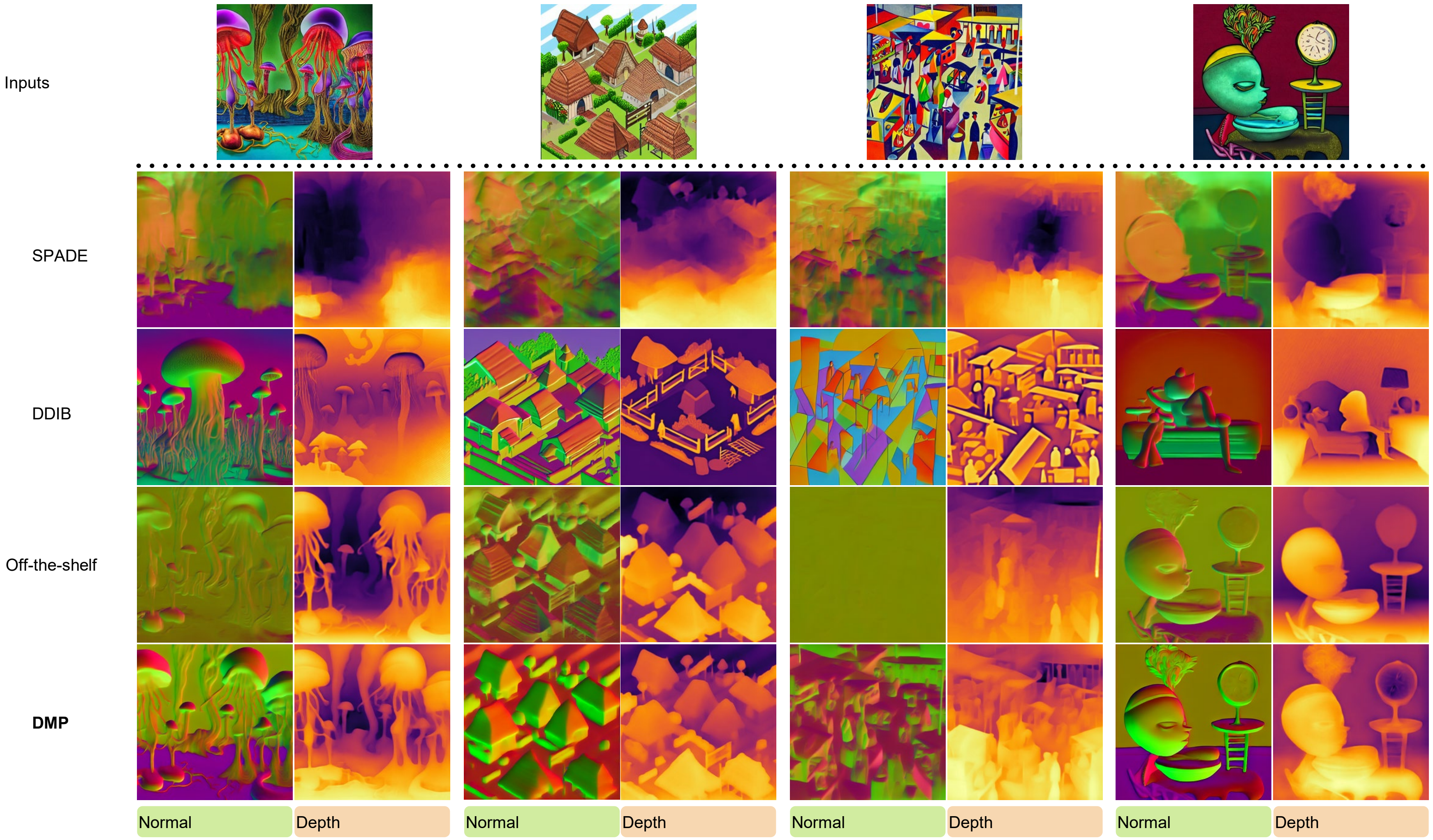


Bedrooms
(in-domain)

Diverse scenes
(out-of-domain)

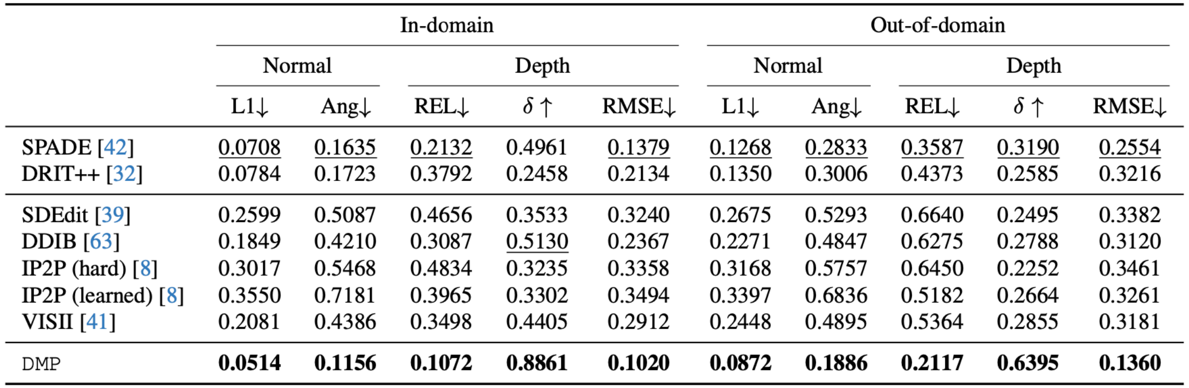
Arbitrary images
(out-of-domain)

Arbitrary Images



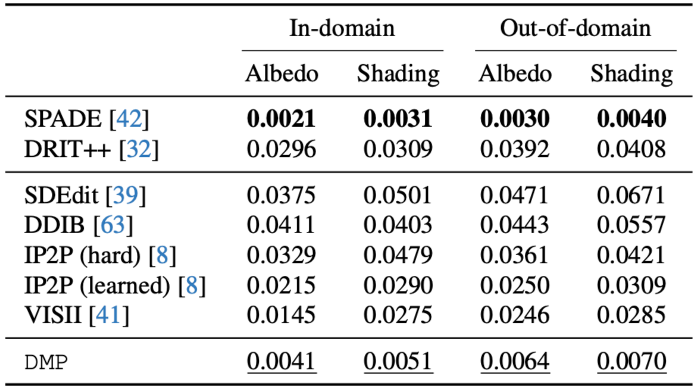
Performance Evaluation

- GAN-based: SPADE, DRIT++
- Diffusion-based: SDEdit, DDIB, IP2P (hard: fixed prompts), IP2P (learned: token inversion), VISII



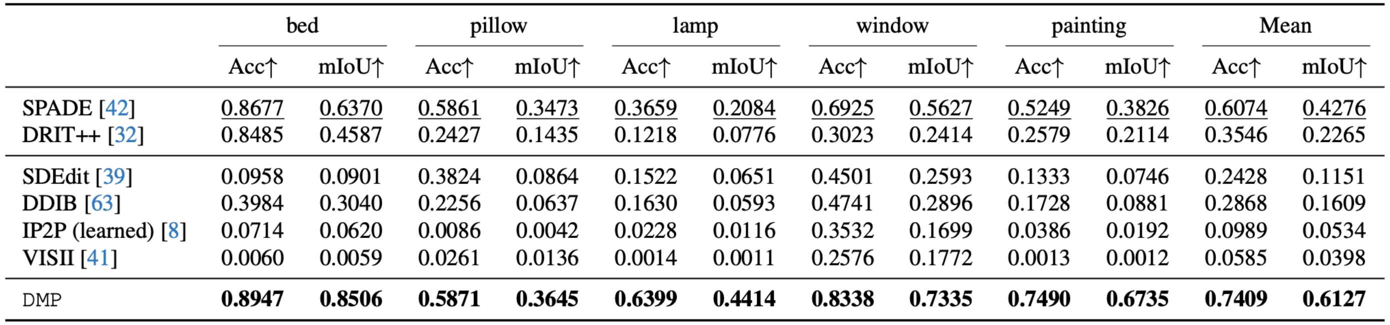
	In-domain					Out-of-domain				
	Normal		Depth			Normal		Depth		
	L1↓	Ang↓	REL↓	δ ↑	RMSE↓	L1↓	Ang↓	REL↓	δ ↑	RMSE↓
SPADE [42]	0.0708	0.1635	0.2132	0.4961	0.1379	0.1268	0.2833	0.3587	0.3190	0.2554
DRIT++ [32]	0.0784	0.1723	0.3792	0.2458	0.2134	0.1350	0.3006	0.4373	0.2585	0.3216
SDEdit [39]	0.2599	0.5087	0.4656	0.3533	0.3240	0.2675	0.5293	0.6640	0.2495	0.3382
DDIB [63]	0.1849	0.4210	0.3087	0.5130	0.2367	0.2271	0.4847	0.6275	0.2788	0.3120
IP2P (hard) [8]	0.3017	0.5468	0.4834	0.3235	0.3358	0.3168	0.5757	0.6450	0.2252	0.3461
IP2P (learned) [8]	0.3550	0.7181	0.3965	0.3302	0.3494	0.3397	0.6836	0.5182	0.2664	0.3261
VISII [41]	0.2081	0.4386	0.3498	0.4405	0.2912	0.2448	0.4895	0.5364	0.2855	0.3181
DMP	**0.0514**	**0.1156**	**0.1072**	**0.8861**	**0.1020**	**0.0872**	**0.1886**	**0.2117**	**0.6395**	**0.1360**

Normal and depth estimation



	In-domain		Out-of-domain	
	Albedo	Shading	Albedo	Shading
SPADE [42]	**0.0021**	**0.0031**	**0.0030**	**0.0040**
DRIT++ [32]	0.0296	0.0309	0.0392	0.0408
SDEdit [39]	0.0375	0.0501	0.0471	0.0671
DDIB [63]	0.0411	0.0403	0.0443	0.0557
IP2P (hard) [8]	0.0329	0.0479	0.0361	0.0421
IP2P (learned) [8]	0.0215	0.0290	0.0250	0.0309
VISII [41]	0.0145	0.0275	0.0246	0.0285
DMP	0.0041	0.0051	0.0064	0.0070

Intrinsic image decomposition

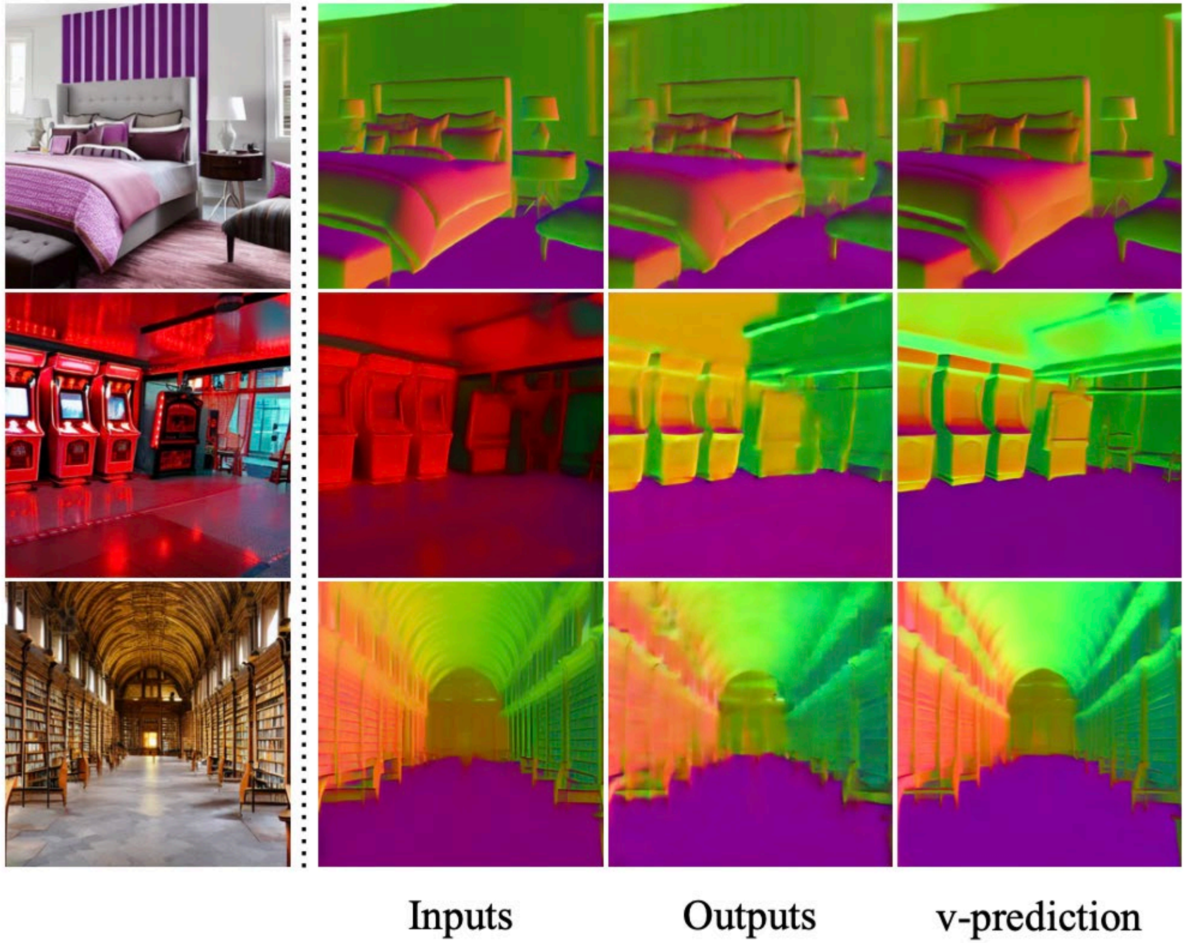


	bed		pillow		lamp		window		painting		Mean	
	Acc↑	mIoU↑	Acc↑	mIoU↑	Acc↑	mIoU↑	Acc↑	mIoU↑	Acc↑	mIoU↑	Acc↑	mIoU↑
SPADE [42]	0.8677	0.6370	0.5861	0.3473	0.3659	0.2084	0.6925	0.5627	0.5249	0.3826	0.6074	0.4276
DRIT++ [32]	0.8485	0.4587	0.2427	0.1435	0.1218	0.0776	0.3023	0.2414	0.2579	0.2114	0.3546	0.2265
SDEdit [39]	0.0958	0.0901	0.3824	0.0864	0.1522	0.0651	0.4501	0.2593	0.1333	0.0746	0.2428	0.1151
DDIB [63]	0.3984	0.3040	0.2256	0.0637	0.1630	0.0593	0.4741	0.2896	0.1728	0.0881	0.2868	0.1609
IP2P (learned) [8]	0.0714	0.0620	0.0086	0.0042	0.0228	0.0116	0.3532	0.1699	0.0386	0.0192	0.0989	0.0534
VISII [41]	0.0060	0.0059	0.0261	0.0136	0.0014	0.0011	0.2576	0.1772	0.0013	0.0012	0.0585	0.0398
DMP	**0.8947**	**0.8506**	**0.5871**	**0.3645**	**0.6399**	**0.4414**	**0.8338**	**0.7335**	**0.7490**	**0.6735**	**0.7409**	**0.6127**

