Beyond Patches:
Learning Dense Visual Features

BLISS Speaker Series
June 16, 2026

Thomas Wimmer
PhD at Max Planck Institute for Informatics & ETH Zurich
currently: Student Researcher at Google

Outline

Introduction

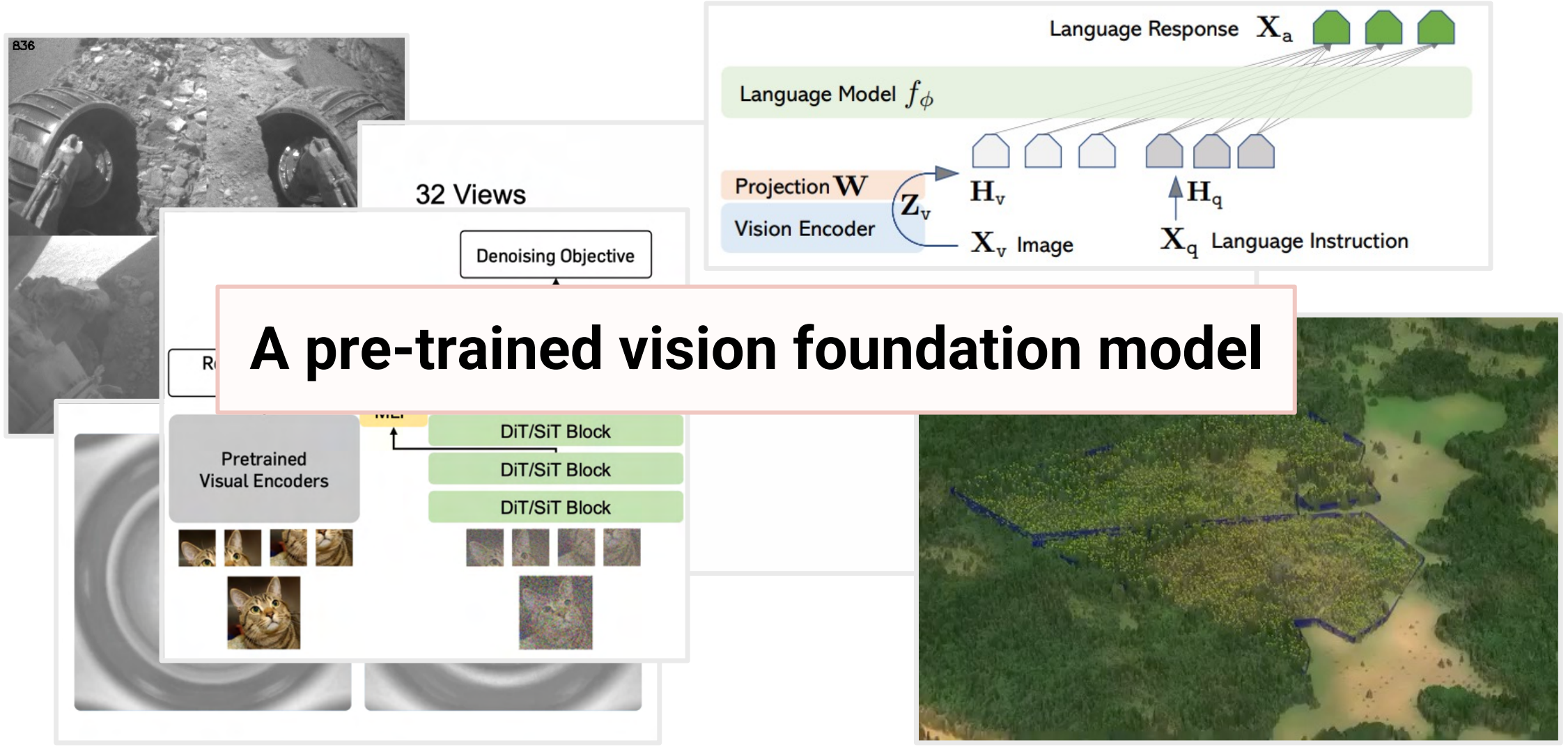
DIY-SC: Improving Dense Features with Pseudo-Labels

AnyUp: Universal Feature Upsampling

Applications of Feature Upsampling

Behind the Scenes

What have all of these in common?



A pre-trained vision foundation model

The Bitter Lesson

Rich Sutton

March 13, 2019

The biggest lesson tha
generalization of cont
of the only ways to in
researchers seek to le
on one is time not spe
advantage of general

In computer chess, th
methods that leverag
were not good losers.
were disappointed w

A similar pattern of r
all those efforts prove
did not play a big rol
important classes of t
needed) and only mu

In speech recognition
tract, etc. On the othe
methods. This led to
step in this consistent
in the games, researc
colossal waste of rese

In computer vision, t
use only the notions

This is a big lesson. A
the bitter lesson that
helps in the short ter
computation by searc

One thing that shoul
methods that seem to

The second general p

re's law, or rather its
an knowledge would be one
fference in the shorter term,
tice they tend to. Time spent
them less suited to taking

researchers who had pursued
edge-based chess researchers
on human input to win and

ecial features of the game, but
in chess, although learning
nd learning are the two most
ding (so that less search was

onemes, of the human vocal
the human-knowledge-based
gnition is the most recent
ech recognition systems. As
unterproductive, and a