Semantic segmentation

Surface Normal Estimation

Fine-tune U-Net to predict signals

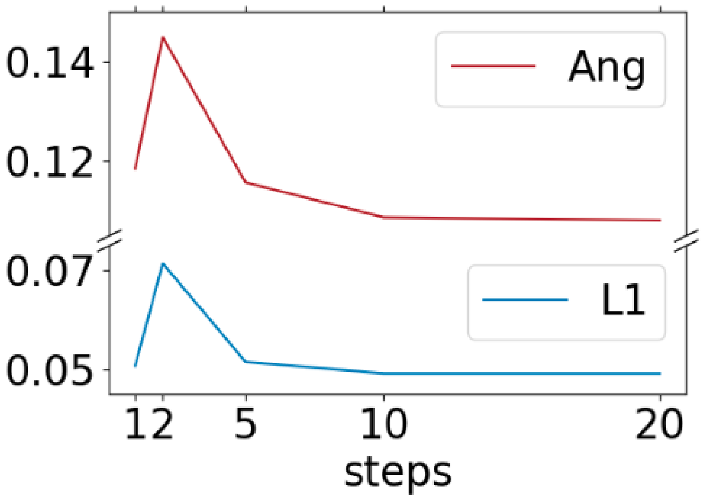


	In-domain		Out-of-domain	
	L1↓	Ang↓	L1↓	Ang↓
Predicting x	0.0736	0.1629	0.1319	0.2764
Predicting y	0.0590	0.1291	0.0888	0.1914
v-prediction	**0.0514**	**0.1156**	**0.0872**	**0.1886**

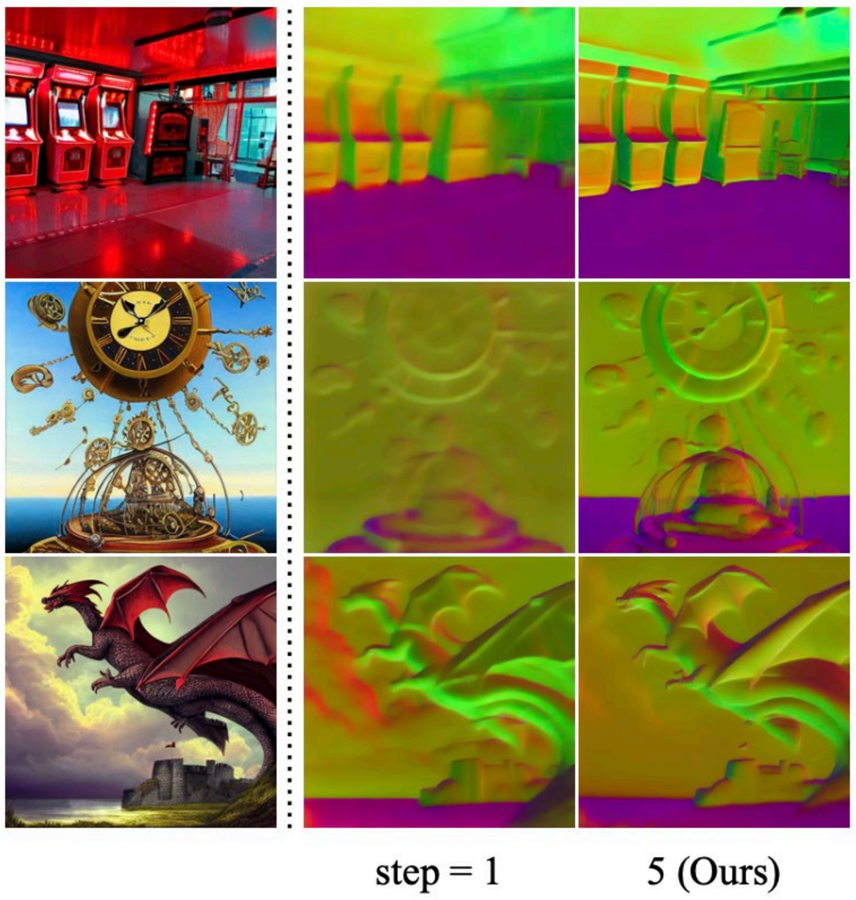
Predicting inputs obtains high-quality images with low generalizability.
Predicting outputs obtains blurred images but generalizes well.
v-prediction has both of their benefits.

Surface Normal Estimation

Generation Steps



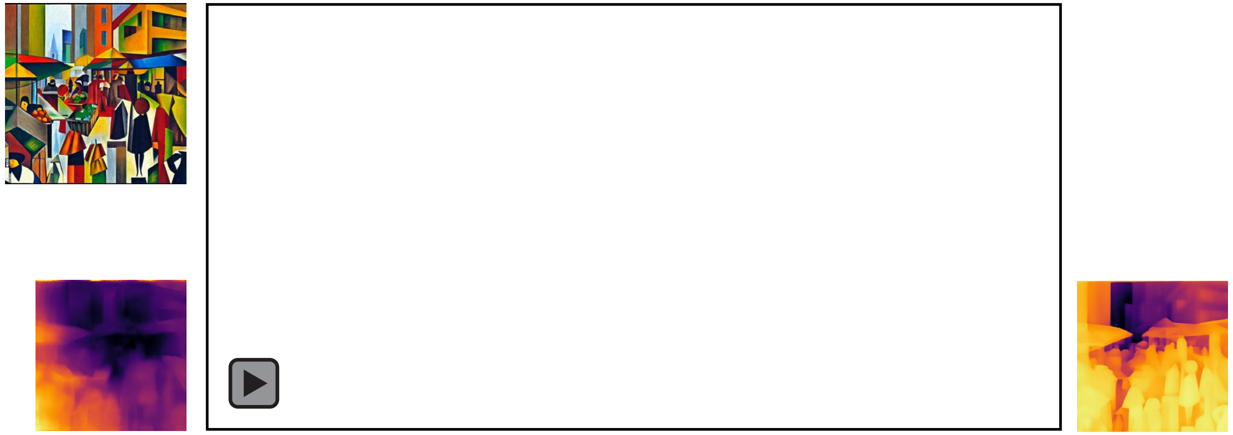
5 steps maintains a good balance between the quality and generalizability.



Single-step model is the denoising U-Net trained to directly predict outputs from inputs images, which struggles to handle arbitrary images.

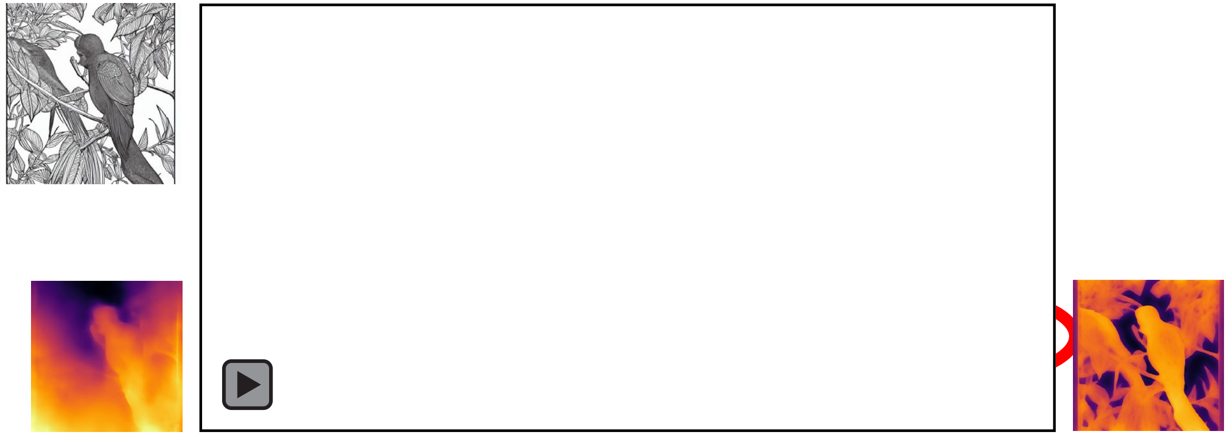
Applications

3D Photo Inpainting



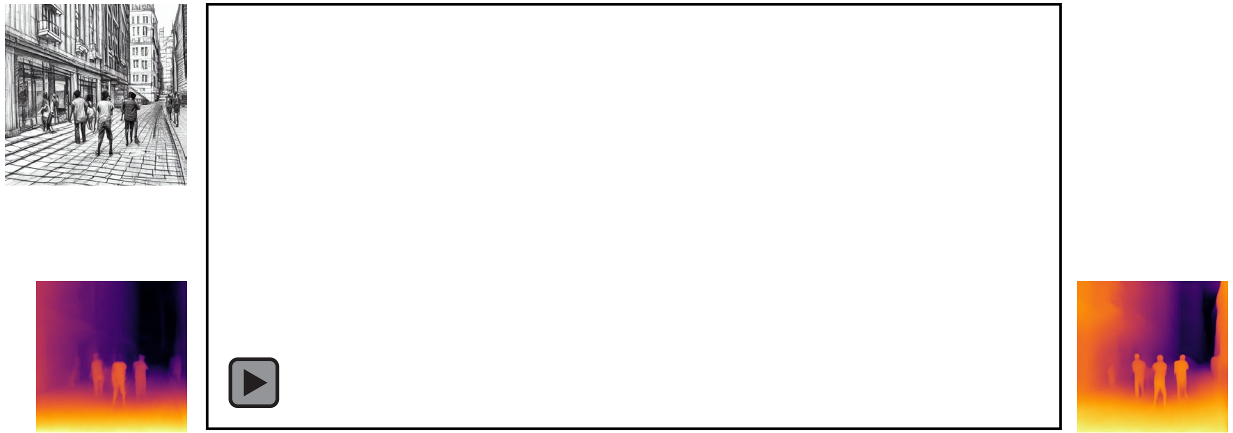
Default

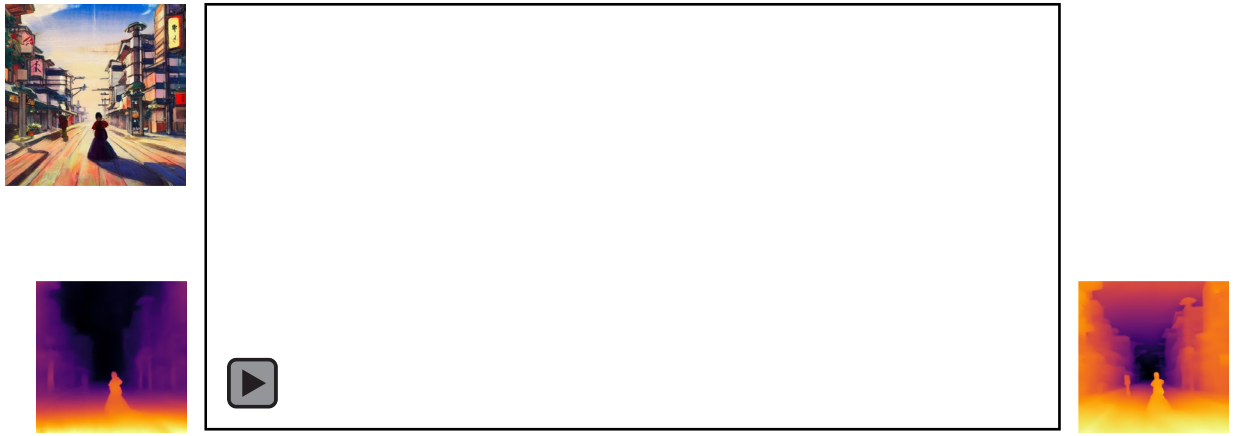
DMP



Default

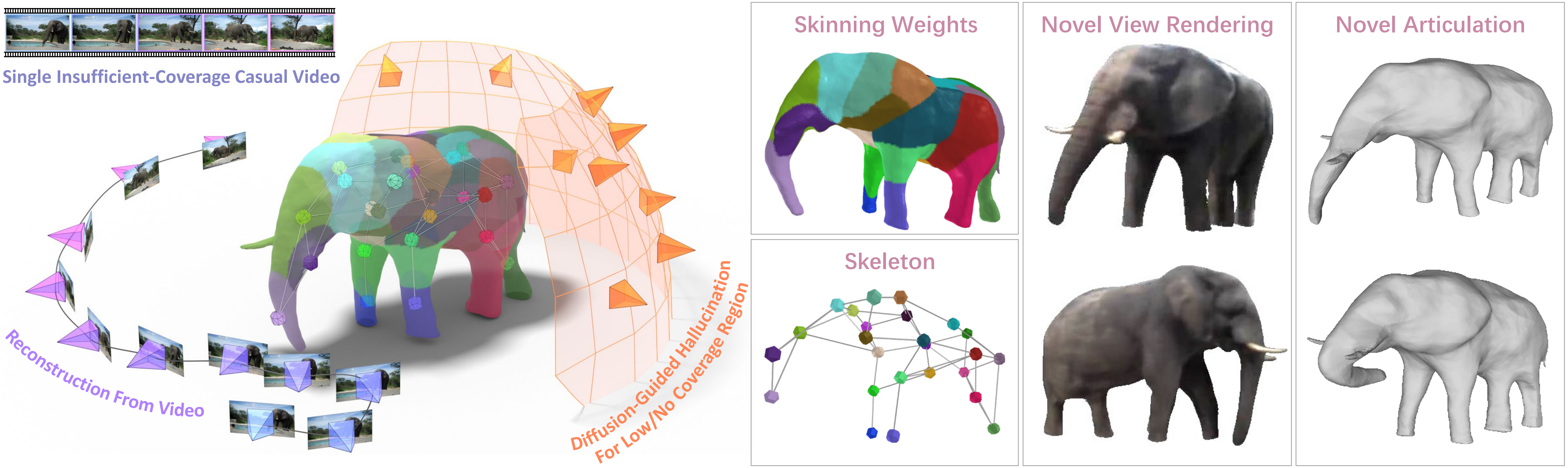
DMP





Summary

- Adapt diffusion models for dense prediction across domains and achieve generalizability
- Address stochastic nature in diffusion process
- Reformulate the diffusion process a series of interpolation
- Achieve state-of-the-art results in dense prediction


Single Insufficient-Coverage Casual Video
Reconstruction From Video
Diffusion-Guided Hallucination For Low/No Coverage Region
Skinning Weights
Skeleton
Novel View Rendering
Novel Articulation

DreaMo: Articulated 3D Reconstruction From A Single Casual Video

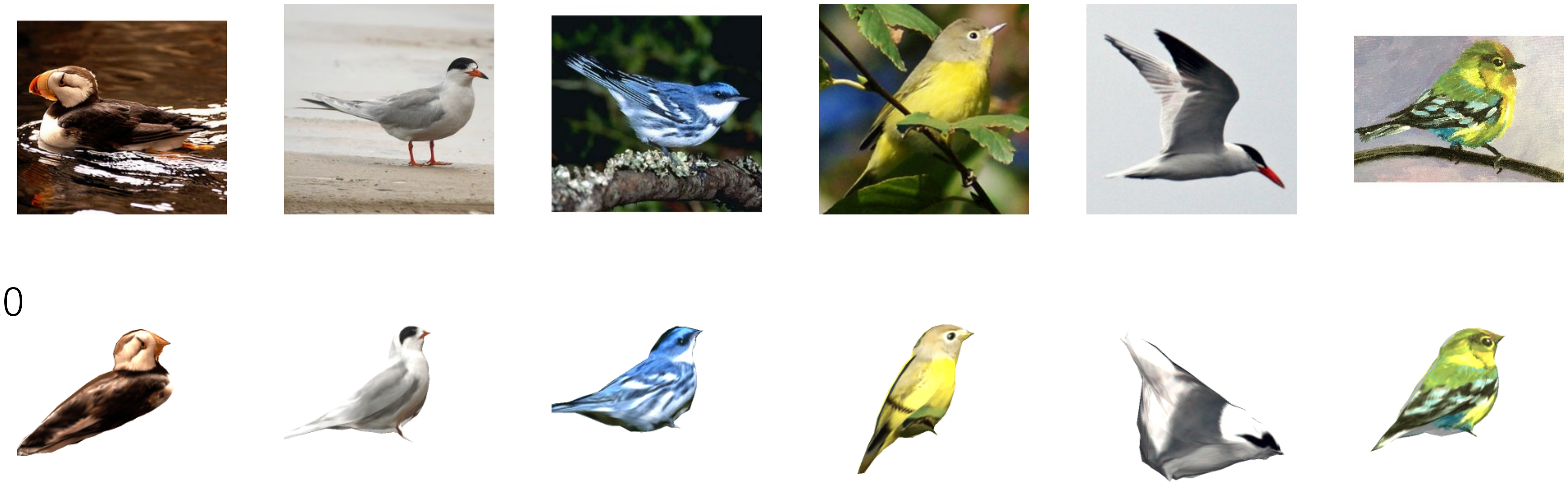
arxiv 2024

Tao Tu[1] Ming-Feng Li[2] Chieh Hubert Lin[3] Yen-Chi Cheng[4] Min Sun[1] Ming-Hsuan Yang[4,5]

[1]NTHU [2]CMU [3]UC Merced [4]UIUC [5]Google

Learning to Recover Non-Rigid 3D Shape

- Self-Supervised Co-Part Segmentation. W.-C. Hung et al. CVPR 2019
- Self-Supervised Single-View 3D Reconstruction X. Li et al. ECCV 2020
- Online Adaption for Consistent Mesh Reconstruction in the Wild. X. Li et al. NeurIPS 2020
- LASSIE: Learning Articulated Shapes from Sparse Image Ensemble. C.-H. Yao et al. NeurIPS 2022
- HI-LASSIE: High-Quality Articulated Shape and Skelton Discovery from Sparse Image Ensemble. C.-H. Yao. et al. CVPR 2023
- ARTIC3D: Learning Robust Articulated 3D Shapes from Noisy Web Image Collection. C.-H. Yao et al. NeurIPS 2023

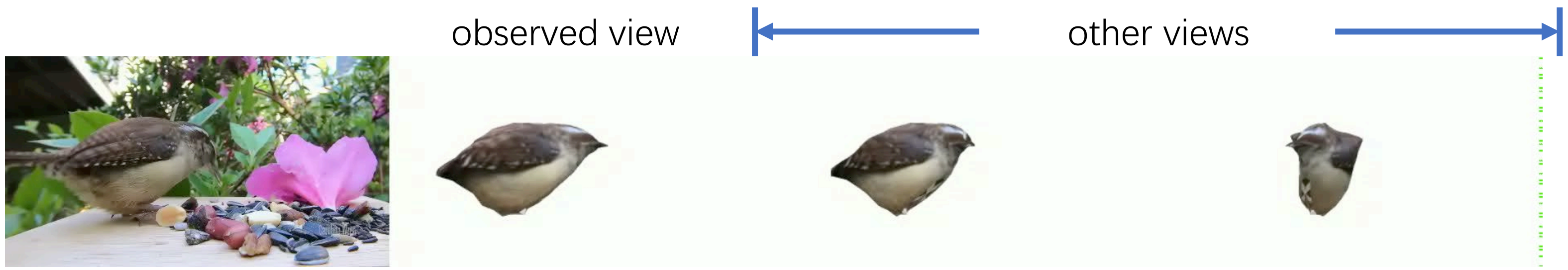


Single image

CMR, ECCV 2020

No annotation
No known
camera pose
Without 3D template

ACMR-vid
test-time
trained



Single video

ACMR, NeurIPS 2020

Temporal consistency

After test-time tuning

LASSIE: Learning Articulated Shape from Sparse Image Ensemble via 3D Part Discovery

NeurIPS 2022

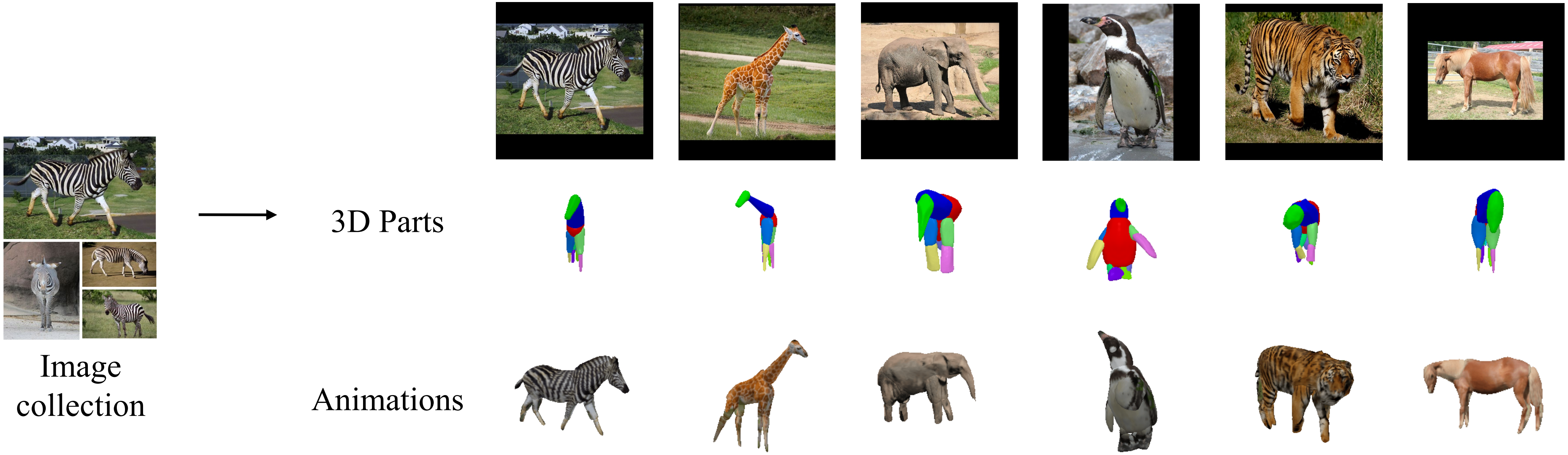


Image collection

3D Parts

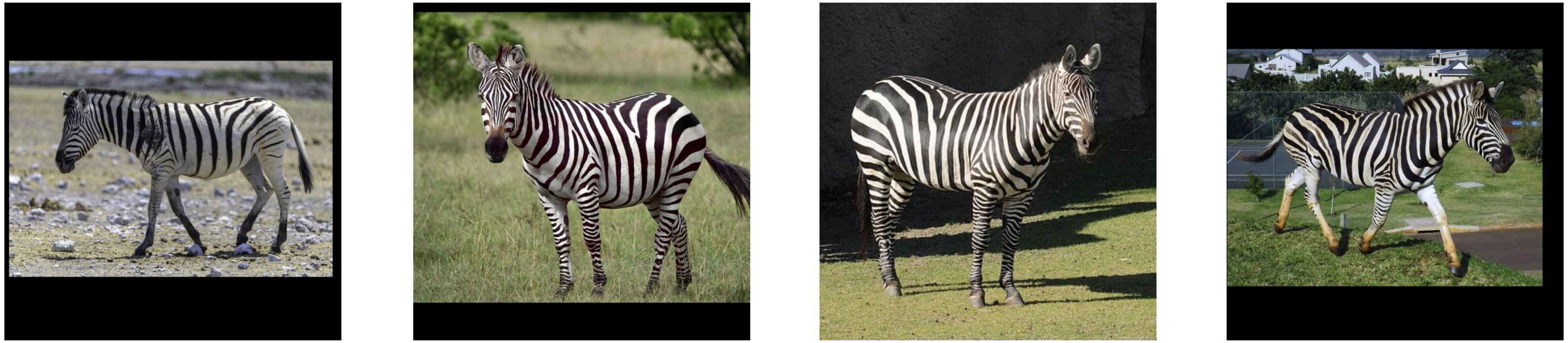
Animations

Based on a few (10-30) images in the wild
Using a generic 3D skeleton
Discover 3D parts in a self-supervised manner

Hi-LASSIE: High-Fidelity Articulated Shape and Skeleton Discovery from Sparse Image Ensemble

CVPR 2023

Image ensemble

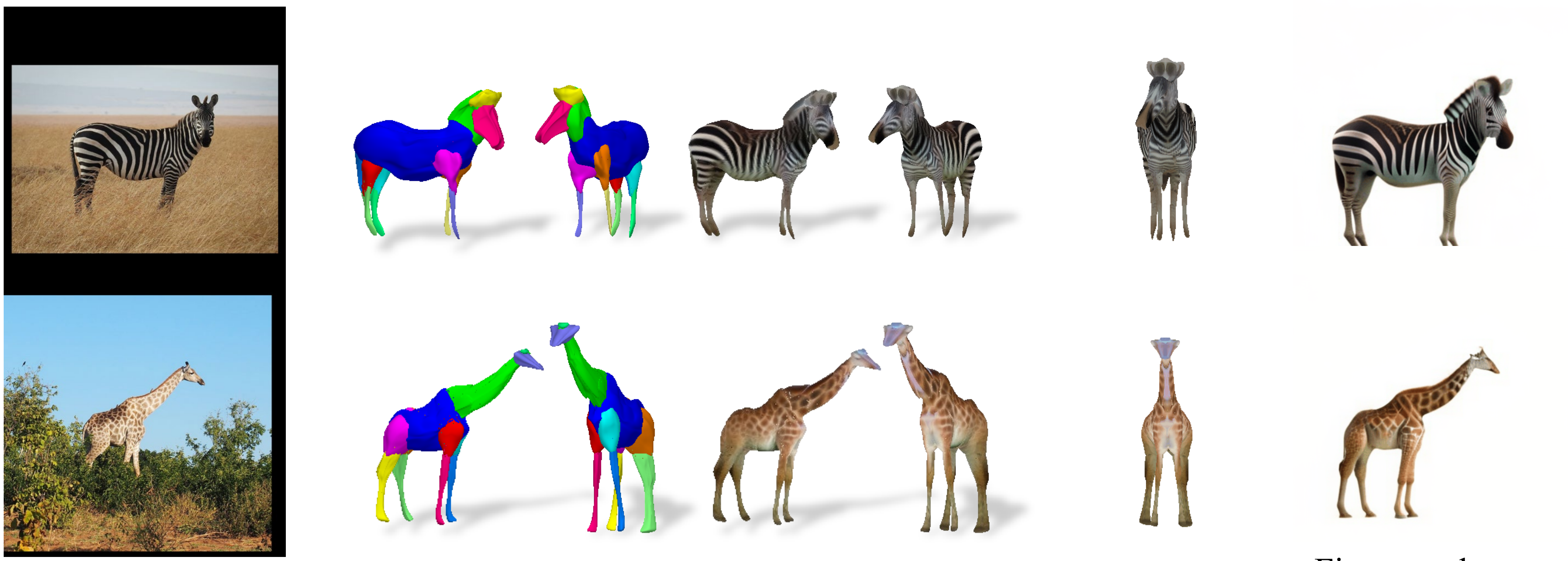


3D articulated shapes
(per-instance)

Automatically estimate class-specific skeleton
Instance-specific optimization for high-quality results