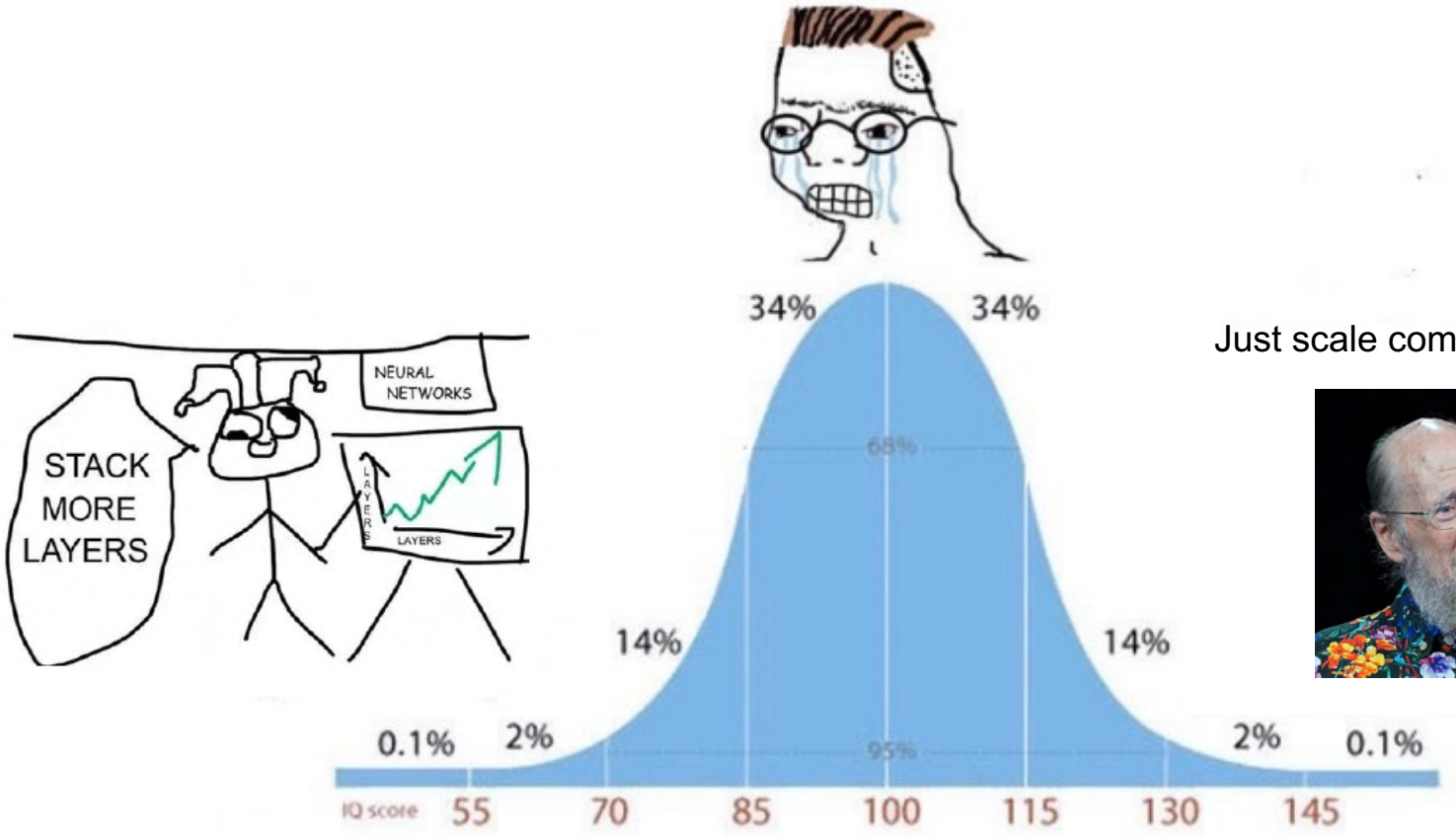
ep-learning neural networks

e mistakes. We have to learn
their agents, 2) this always
roach based on scaling

es very great. The two

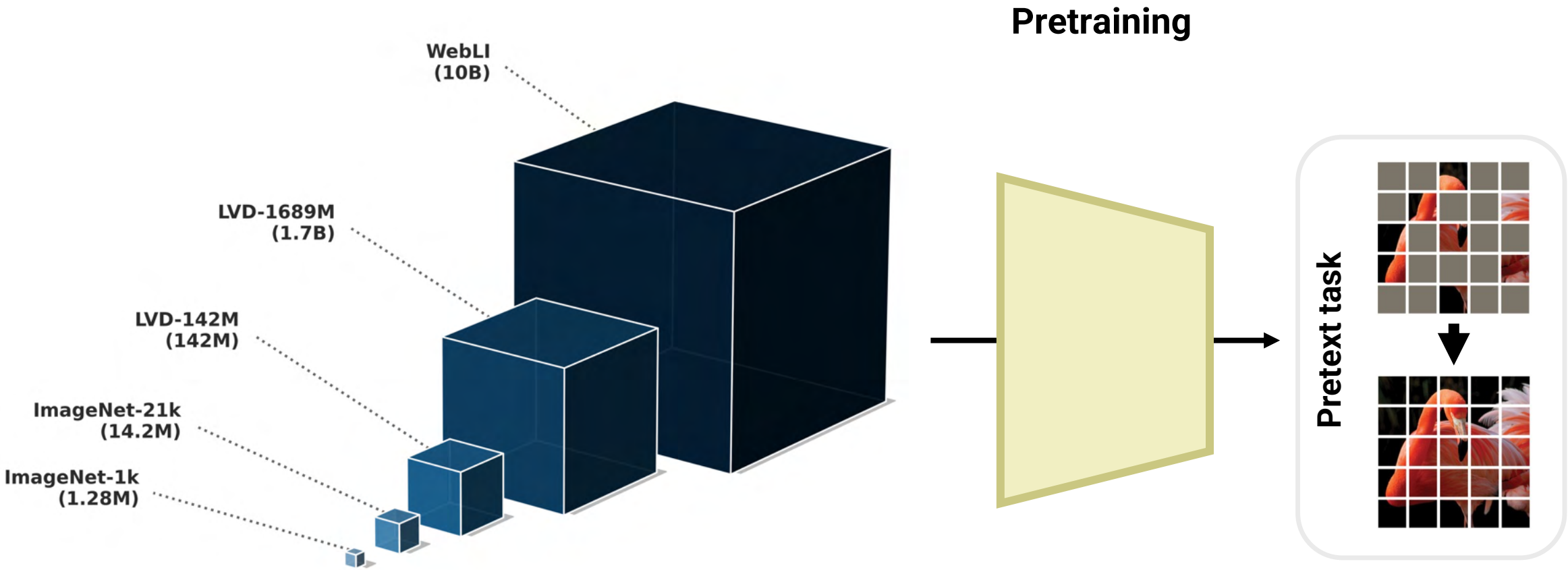
nts of minds, such as simple
ways to think about space, objects, multiple agents, or symmetries. All these are part of the arbitrary, intrinsically-complex, outside world. They are not what should be built in, as their complexity is endless; instead we should build in only the meta-methods that can find and capture this arbitrary complexity. Essential to these methods is that they can find good approximations, but the search for them should be by our methods, not by us. We want AI agents that can discover like we can, not which contain what we have discovered. Building in our discoveries only makes it harder to see how the discovering process can be done.

Noooo… The optimal inductive biases and architectural priors must be carefully engineered



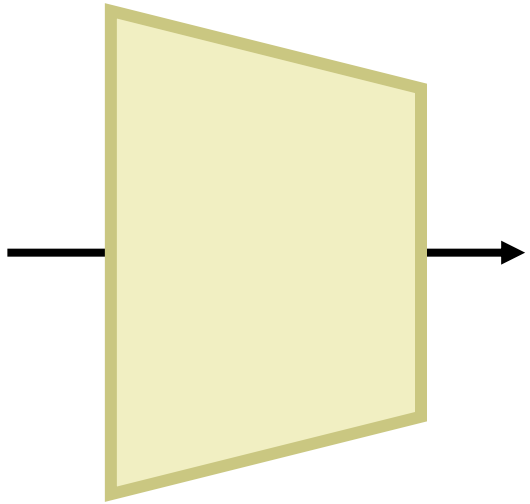
Just scale compute and data.

(Self-supervised) Large-Scale Pretraining



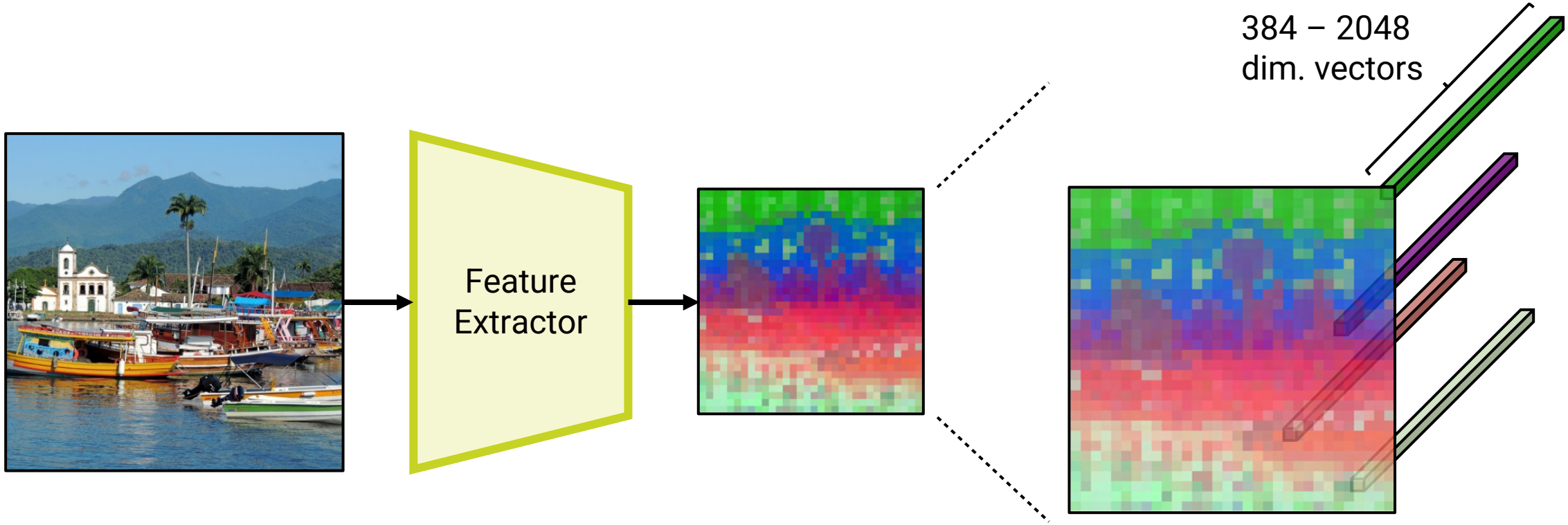
(Self-supervised) Large-Scale Pretraining

Fine-tuning

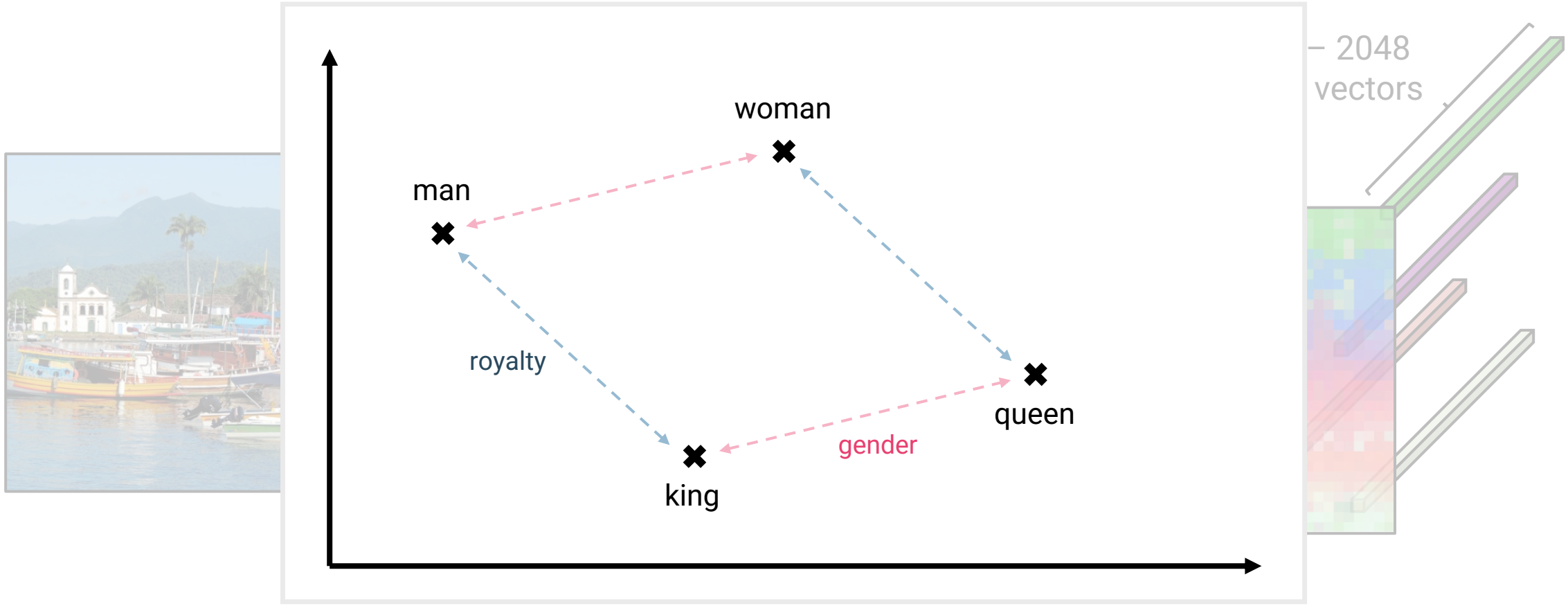


Night / Day?

What makes a good representation?



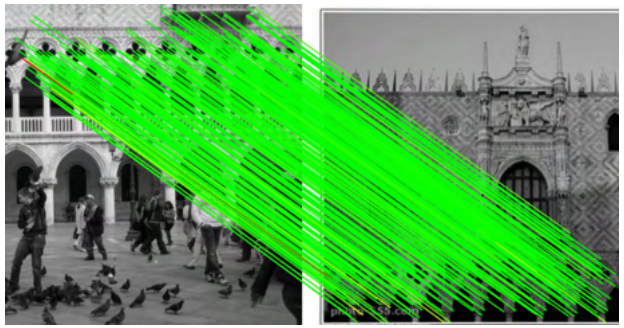
What makes a good representation?



Other important aspects: Versatility, geometric awareness, robustness, …

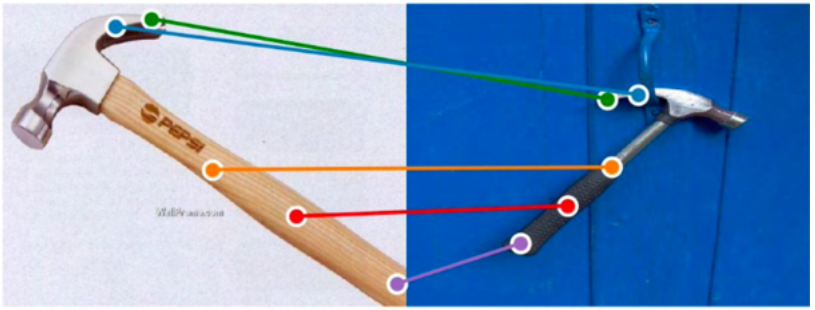
What makes dense features useful?
Can we optimize for it?