ARTIC3D: Learning Robust Articulated 3D Shapes from Noisy Web Image Collections

NeurIPS 2023



Noisy web images

3D articulated shapes and texture

Animated

Fine-tuned animation

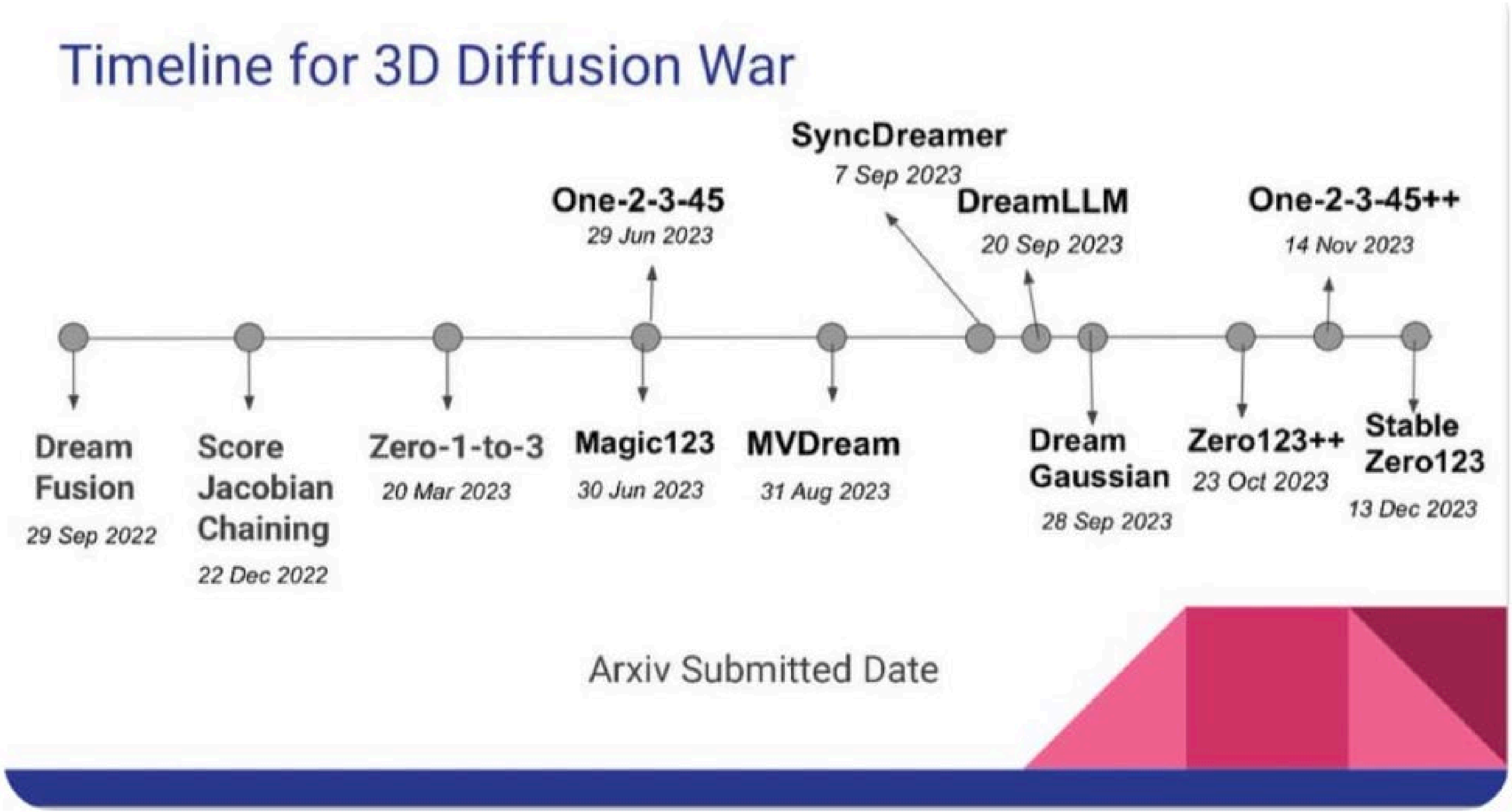
Handle occlusion

Diffusion-guided optimization

Incorporate 3D diffusion priors in 3D surface optimization

Diffusion Models for 3D

- Pre-trained T2I diffusion models generate realistic 3D images



By Seoyeon Stella Yang

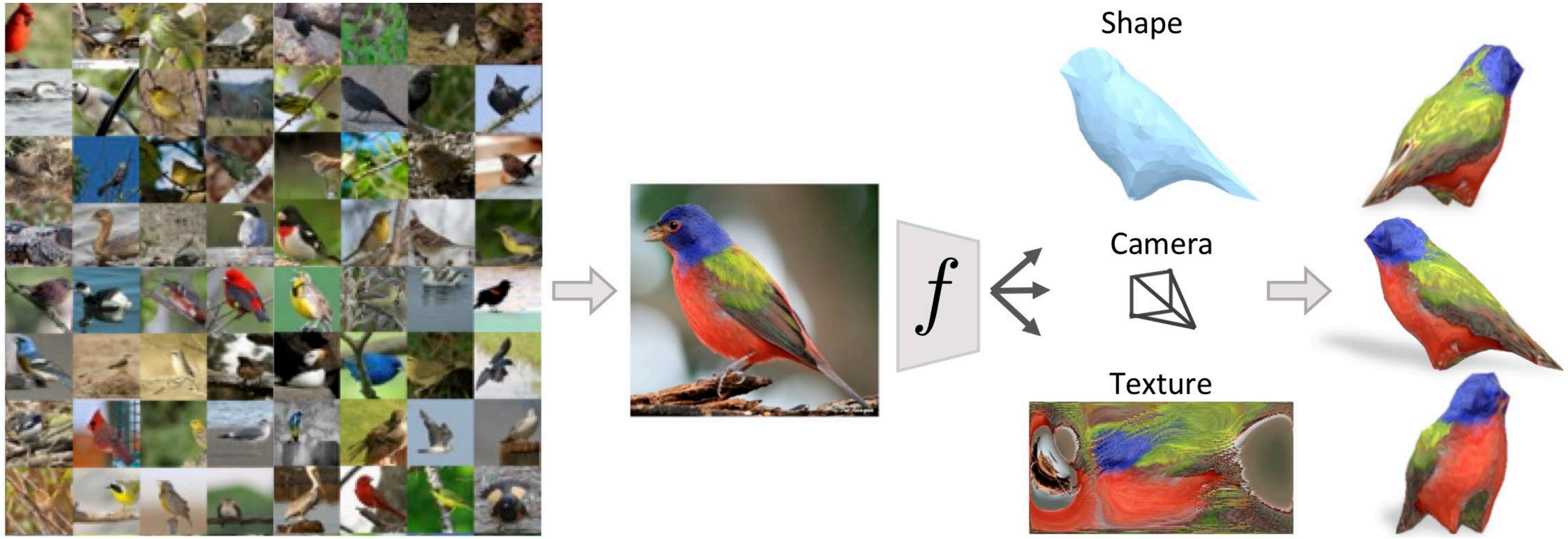
Existing Methods for 3D Animal Reconstruction

[1] A. Kanazawa et al. “Learning Category-Specific Mesh Reconstruction from Image Collections.” ECCV 2018
[2] S. Goel, et al. “Shape and viewpoints without keypoints.” ECCV 2020
[3] X. Li, et al. “Self-supervised Single-view 3D Reconstruction via Semantic Consistency.” ECCV 2020
[4] S. Wu et al. “DOVE: Learning Deformable 3D Objects by Watching Videos.” IJCV 2023
[5] S. Wu et al. “MagicPony: Learning Articulated 3D Animals in the Wild.” CVPR 2023
[6] C.-H. Yao, et al. “LASSIE: Learning Articulated Shape from Sparse Image Ensemble via 3D Part Discovery.” NeurIPS 2022
[7] C.-H. Yao et al. “Hi-LASSIE: High-Fidelity Articulated Shape and Skeleton Discovery from Sparse Image Ensemble.” CVPR 2023
[8] G. Yang, et al. “Building animatable 3d neural models from many casual videos.” CVPR. 2022.
[9] R. Liu, et al. “Zero-1-to-3: Zero-shot One Image to 3D Object.” ICCV 2023

3D Animal Reconstruction

- Existing methods learn to reconstruct from large image collections [1-5]
  - Require a large dataset for each category (e.g., 6000 images for bird [1])

6000 images!!



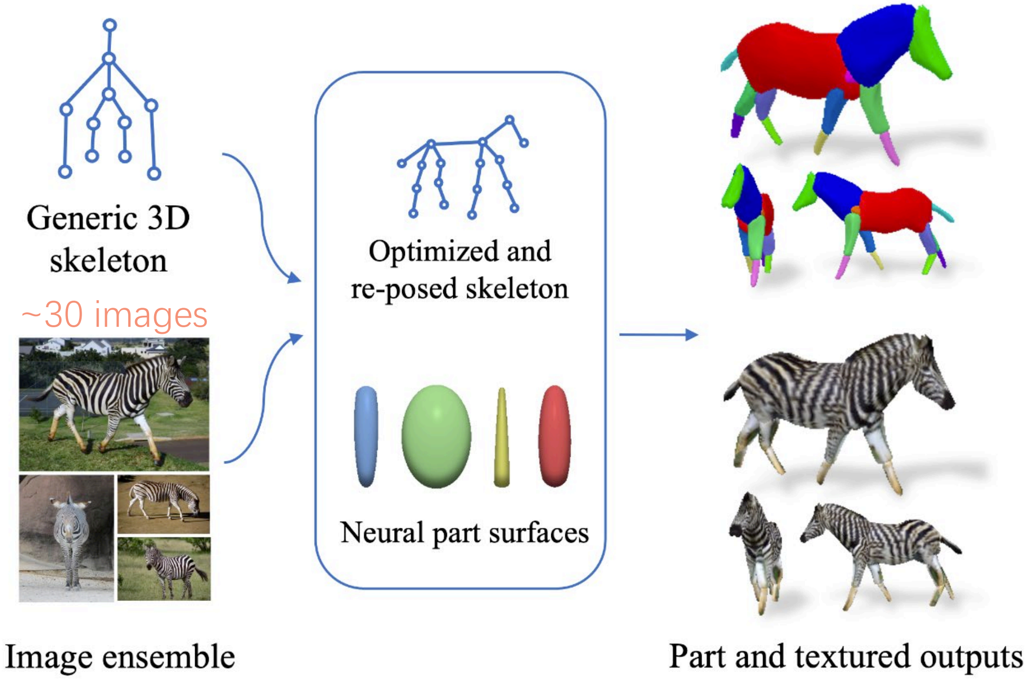
[1] A. Kanazawa, et al. “Learning Category-Specific Mesh Reconstruction from Image Collections.” ECCV 2018

Existing Methods for 3D Animal Reconstruction

[1] A. Kanazawa et al. “Learning Category-Specific Mesh Reconstruction from Image Collections.” ECCV 2018
[2] S. Goel, et al. “Shape and viewpoints without keypoints.” ECCV 2020
[3] X. Li, et al. “Self-supervised Single-view 3D Reconstruction via Semantic Consistency.” ECCV 2020
[4] S. Wu et al. “DOVE: Learning Deformable 3D Objects by Watching Videos.” IJCV 2023
[5] S. Wu et al. “MagicPony: Learning Articulated 3D Animals in the Wild.” CVPR 2023
[6] C.-H. Yao, et al. “LASSIE: Learning Articulated Shape from Sparse Image Ensemble via 3D Part Discovery.” NeurIPS 2022
[7] C.-H. Yao et al. “Hi-LASSIE: High-Fidelity Articulated Shape and Skeleton Discovery from Sparse Image Ensemble.” CVPR 2023
[8] G. Yang, et al. “Building animatable 3d neural models from many casual videos.” CVPR. 2022.
[9] R. Liu, et al. “Zero-1-to-3: Zero-shot One Image to 3D Object.” ICCV 2023

3D Animal Reconstruction

- Existing methods learn to reconstruct from large image collections [1-5]
  - Require a large dataset for each category (e.g., 6000 images for bird [1])
- How to reduce the amount of required data?
  - [6-7] target on recontructing animals with sparse image collections
  - Less natural result shapes and articulations



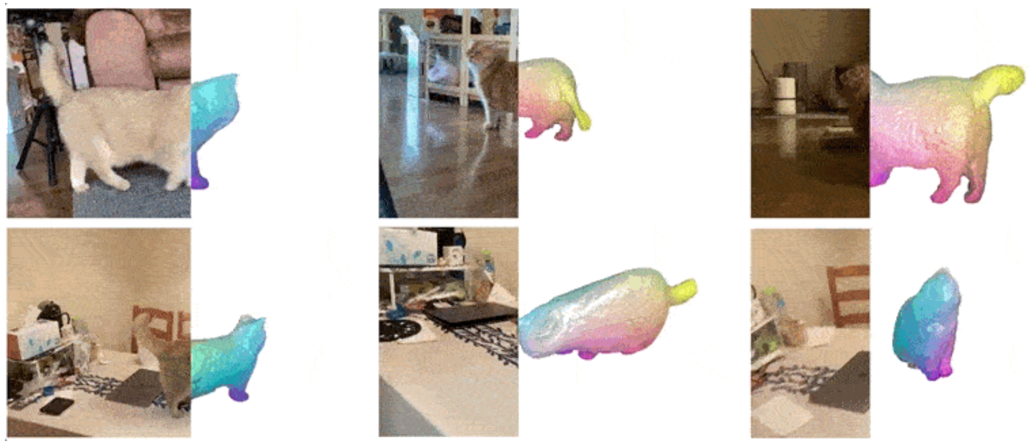
[6] C.-H. Yao, et al. “LASSIE: Learning Articulated Shape from Sparse Image Ensemble via 3D Part Discovery.” NeurIPS 2022

Existing Methods for 3D Animal Reconstruction

[1] A. Kanazawa et al. “Learning Category-Specific Mesh Reconstruction from Image Collections.” ECCV 2018
[2] S. Goel, et al. “Shape and viewpoints without keypoints.” ECCV 2020
[3] X. Li, et al. “Self-supervised Single-view 3D Reconstruction via Semantic Consistency.” ECCV 2020
[4] S. Wu et al. “DOVE: Learning Deformable 3D Objects by Watching Videos.” IJCV 2023
[5] S. Wu et al. “MagicPony: Learning Articulated 3D Animals in the Wild.” CVPR 2023
[6] C.-H. Yao, et al. “LASSIE: Learning Articulated Shape from Sparse Image Ensemble via 3D Part Discovery.” NeurIPS 2022
[7] C.-H. Yao et al. “Hi-LASSIE: High-Fidelity Articulated Shape and Skeleton Discovery from Sparse Image Ensemble.” CVPR 2023
[8] G. Yang, et al. “Building animatable 3d neural models from many casual videos.” CVPR. 2022.
[9] R. Liu, et al. “Zero-1-to-3: Zero-shot One Image to 3D Object.” ICCV 2023

3D Animal Reconstruction

- Existing methods learn to reconstruct from large image collections [1-5]
  - Require a large dataset for each category (e.g., 6000 images for bird [1])
- How to reduce the amount of required data?
  - [6-7] target on recontructing animals with sparse image collections
  - Less natural shapes and articulations
- Learn from videos
  - Videos offer more temporal information than still images, aiding the model in learning articulation through continuous movements
  - BANMO [8] shows promising reconstruction results from ~10 casually captured videos
  - Require videos with dense camera viewpoint coverage



[8] G. Yang, et al. “Building animatable 3d neural models from many casual videos.” CVPR. 2022.