"[The three most important problems in computer vision are]
Correspondence, correspondence, correspondence!"
- Takeo Kanade



Geometric

Temporal



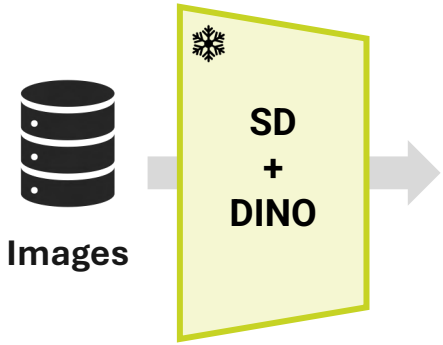
Semantic

Do It Yourself: Learning Semantic Correspondence from Pseudo-Labels

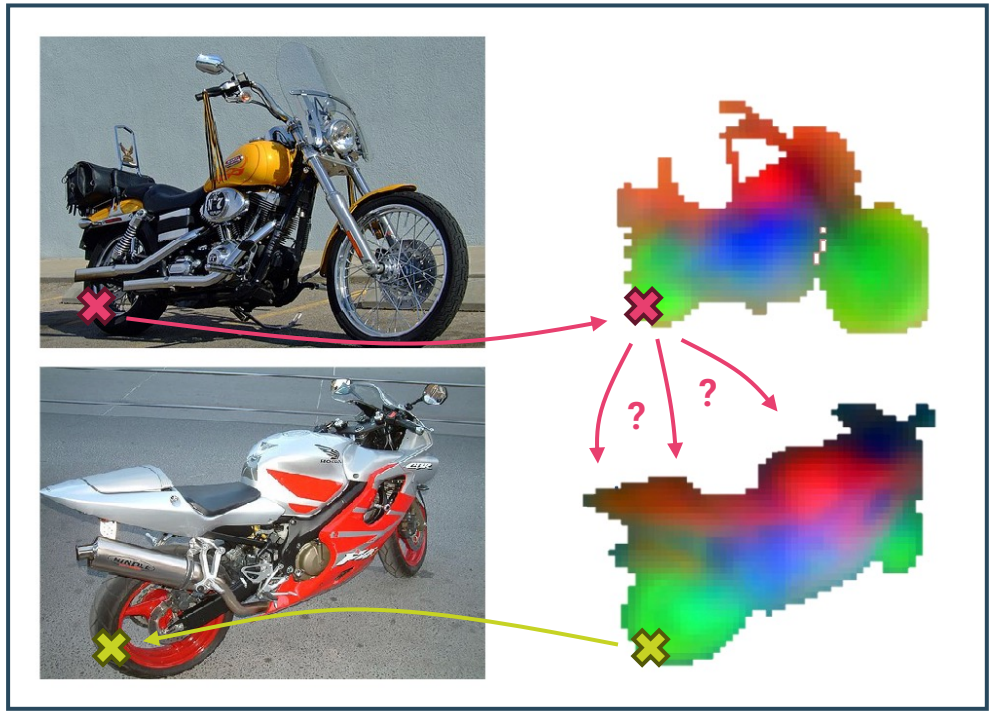
Olaf Dünkel[1], Thomas Wimmer[1,2], Christian Theobalt[1], Christian Rupprecht[3], Adam Kortylewski[1,4]
[1]Max Planck Institute for Informatics, [2]ETH Zurich, [3]University of Oxford, [4]University of Freiburg

ICCV 2025

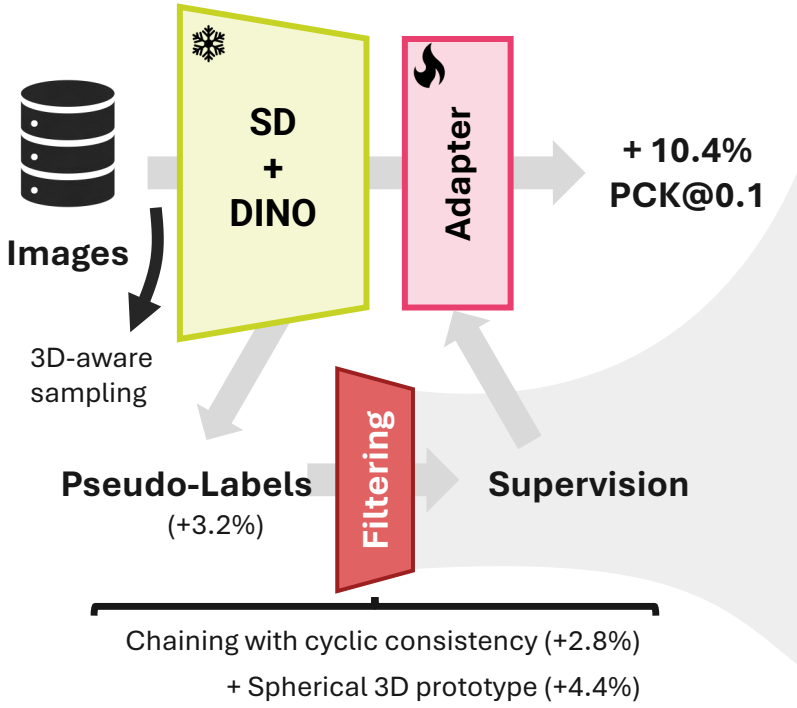
DIY-SC: Learning Semantic Correspondence from Pseudo-Labels

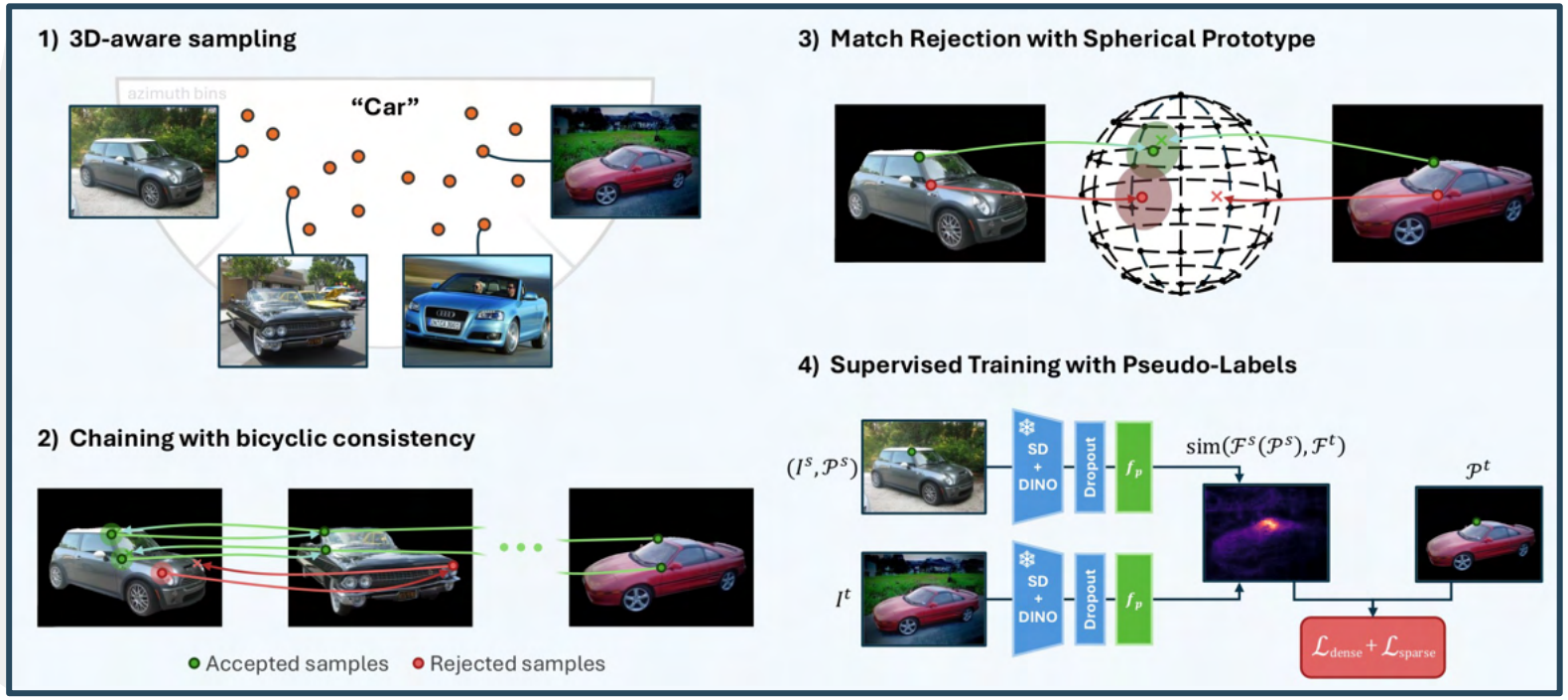


Nearest-neighbor matching in feature space already works well!