Existing Methods for 3D Animal Reconstruction

[1] A. Kanazawa et al. “Learning Category-Specific Mesh Reconstruction from Image Collections.” ECCV 2018
[2] S. Goel, et al. “Shape and viewpoints without keypoints.” ECCV 2020
[3] X. Li, et al. “Self-supervised Single-view 3D Reconstruction via Semantic Consistency.” ECCV 2020
[4] S. Wu et al. “DOVE: Learning Deformable 3D Objects by Watching Videos.” IJCV 2023
[5] S. Wu et al. “MagicPony: Learning Articulated 3D Animals in the Wild.” CVPR 2023
[6] C.-H. Yao, et al. “LASSIE: Learning Articulated Shape from Sparse Image Ensemble via 3D Part Discovery.” NeurIPS 2022
[7] C.-H. Yao et al. “Hi-LASSIE: High-Fidelity Articulated Shape and Skeleton Discovery from Sparse Image Ensemble.” CVPR 2023
[8] G. Yang, et al. “Building animatable 3d neural models from many casual videos.” CVPR. 2022.
[9] R. Liu, et al. “Zero-1-to-3: Zero-shot One Image to 3D Object.” ICCV 2023

3D Animal Reconstruction

- Existing methods learn to reconstruct from large image collections [1-5]
    - Require a large dataset for each category (e.g., 6000 images for bird [1])
- How to reduce the amount of required data?
    - [6-7] target on recontructing animals with sparse image collections
    - Less natural result shapes and articulations
- Learn from videos
    - Videos offer more temporal information than still images, aiding the model in learning articulation through continuous movements
    - BANMO [8] shows promising reconstruction results from ~10 casually captured videos
    - Require videos with dense camera viewpoint coverage
- Could we directly reconstruct animals from single videos?

[8] G. Yang, et al. “Building animatable 3d neural models from many casual videos.” CVPR. 2022.

Existing Methods for 3D Animal Reconstruction

[1] A. Kanazawa et al. “Learning Category-Specific Mesh Reconstruction from Image Collections.” ECCV 2018
[2] S. Goel, et al. “Shape and viewpoints without keypoints.” ECCV 2020
[3] X. Li, et al. “Self-supervised Single-view 3D Reconstruction via Semantic Consistency.” ECCV 2020
[4] S. Wu et al. “DOVE: Learning Deformable 3D Objects by Watching Videos.” IJCV 2023
[5] S. Wu et al. “MagicPony: Learning Articulated 3D Animals in the Wild.” CVPR 2023
[6] C.-H. Yao, et al. “LASSIE: Learning Articulated Shape from Sparse Image Ensemble via 3D Part Discovery.” NeurIPS 2022
[7] C.-H. Yao et al. “Hi-LASSIE: High-Fidelity Articulated Shape and Skeleton Discovery from Sparse Image Ensemble.” CVPR 2023
[8] G. Yang, et al. “Building animatable 3d neural models from many casual videos.” CVPR. 2022.
[9] R. Liu, et al. “Zero-1-to-3: Zero-shot One Image to 3D Object.” ICCV 2023

3D Animal Reconstruction

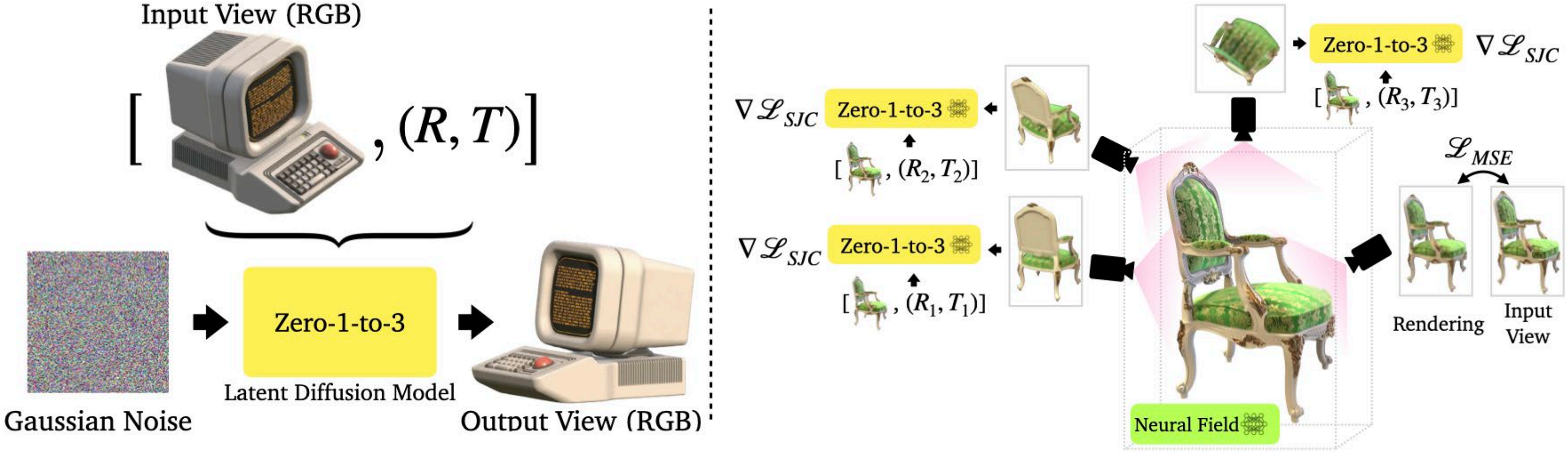
- Existing methods learn to reconstruct from large image collections [1-5]
  - Require a large dataset for each category (e.g., 6000 images for bird [1])
- How to reduce the amount of required data?
  - [6-7] target on recontructing animals with sparse image collections
  - Less natural result shapes and articulations
- Learn from videos
  - Videos offer more temporal information than still images, aiding the model in learning articulation through continuous movements
  - [8] shows promising reconstruction results from ~10 casually captured videos
  - Require videos with dense camera viewpoint coverage
- Could we directly reconstruct animals from single videos?
  - Hallucinating low-coverage regions using a view-conditioned diffusion model— Zero-1-to-3 [9]

[8] G. Yang, et al. “Building animatable 3d neural models from many casual videos.” CVPR. 2022.

Existing Methods for 3D Animal Reconstruction

Zero-1-to-3

- Diffusion model for control of camera viewpoint in novel view synthesis
- 3D reconstruction of a static object from a single image by imagining different views
- Difficult to apply to moving and deformable objects such as animals



Novel View Synthesis

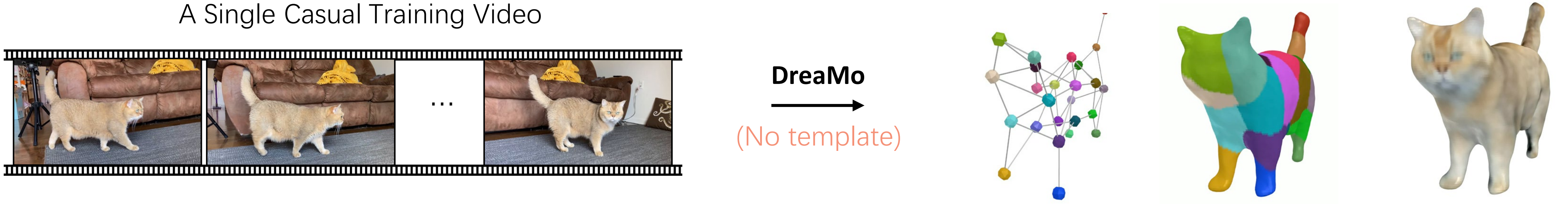
3D Reconstruction

[9] R. Liu, et al. "Zero-1-to-3: Zero-shot One Image to 3D Object." ICCV 2023

Articulated 3D Reconstruction From A Single Video

Problem Formulation

- Goal: reconstructing an articulated 3D model from an Internet video
- Input: an Internet video capturing a deformable object
- Output: an articulated 3D model with skeleton, skinning weight, 3D shape, and color
- No pre-defined shape template from 3D scans



(Inadequate View Coverage)

A Single Casual Training Video

DreaMo

(No template)

Skeleton

Skinning Weights

Shape & Color

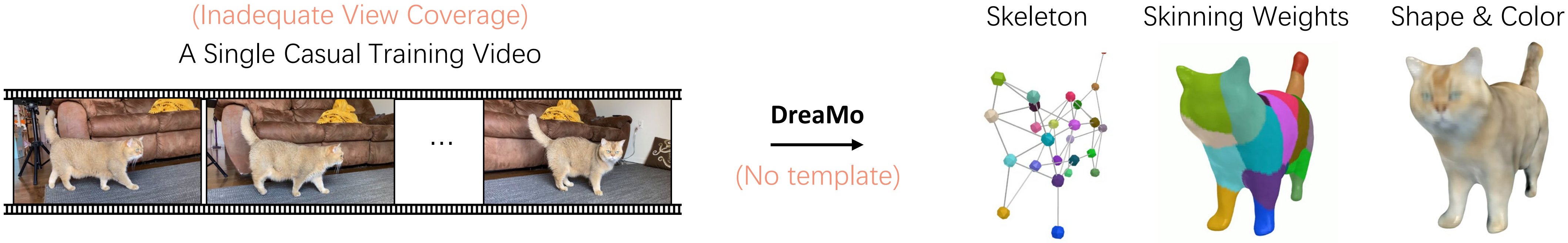
Articulated 3D Reconstruction From A Single Video

Problem Formulation

- Goal: reconstructing an articulated 3D model from an Internet video
- Input: an Internet video capturing a deformable object
- Output: an articulated 3D model with skeleton, skinning weight, 3D shape, and color
- No pre-defined shape template from 3D scans

Why Is It Important?

- Reconstructing arbitrary animals in 3D from the Internet can provide diverse 3D assets for various applications such as movie production, gaming, and virtual reality



(Inadequate View Coverage)

A Single Casual Training Video

DreaMo

(No template)

Skeleton

Skinning Weights

Shape & Color

Articulated 3D Reconstruction From A Single Video

Problem Formulation

- Goal: reconstructing an articulated 3D model from an Internet video
- Input: an Internet video capturing a deformable object
- Output: an articulated 3D model with skeleton, skinning weight, 3D shape, and color
- No pre-defined shape template from 3D scans

Why Is It Important?

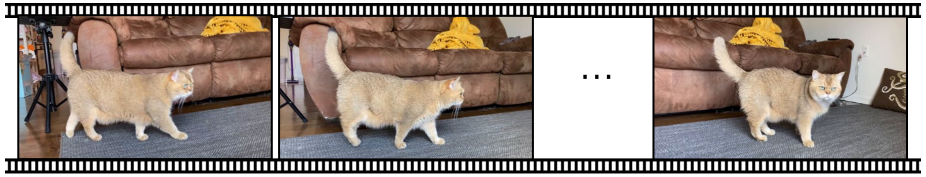
- Reconstructing arbitrary animals in 3D from the Internet can provide diverse 3D assets for various applications such as movie production, gaming, and virtual reality

Why Is It Challenging?

- Internet videos often lack sufficient view coverage for 3D reconstruction
- Reconstructing plausible 3D shapes without shape templates in such a demanding setting is difficult

(Inadequate View Coverage)

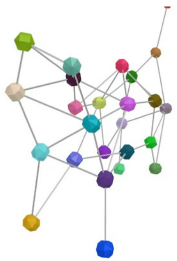
A Single Casual Training Video



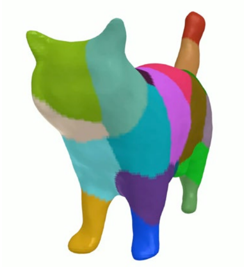
DreaMo →

(No template)

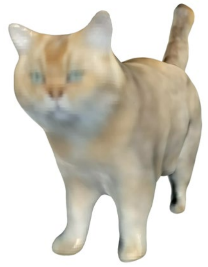
Skeleton



Skinning Weights



Shape & Color



DreaMo Overview

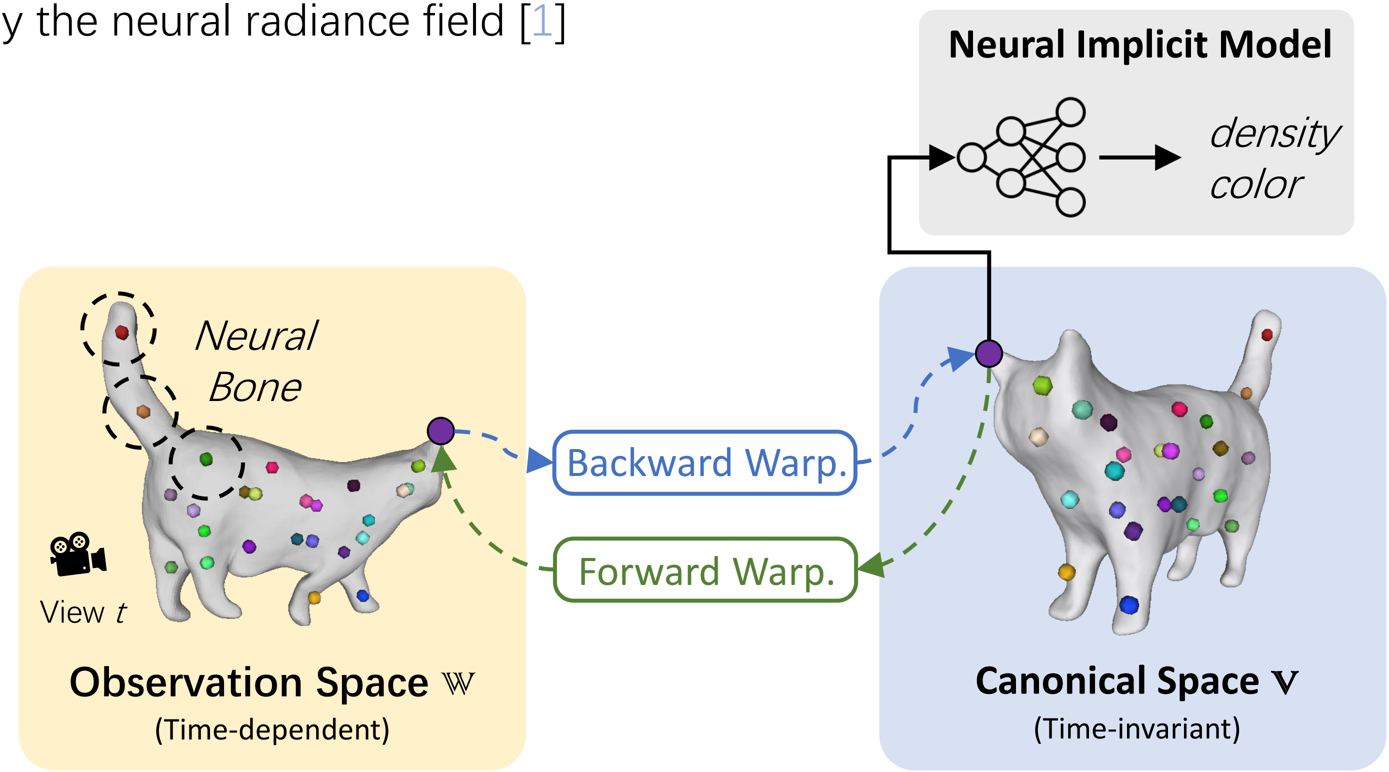
[1] B. Mildenhall, et al. Nerf: Representing scenes as neural radiance fields for view synthesis. ECCV. 2020.
[2] G. Yang, et al. Building animatable 3d neural models from many casual videos. CVPR. 2022.
[3] A. Jacobson, et al. Skinning: Real-time shape deformation. SIGGRAPH Courses. 2014.