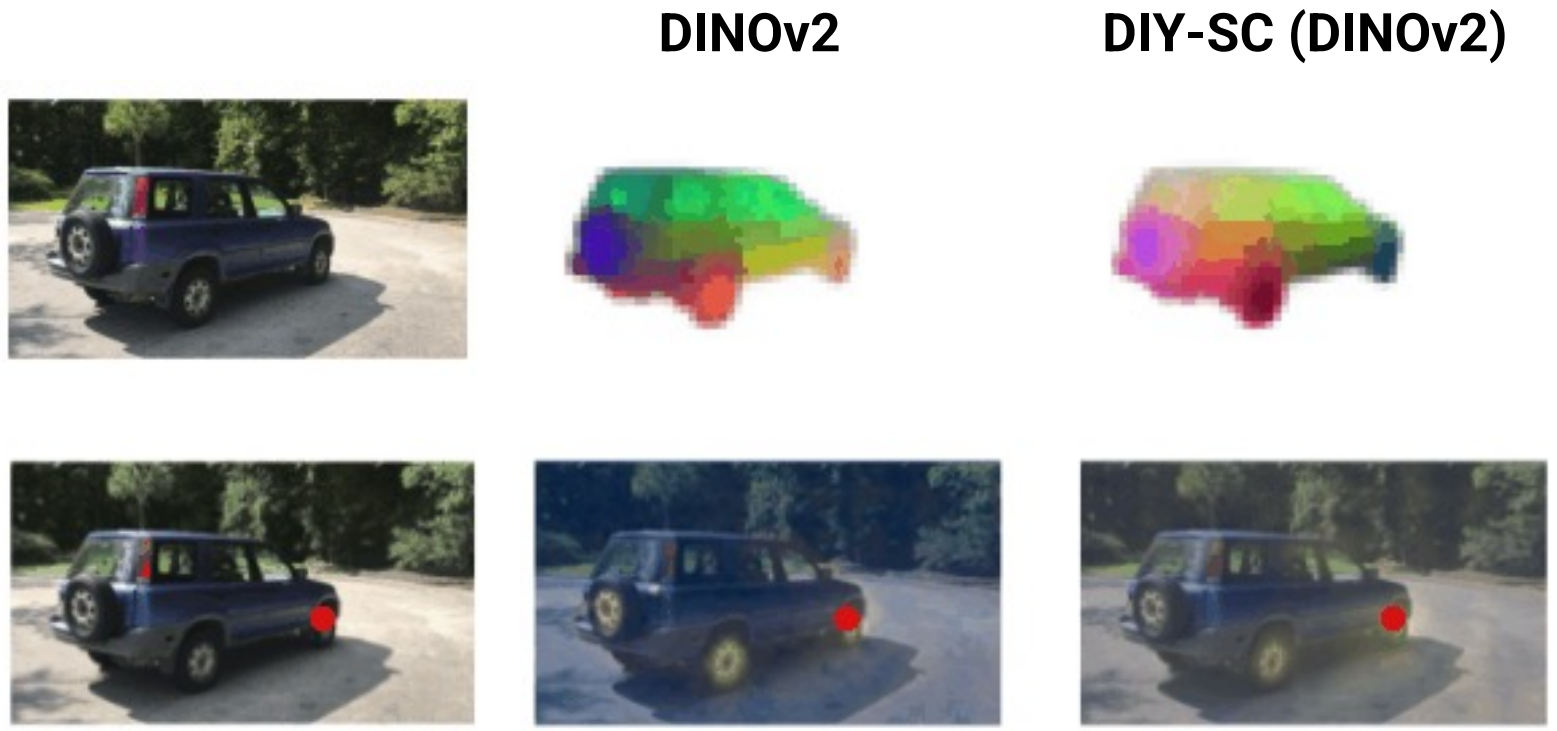


DIY-SC: Learning Semantic Correspondence from Pseudo-Labels





DIY-SC: Learning Semantic Correspondence from Pseudo-Labels

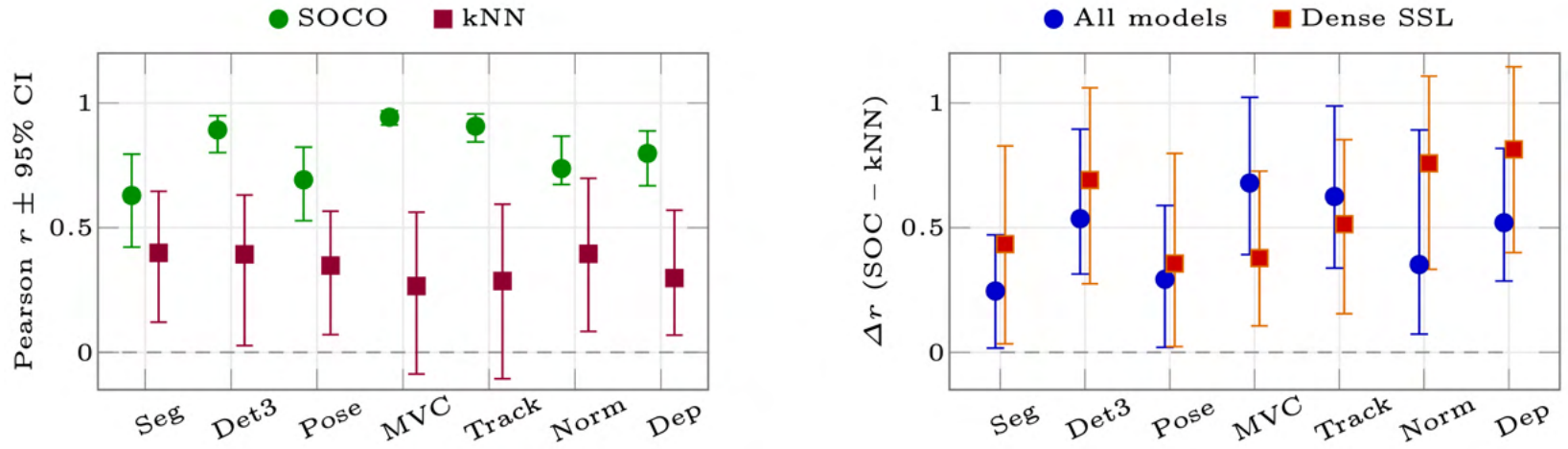


→ Increased robustness of object-centric features

DIY-SC: Learning Semantic Correspondence from Pseudo-Labels

- The optimized features improve performance for all correspondence-based tasks (geometric corr., tracking, …) while preserving quality of backbone features for other tasks.

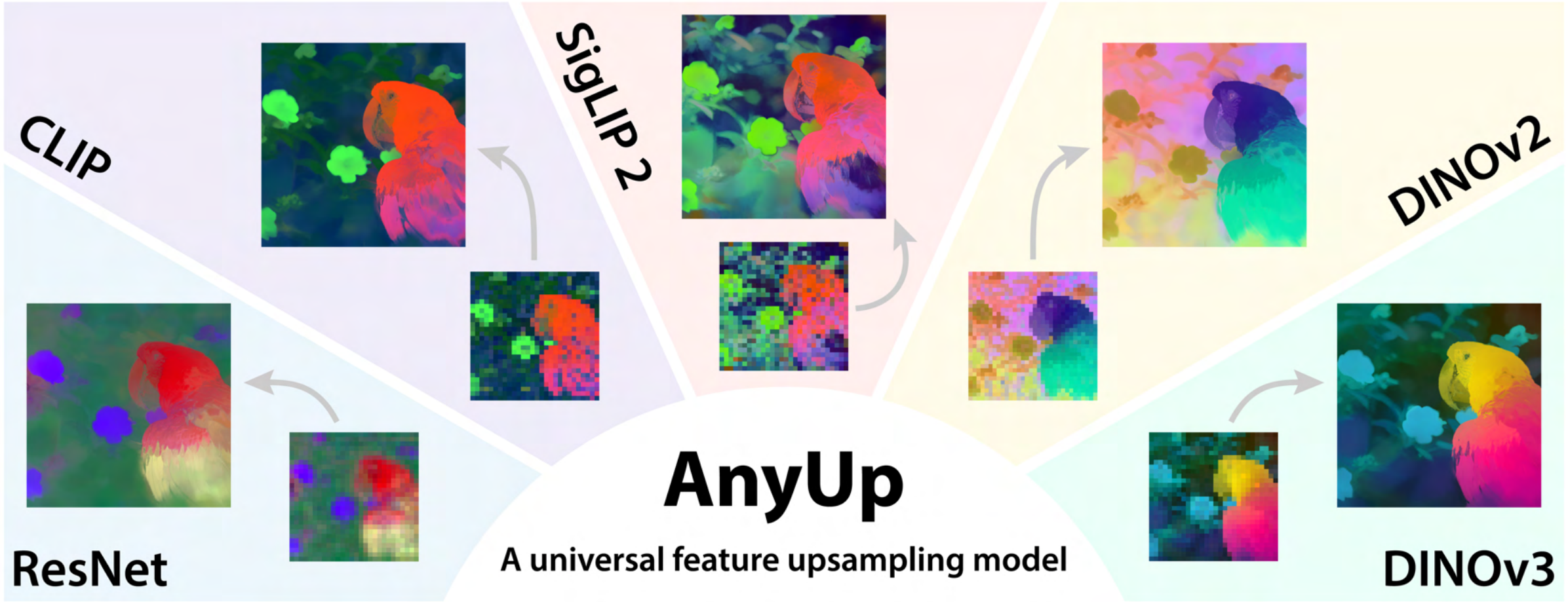
- Easy to probe and gives good signal on performance for other tasks.



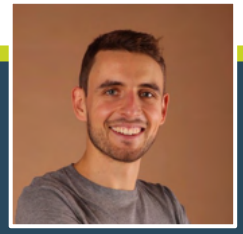
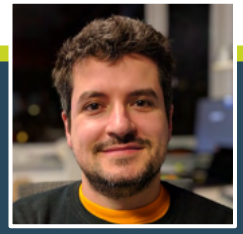
Dünkel et al. “SOCO: Benchmarking Semantic Object Correspondence in Vision Foundation Models”. Preprint 2026

So far, our dense features are still not *really dense* due to the patchification in ViTs.

How can we obtain *actually* dense features?

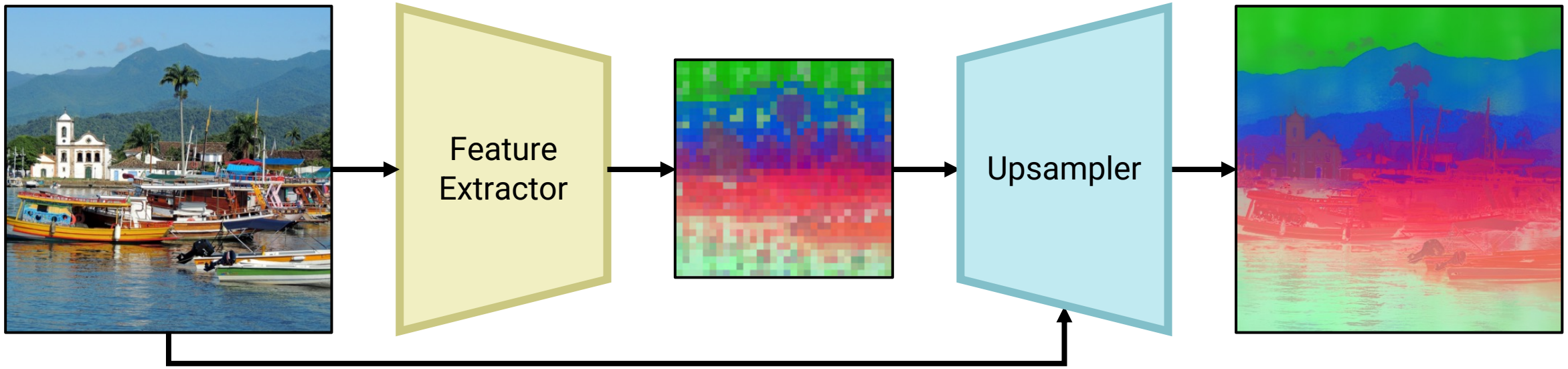




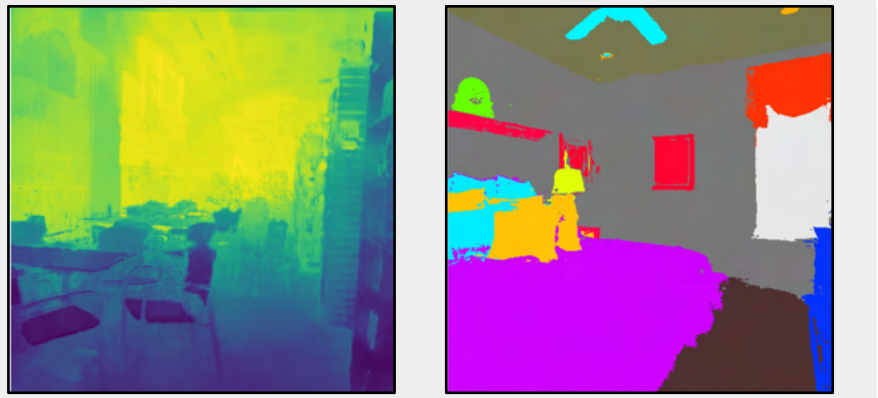
Thomas Wimmer[1,2], Prune Truong[3], Marie-Julie Rakotasoana[3], Michael Oechsle[3], Federico Tombari[3,4], Bernt Schiele[1], Jan Eric Lenssen[1]
[1]Max Planck Institute for Informatics, [2]ETH Zurich, [3]Google, [4]Technical University of Munich

ICLR 2026



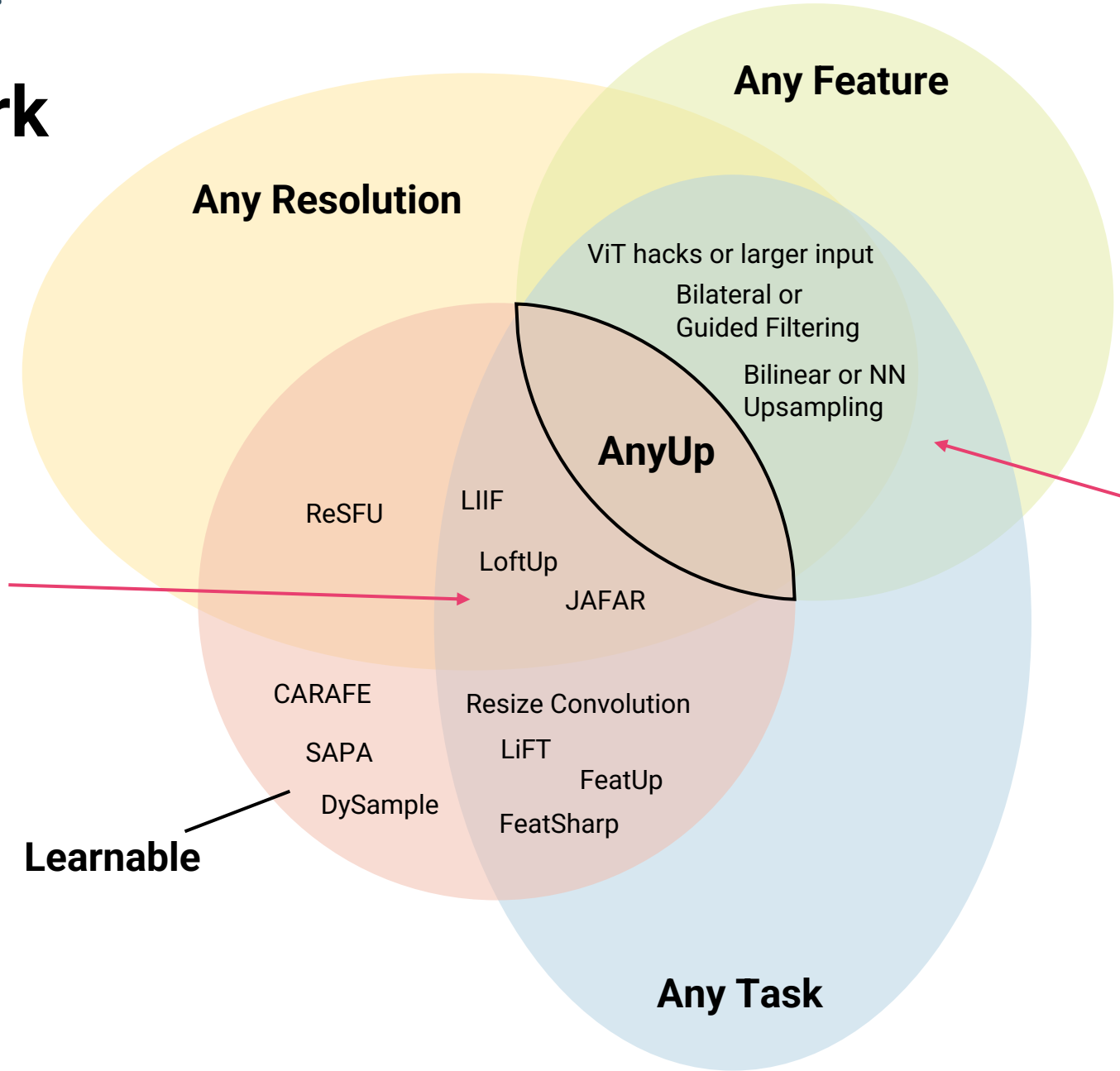
Why do we need dense features?

- Pixel-wise predictions:
  - depth estimation, segmentation, correspondences
- Dense explanations
- Zero-shot predictions of VLMs
- Projecting features into 3D (or anywhere else)
- High-res teacher in model agglomeration



Depth Estimation **Sem. Segmentation**

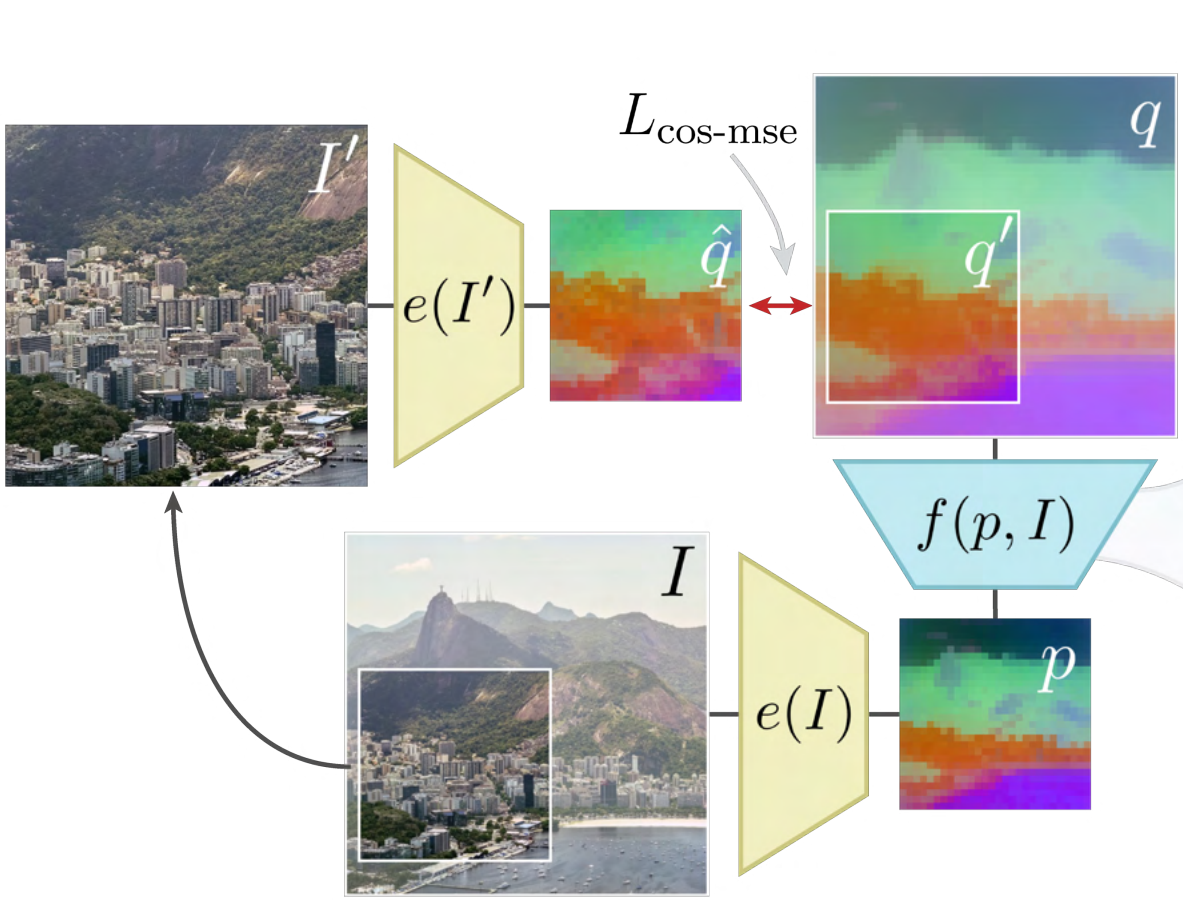
Related Work



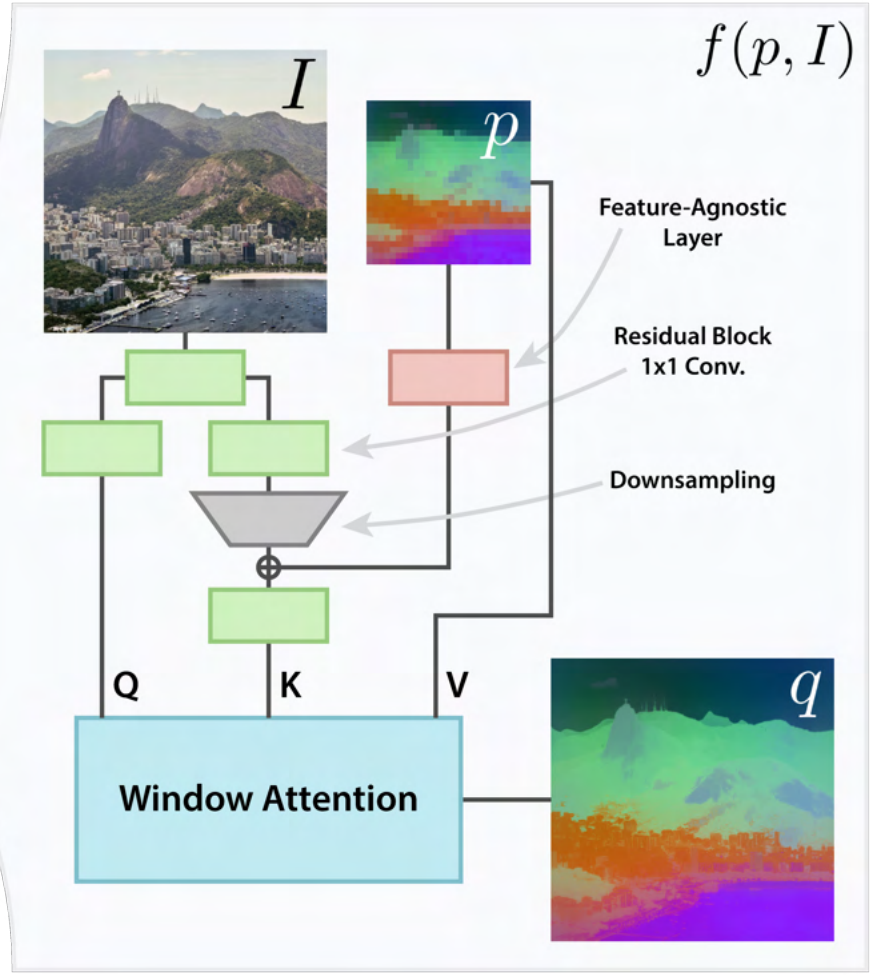
Powerful but need to be trained from scratch for every image encoder