Implicit 3D Model (Canonical Space)

- Represent a 3D reconstruction target in a resting pose
- Implicitly model the 3D shape and color by the neural radiance field [1]

Warping Model

- Mapping between observation space and canonical space
- Object deformation
  - Neural bones represent articulation [2]
  - Linear blend skinning [3]
- Camera transformation
  - Global transformation model



How to Address The Low-Coverage Regions?

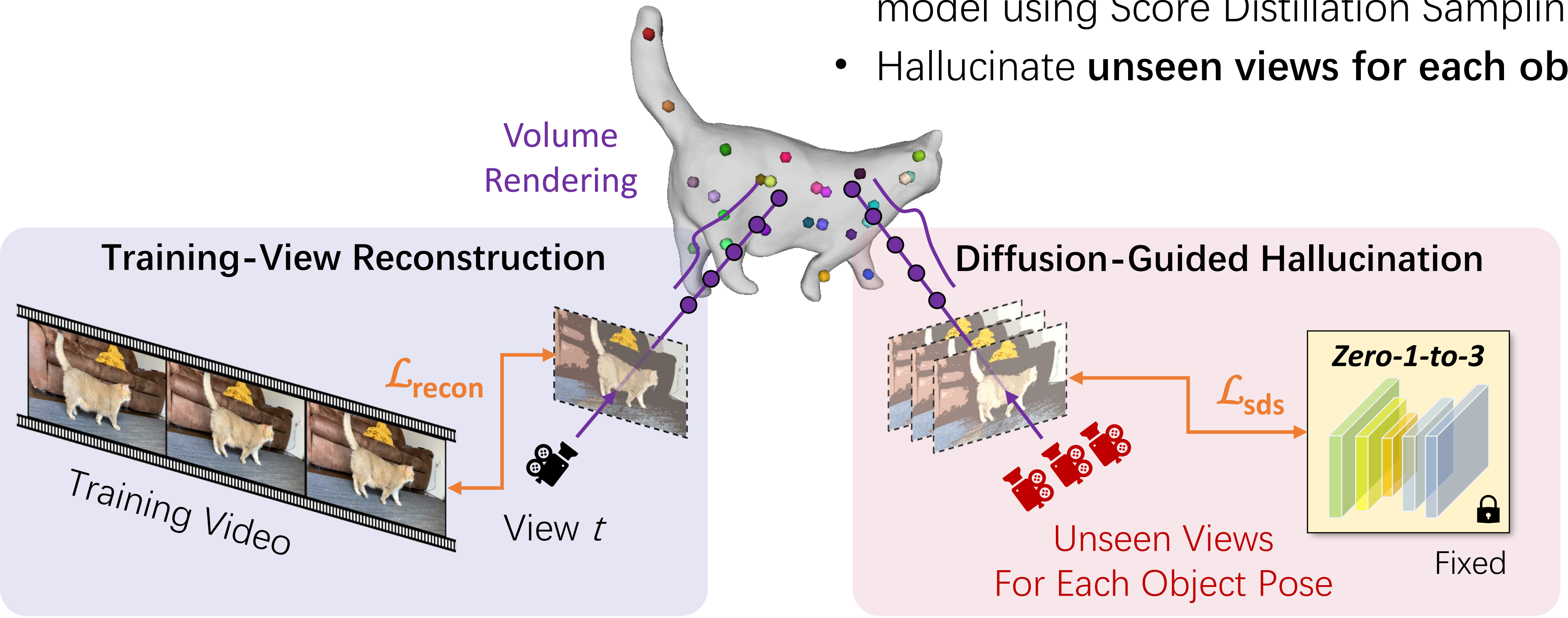
[4] R. Liu, et al. "Zero-1-to-3: 597 Zero-shot one image to 3d object." ICCV. 2023.
[5] B. Poole, et al. "Dreamfusion: Text-to-3d using 2d diffusion." ICLR. 2023.

Training-view Reconstruction

- Insufficient for reconstructing plausible shapes in low-coverage regions

Diffusion-guided Hallucination

- Leverage zero-1-to-3 conditioned on source video frame and came pose to synthesize novel view
- Distill synthetic supervisions into the 3D reconstruction model using Score Distillation Sampling (SDS)
- Hallucinate **unseen views for each object pose**



How to Avoid Irregular Reconstructed Shapes?

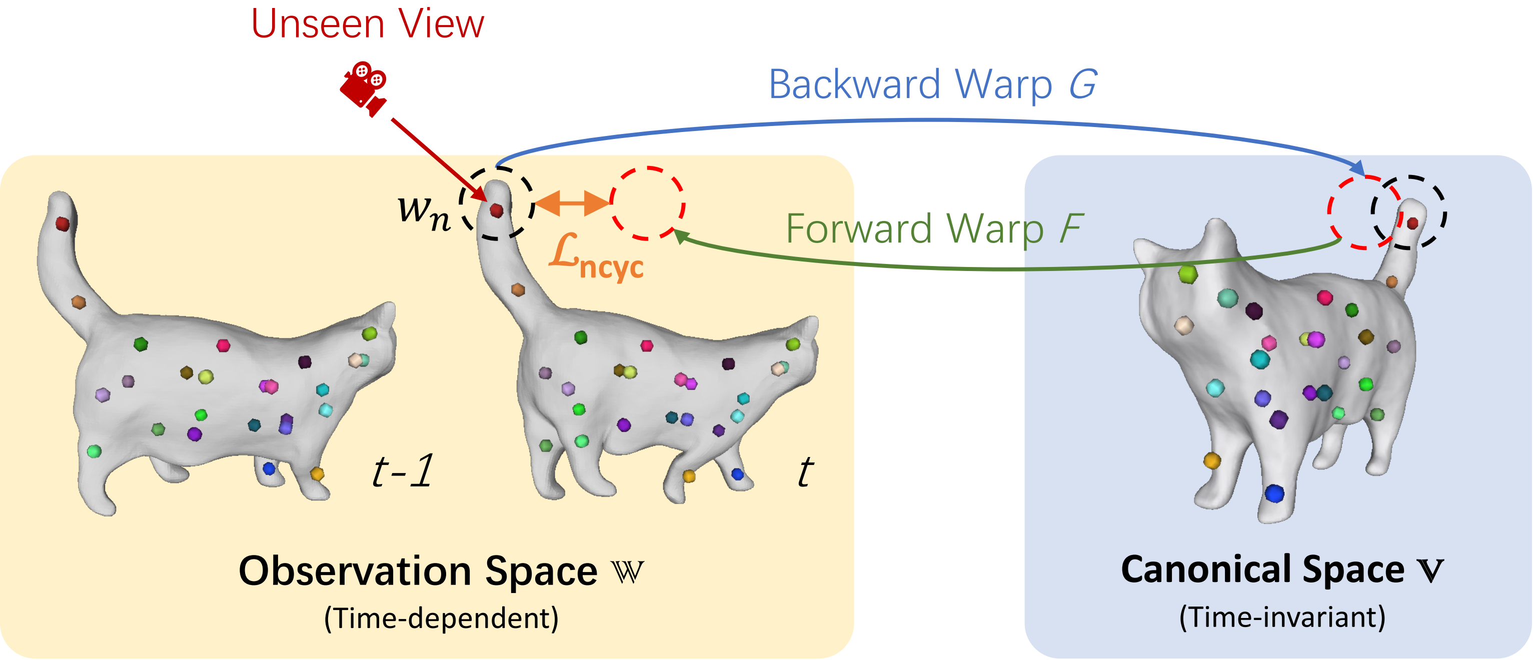
Regularization

- Novel-view cycle consistency
- Smooth articulation
- Surface constraint

Minimize the warping error to encourage the inverse relationship between forward and backward warpings for unseen views

$$\mathcal{L}_{\text{ncyc}} = \sum_{n} \tau_n \| w_n - F(G(w_n, t), t) \|_2$$

Sample a point along the ray from a novel view



How to Avoid Irregular Reconstructed Shapes?

Regularization

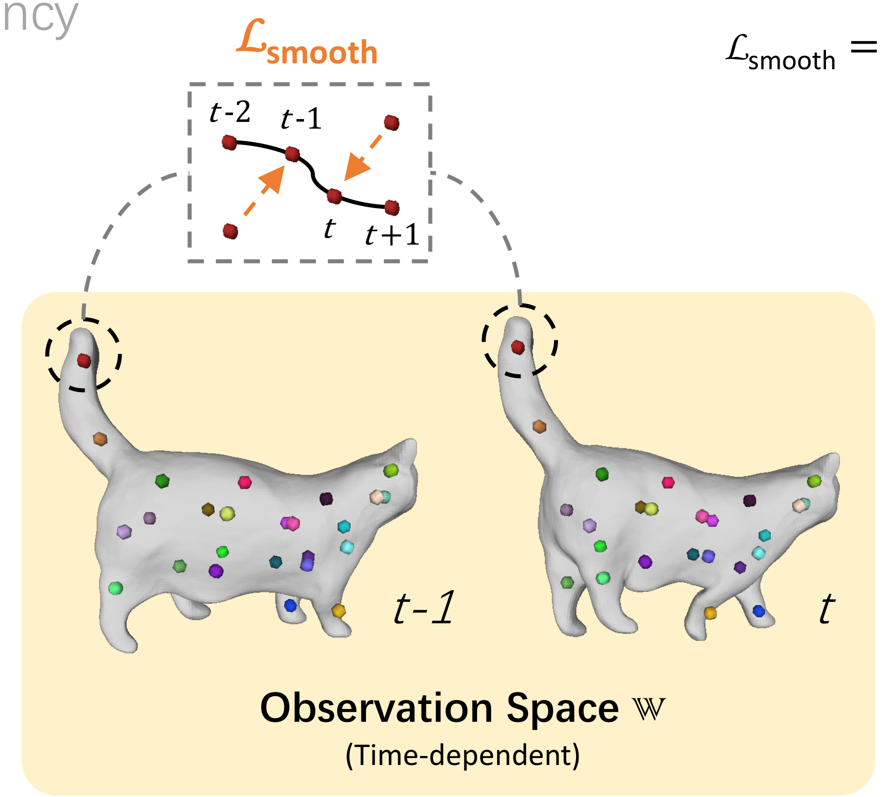
- Novel-view cycle consistency
- Smooth articulation
- Surface constraint

Regularize variations in rotations *R* and translations ***s*** between consecutive time steps

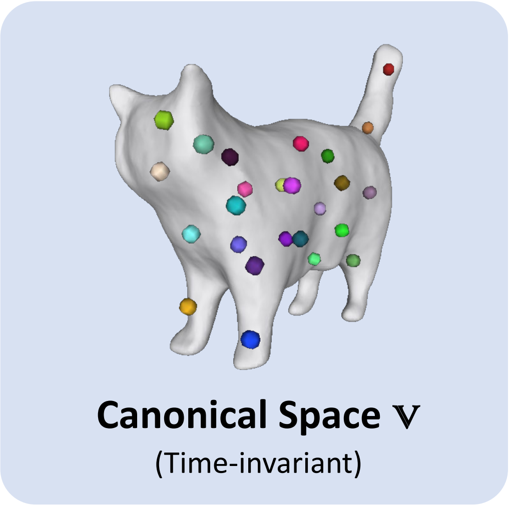
$$\mathcal{L}_{\text{smooth}} = \sum_{b=1,t=1}^{B,T-1} \frac{\operatorname{ang}\left(\boldsymbol{R}_b^t, \boldsymbol{R}_b^{t+1}\right) + \left\|\boldsymbol{s}_b^t - \boldsymbol{s}_b^{t+1}\right\|_2}{B(T-1)}$$

Learned transitions of bones in the low-coverage or self-occluded regions often exhibit unnatural jiggles

Introduce smooth transition for smooth bone motion



Observation Space $\mathbb{W}$
(Time-dependent)



Canonical Space $\mathbb{V}$
(Time-invariant)

How to Avoid Irregular Reconstructed Shapes?