General but lower quality, need to be tuned per sample

Method

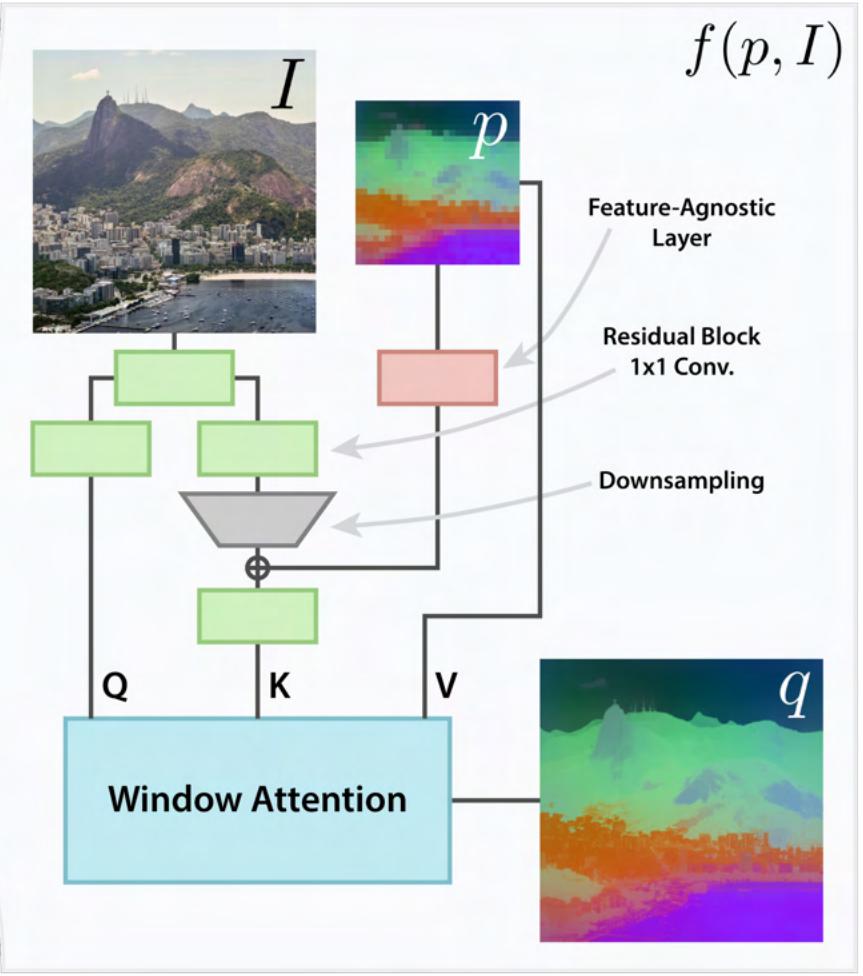


(a) Training Strategy



(b) Architecture

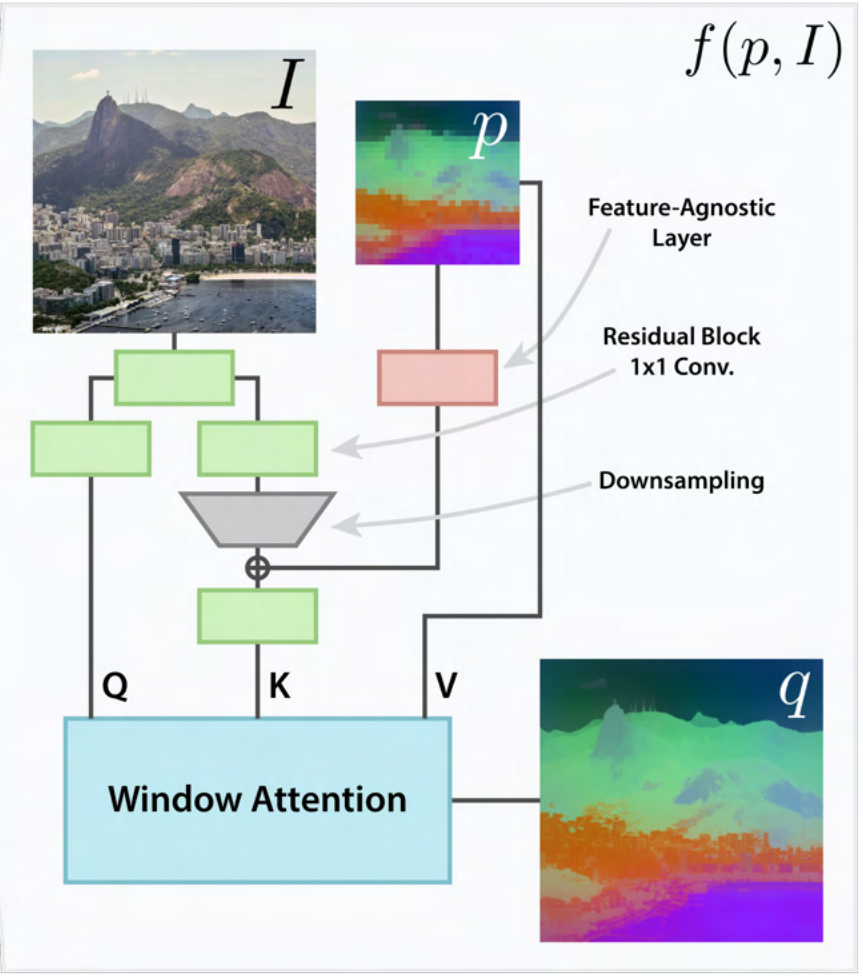
Method



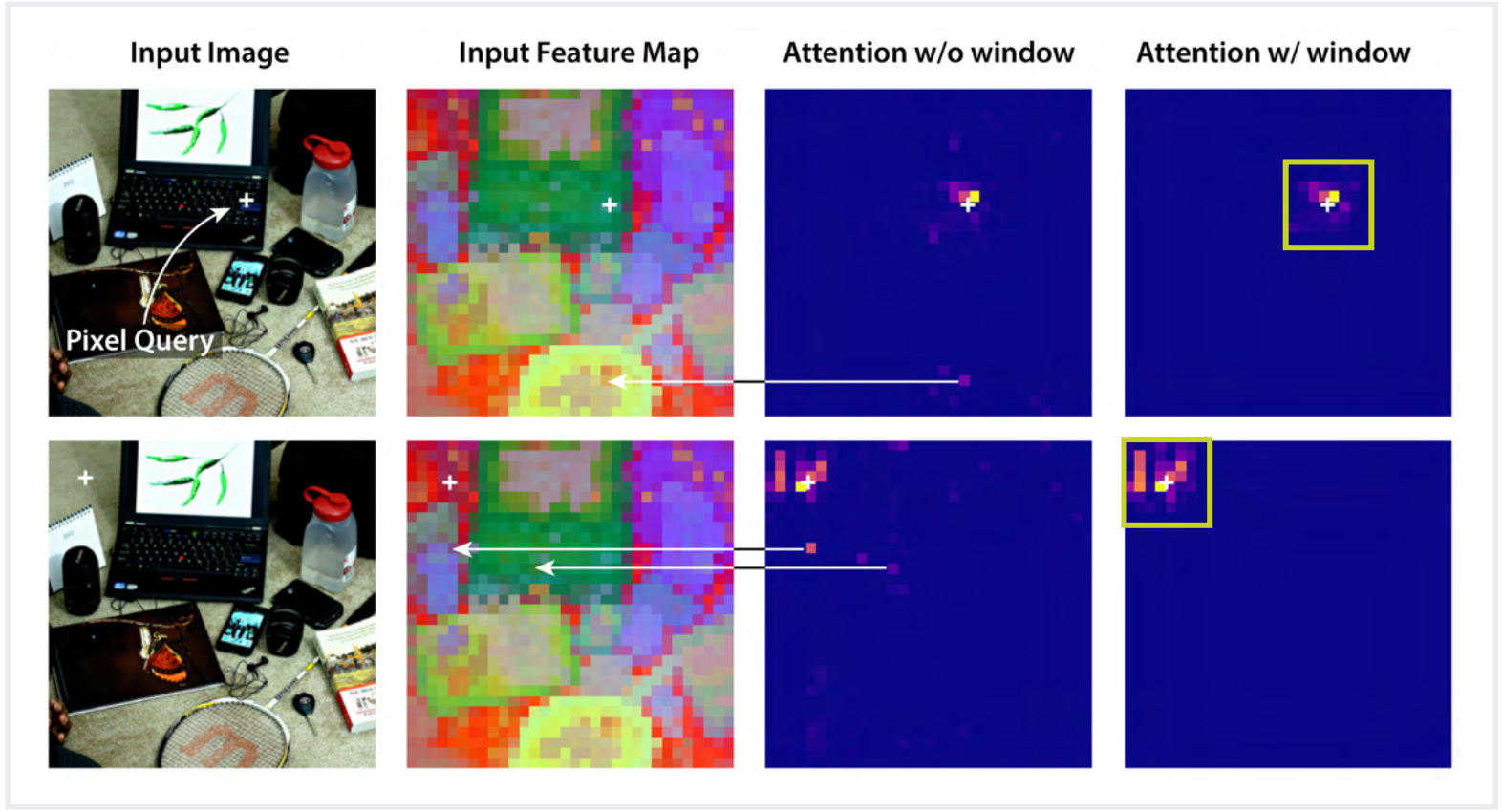
(b) Architecture

(c) Feature-Agnostic Layer

Method

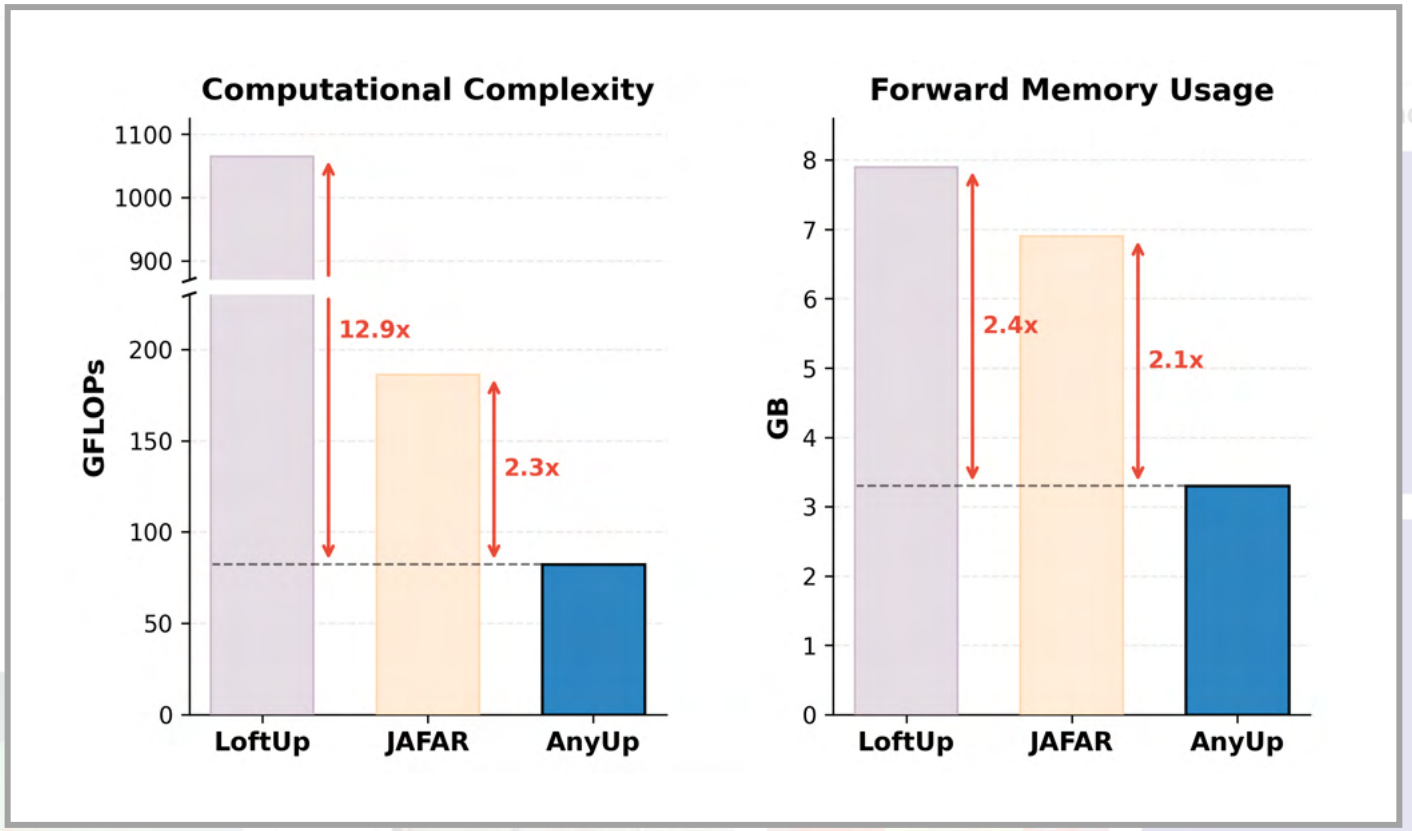


(b) Architecture



(d) Window Attention

Method

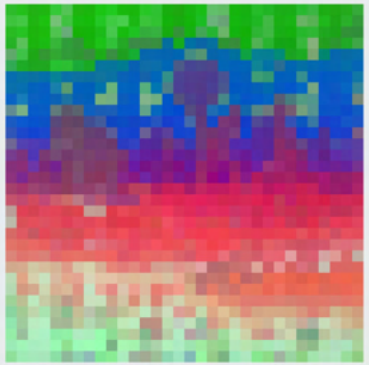


Local Window Attention **prevents relying on far-away, unrelated objects** during upsampling, while making upsampling more **computationally efficient**

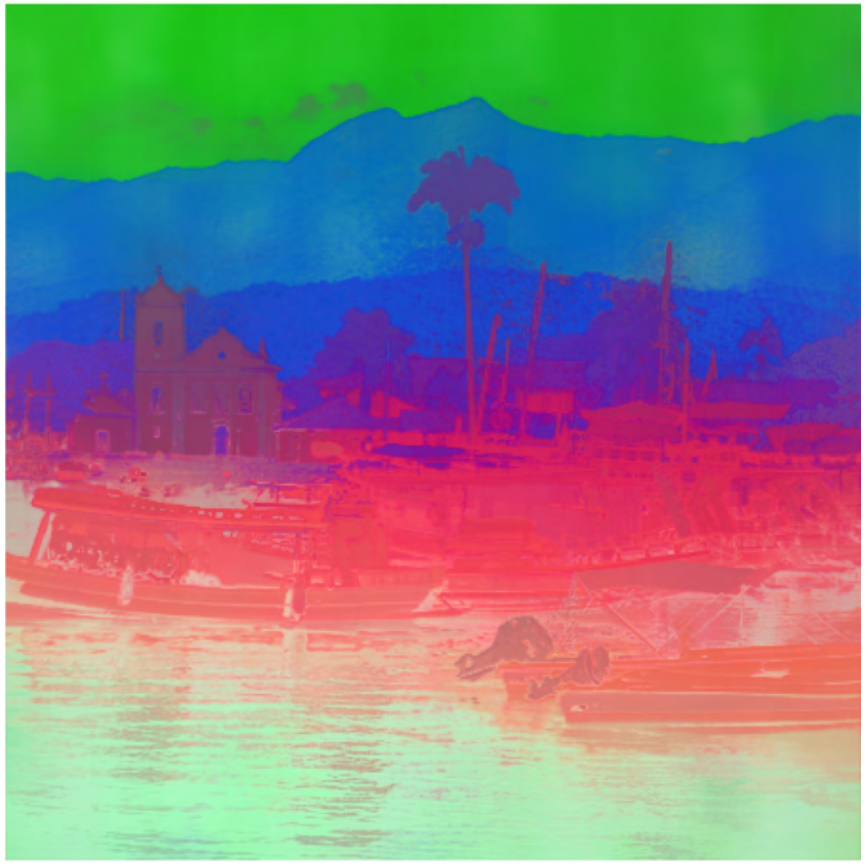
Qualitative Results

RGB Image

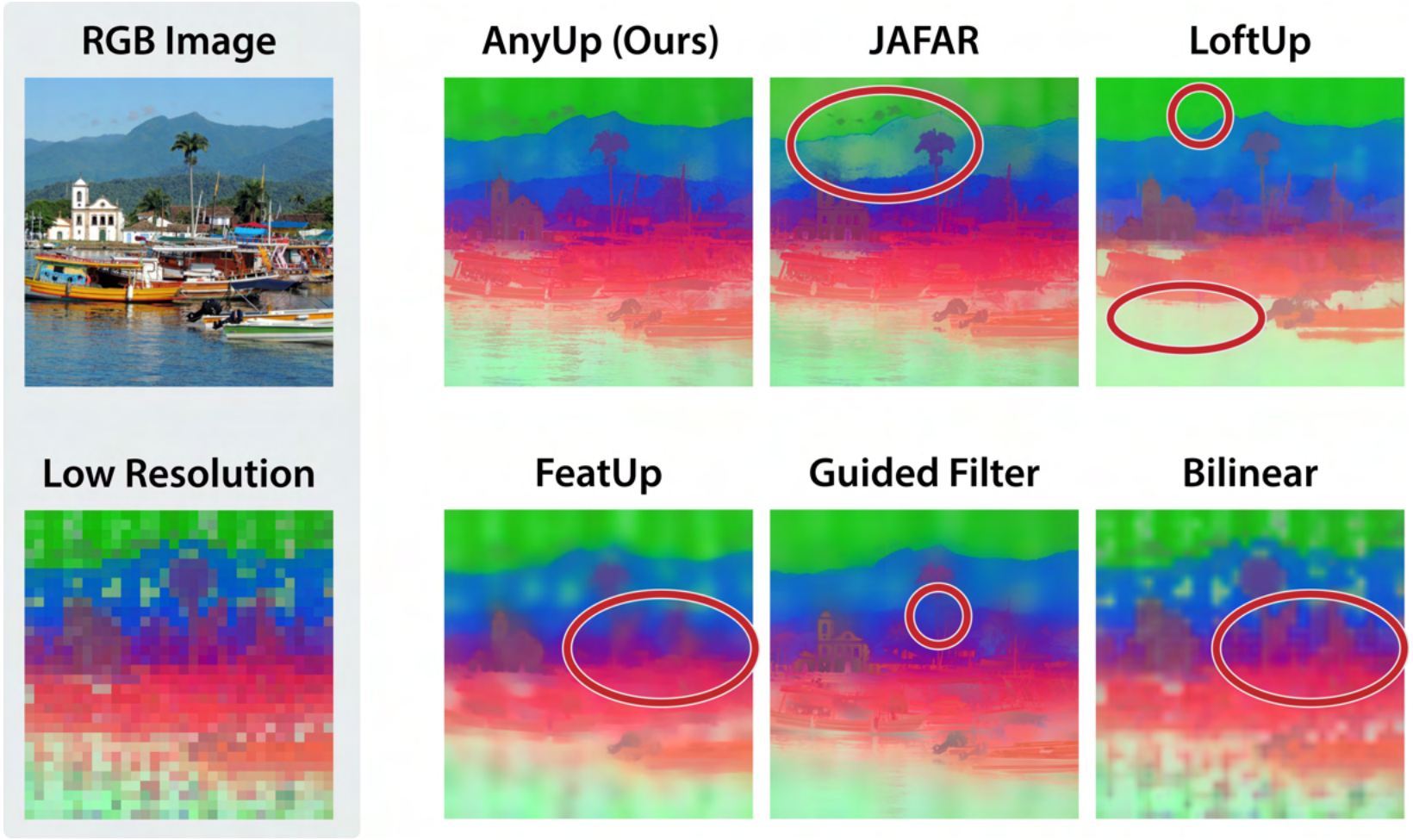
Low Resolution



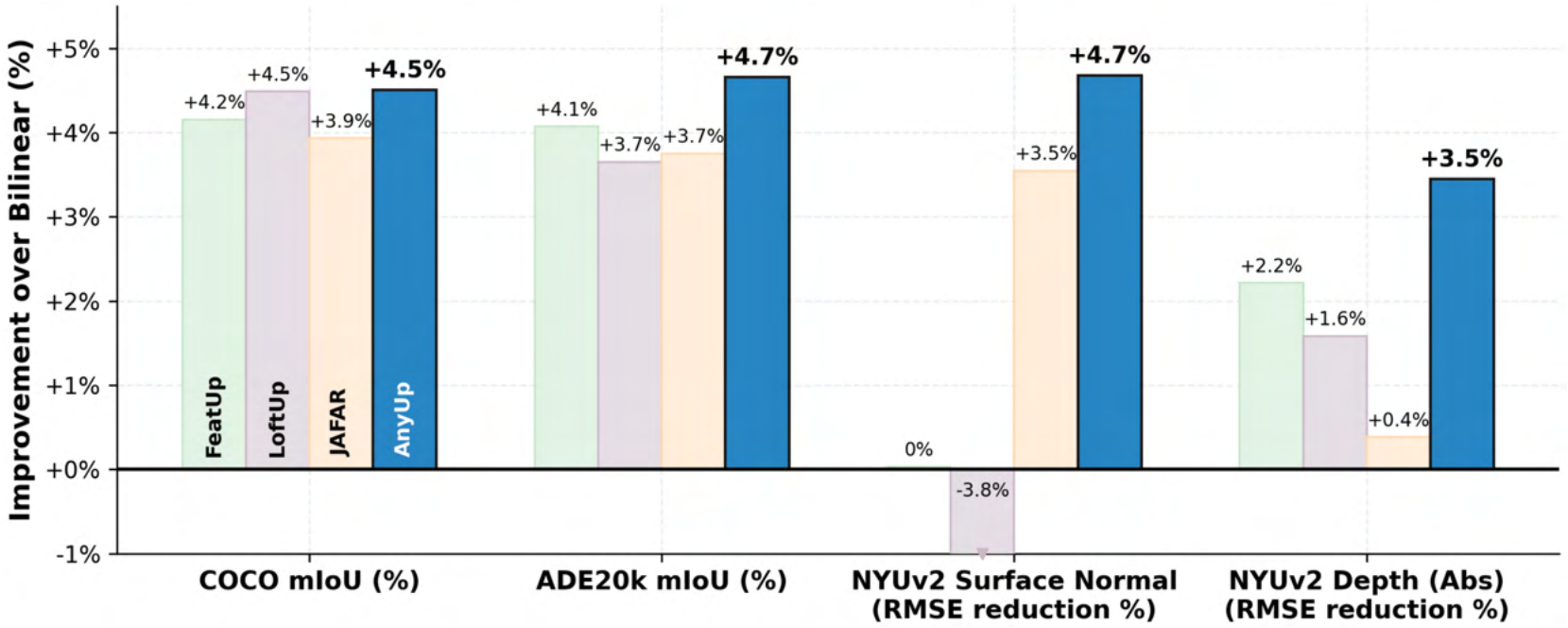
AnyUp (Ours)



Qualitative Results

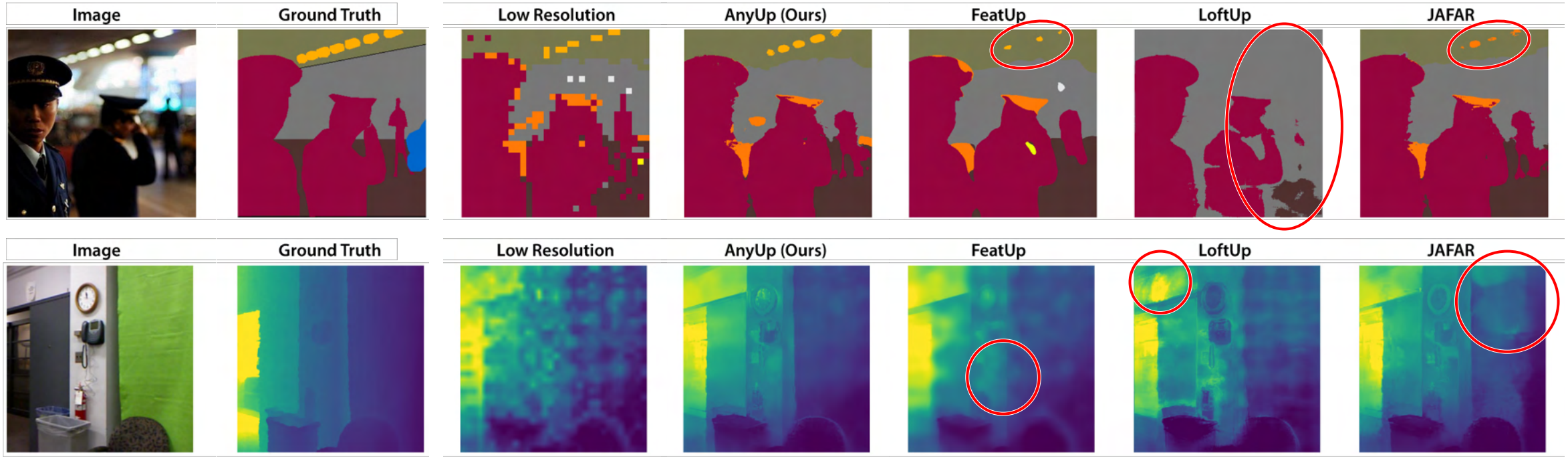


Quantitative Results



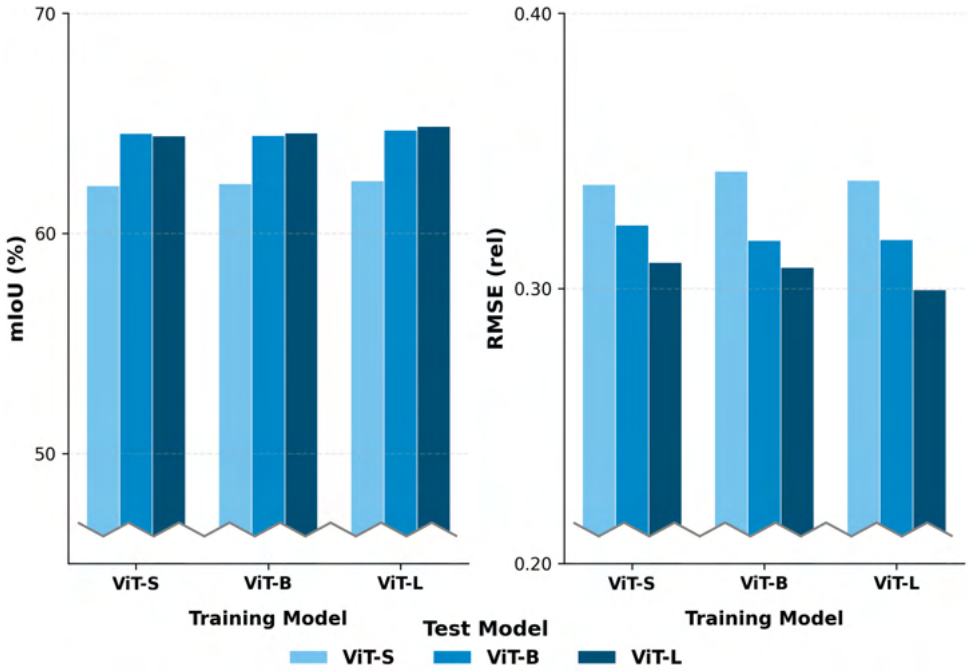