Regularization

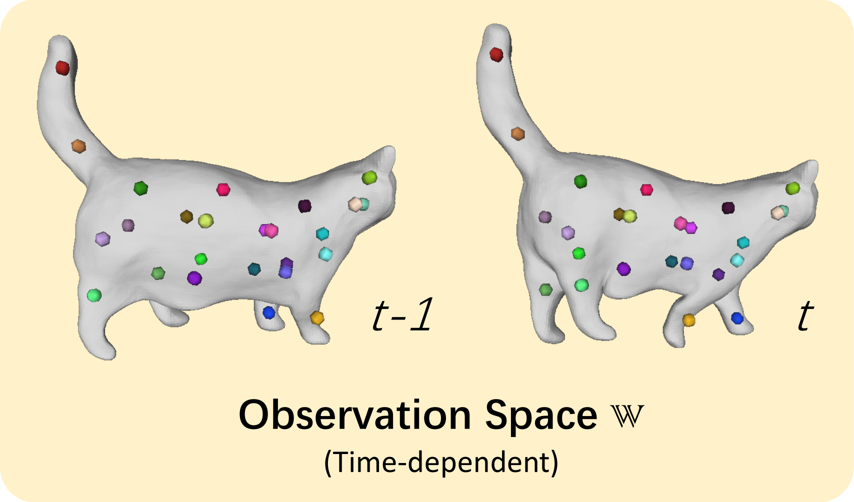
- Novel-view cycle consistency
- Smooth articulation
- Surface constraint

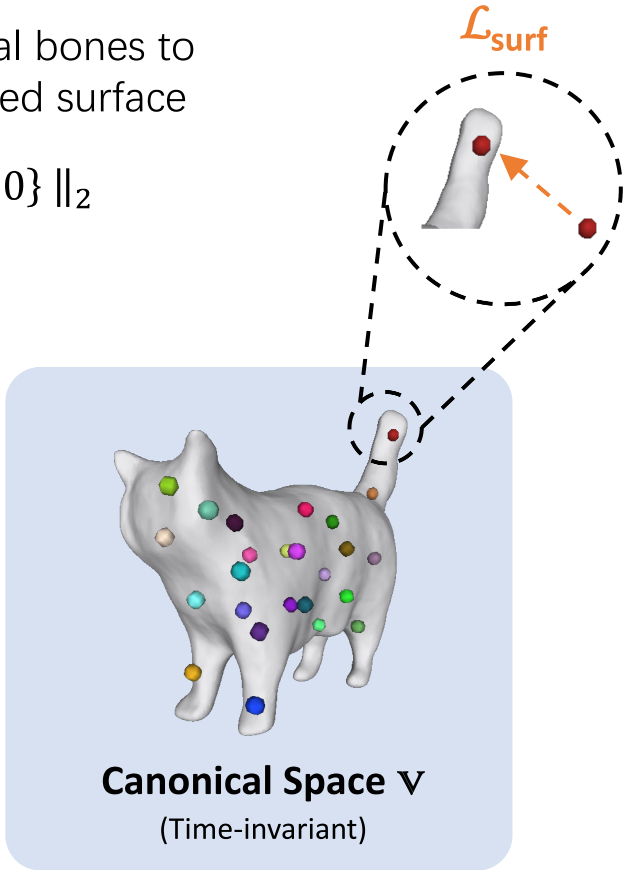
Encourage the neural bones to stay within the learned surface

$$\mathcal{L}_{\text{surf}} = \|\max\{\delta, 0\}\|_2$$

Neural bones may scatter all over the space

Add constraints to keep neural bones beneath learned surface

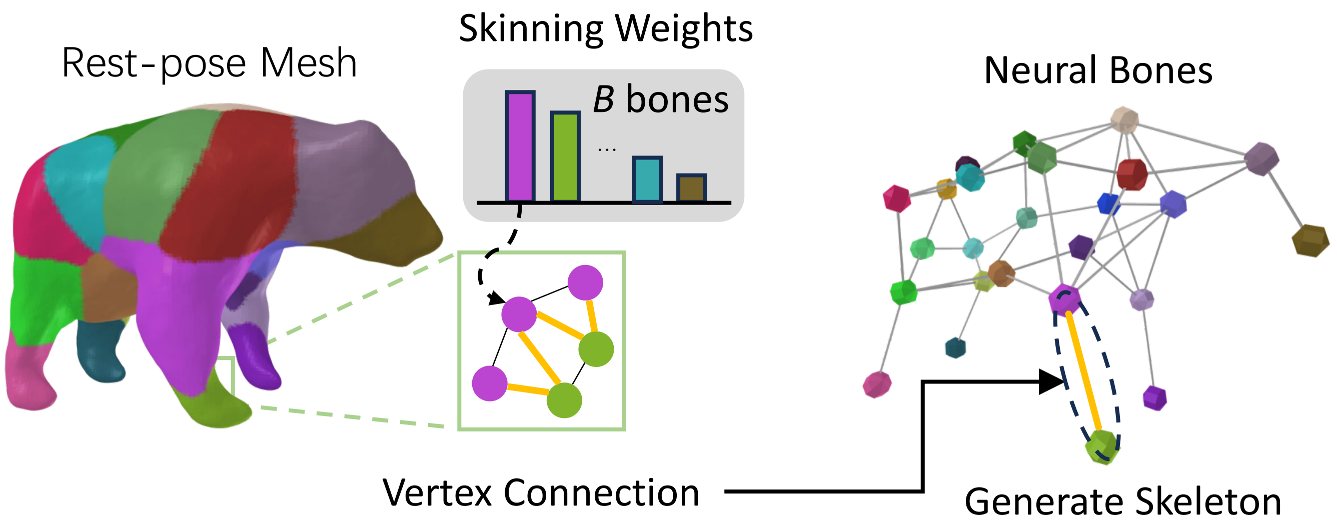




Interpretable Skeleton Generation

Skeleton Extraction From The Implicit Model

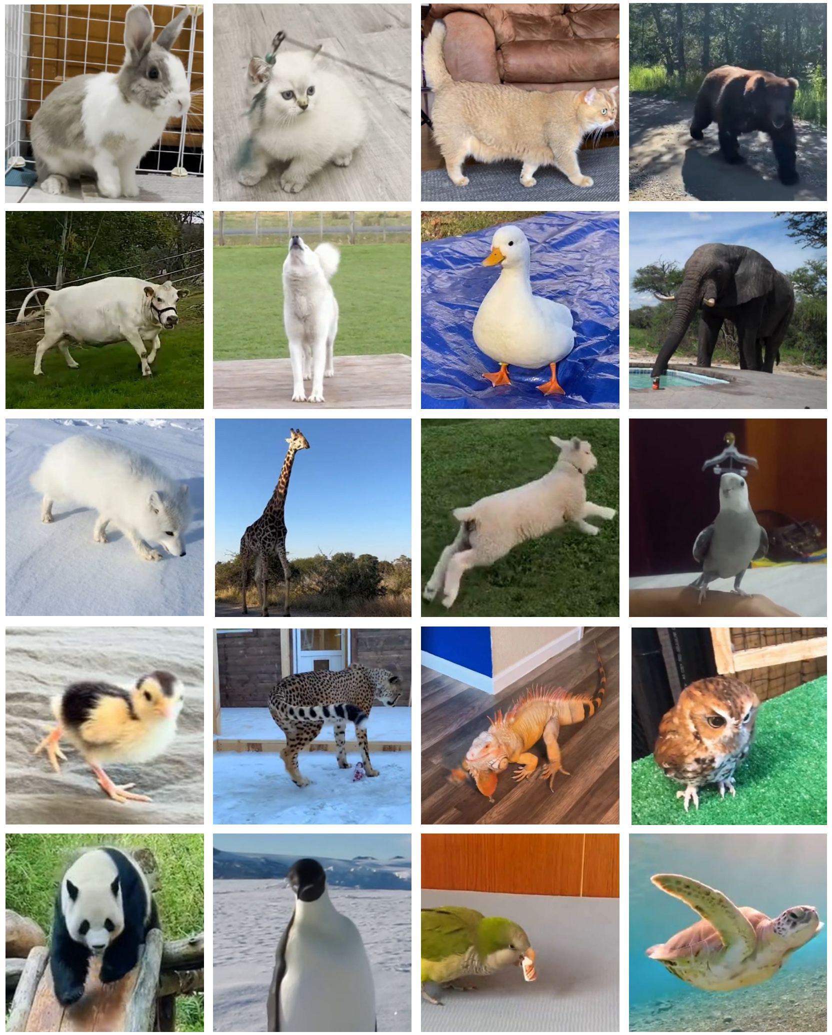
- Extract rest-pose mesh from the implicit 3D model using marching cube
- Assign each vertex (skin point) to the neural bone of the highest skinning weight
- Establish an edge between bones if there is a sufficient skin point connection



Experimental Results

Self-collected Dataset

- 42 animal video clips from the Internet
- 28 different species
- Average azimuth viewpoint coverage: 31%
- Average video duration: 15.7 seconds

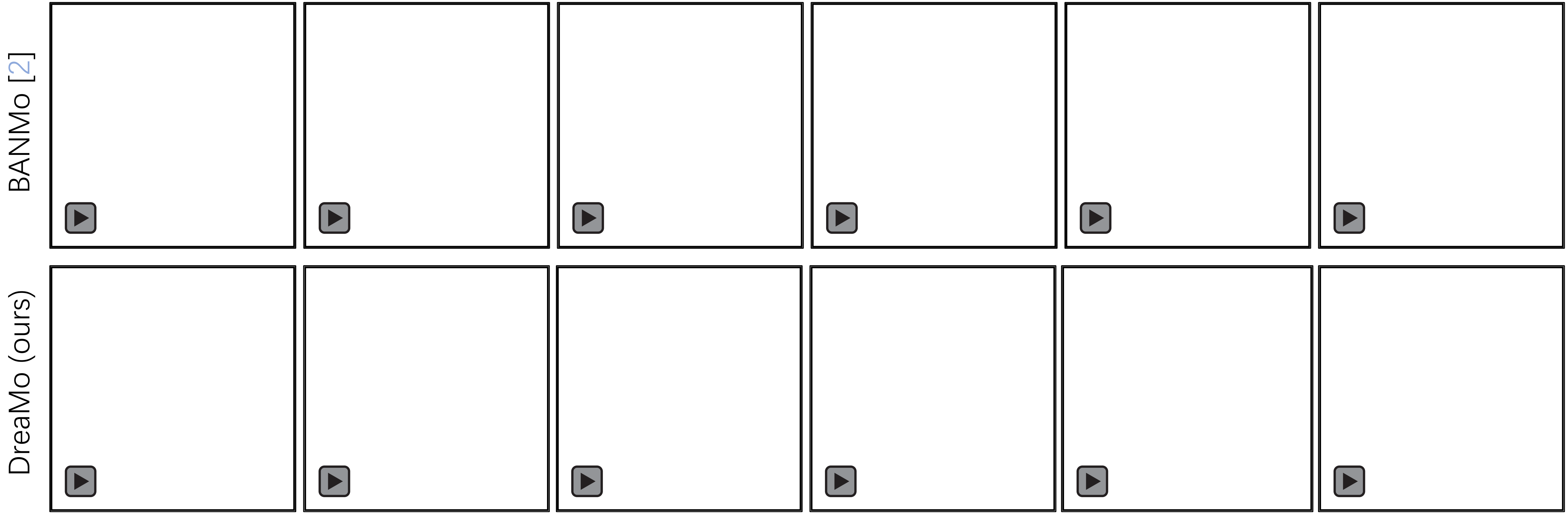


Some image crops of the target subjects from the dataset.

Experimental Results

[2] G. Yang, et al. Building animatable 3d neural models from many casual videos. CVPR. 2022.

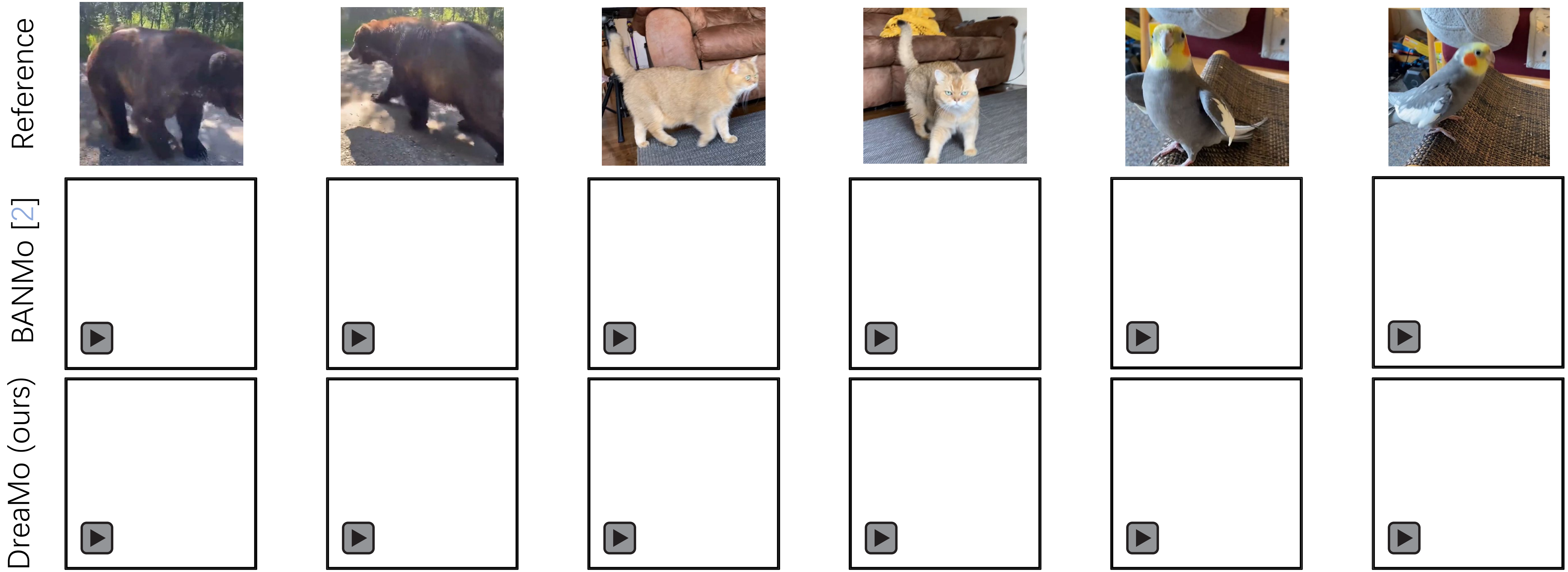
Novel View Rendering (RGB & Depth)



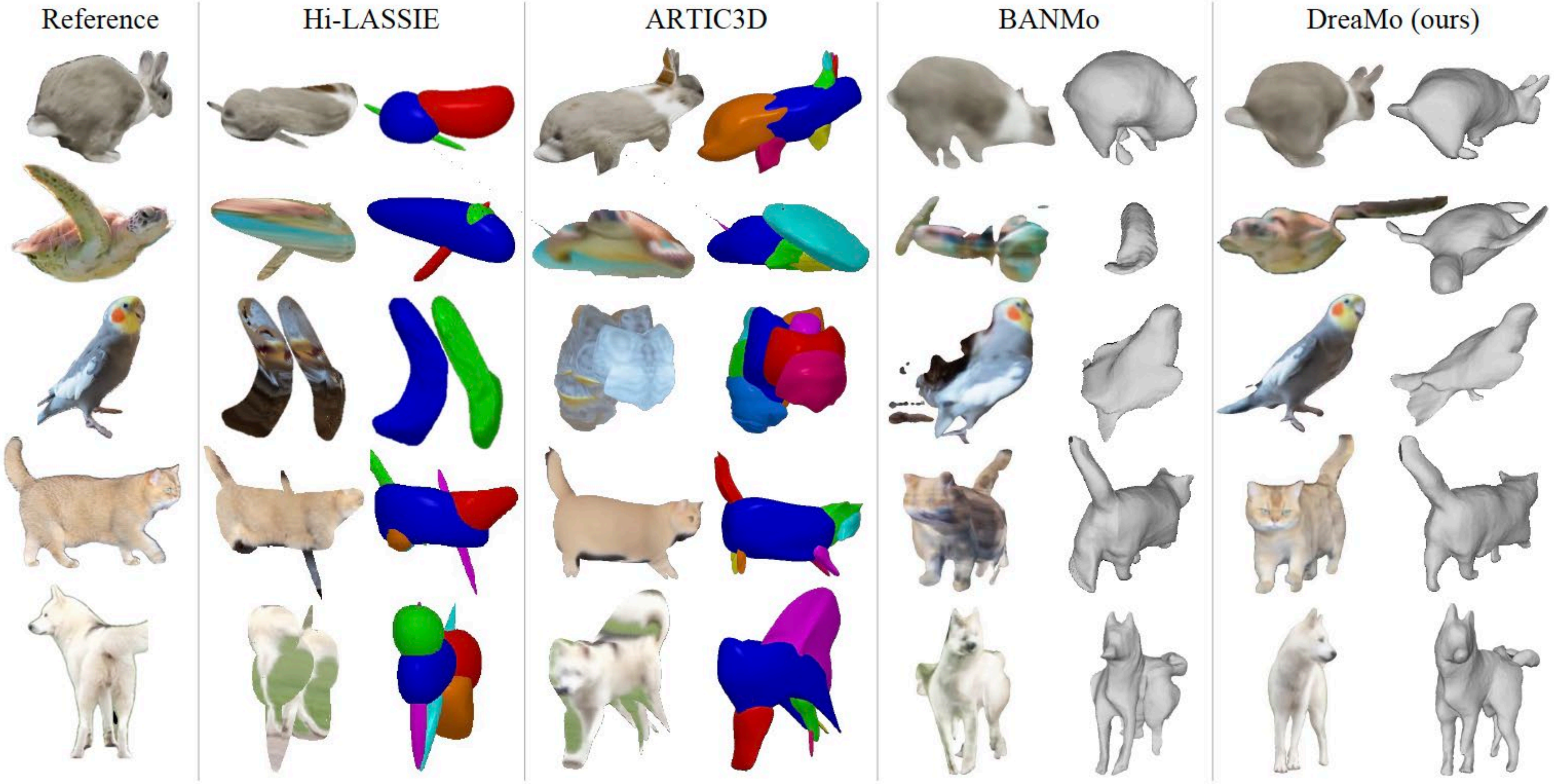
Experimental Results

[2] G. Yang, et al. Building animatable 3d neural models from many casual videos. CVPR. 2022.

3D Reconstructed Shape

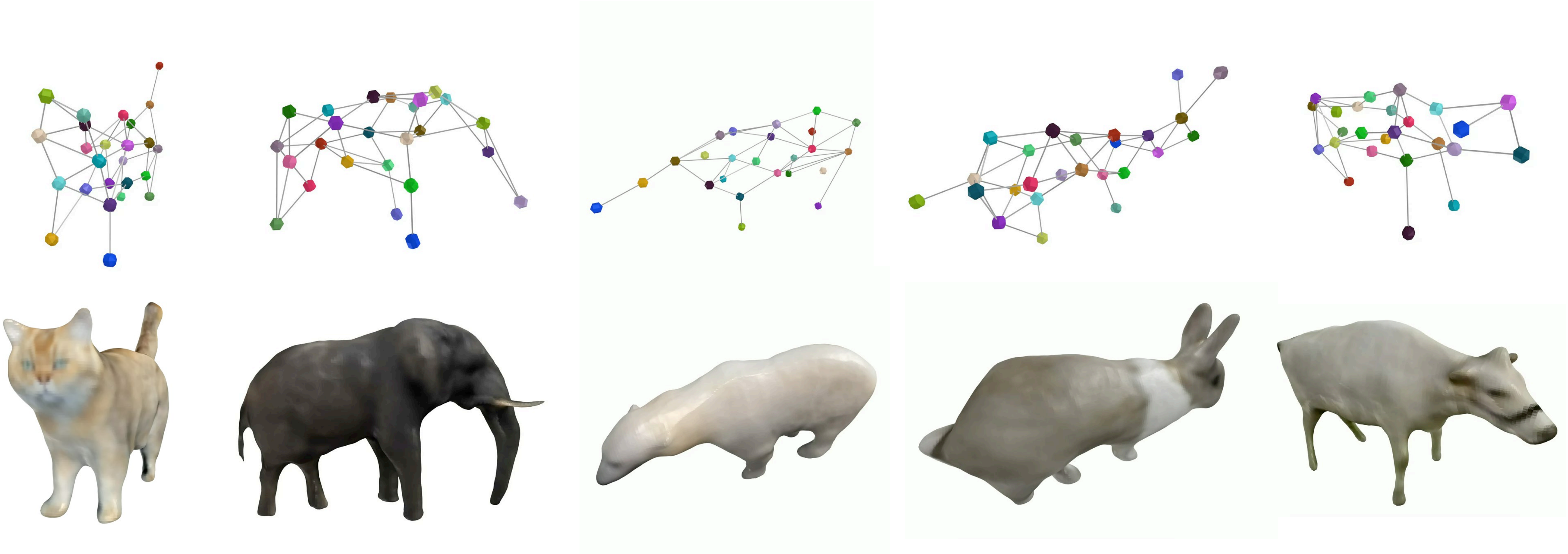


Experimental Results



Experimental Results

Skeleton Generation & 3D Model Manipulation



Summary

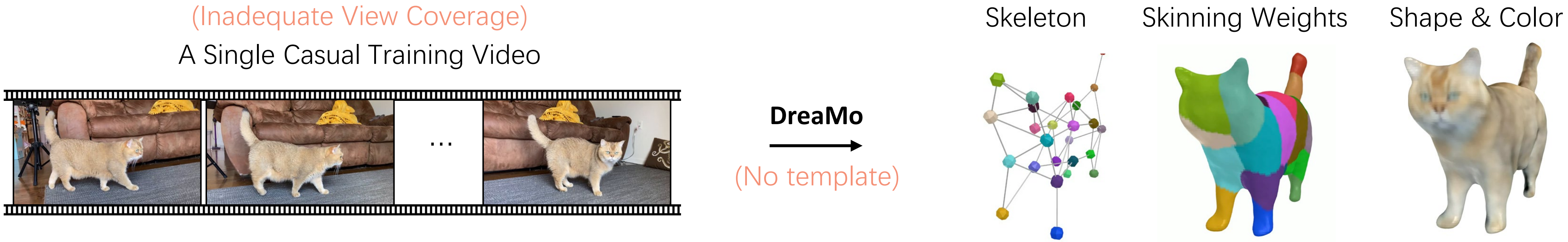
Articulated 3D reconstruction through jointly training-view reconstruction, unseen view hallucination, and tailored regularizations from a single casual video with inadequate view coverage

Advantages

- Require only one easily accessible casual video from the Internet
- Template-free, eliminating the need for 3D scans

Main Technical Contributions

- Diffusion-guided hallucination
- Tailored regularizations to prevent irregular 3D shape
- Simple strategy for interpretable skeleton generation



(Inadequate View Coverage)

A Single Casual Training Video

DreaMo

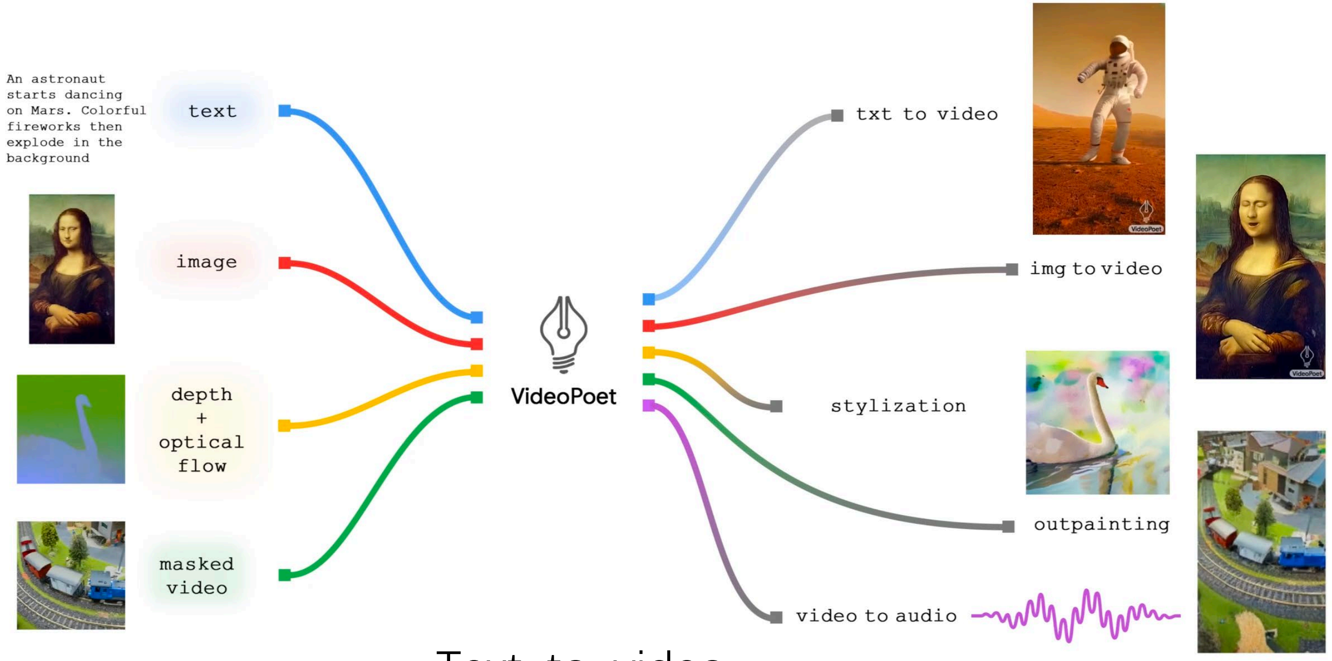
(No template)

Skeleton

Skinning Weights

Shape & Color

Google VideoPoet



An astronaut starts dancing on Mars. Colorful fireworks then explode in the background

text

image

depth + optical flow

masked video

VideoPoet

txt to video

img to video

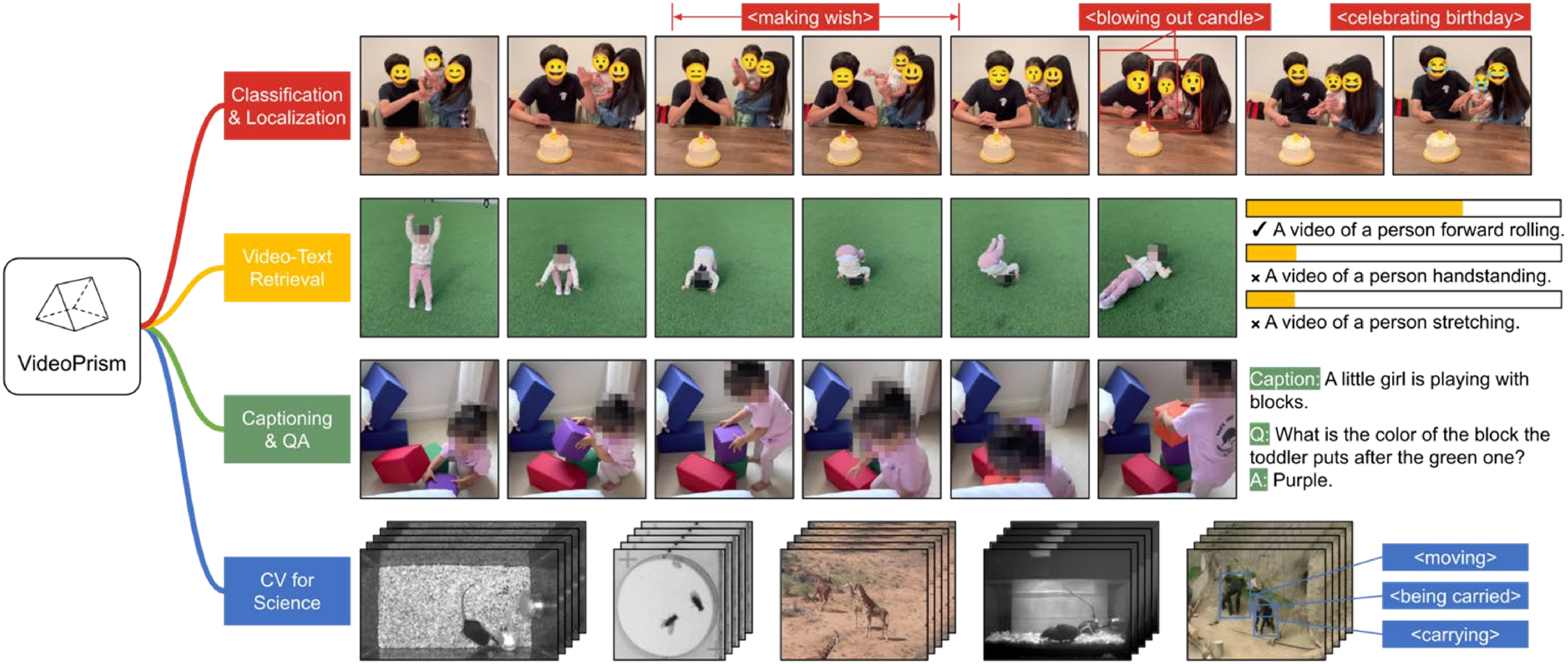
stylization

outpainting

video to audio

Text-to-video
Image-to-video
Video editing
Stylization
Inpainting

Google VideoPrism



<making wish>
<blowing out candle>
<celebrating birthday>
VideoPrism
Classification & Localization
Video-Text Retrieval
Captioning & QA
CV for Science
✓ A video of a person forward rolling.
× A video of a person handstanding.
× A video of a person stretching.
Caption: A little girl is playing with blocks.
Q: What is the color of the block the toddler puts after the green one?
A: Purple.
<moving>
<being carried>
<carrying>