AnyUp achieves **SOTA performance** on downstream tasks in comparison against *encoder-specific* models.

Feature Space Preservation



AnyUp **retains the input feature distribution** while improving upsampling quality.
LoftUp introduces extra transformations that distort the distribution and degrade results.

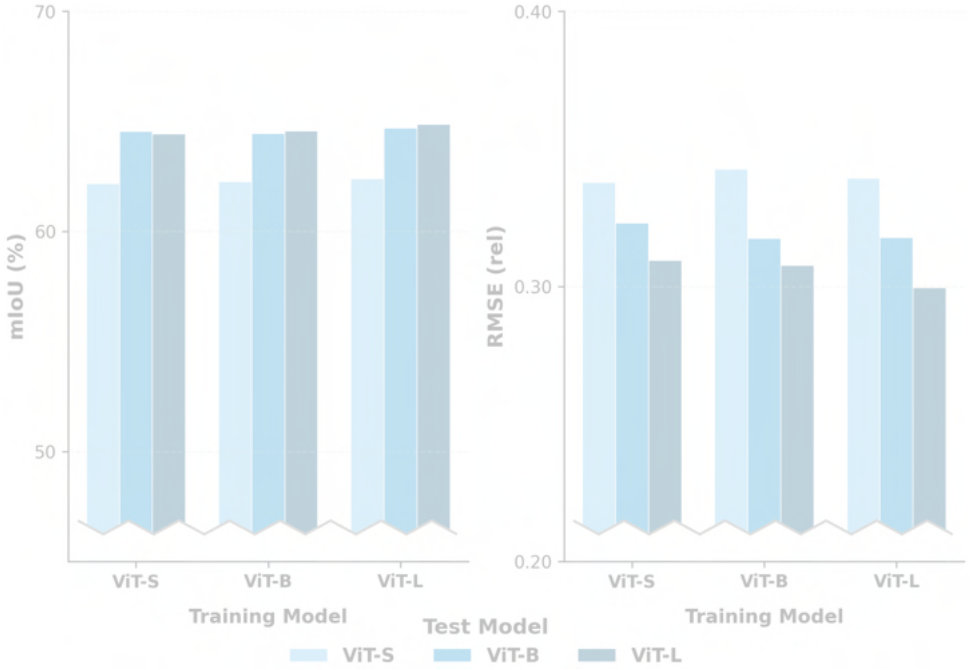
Generalization

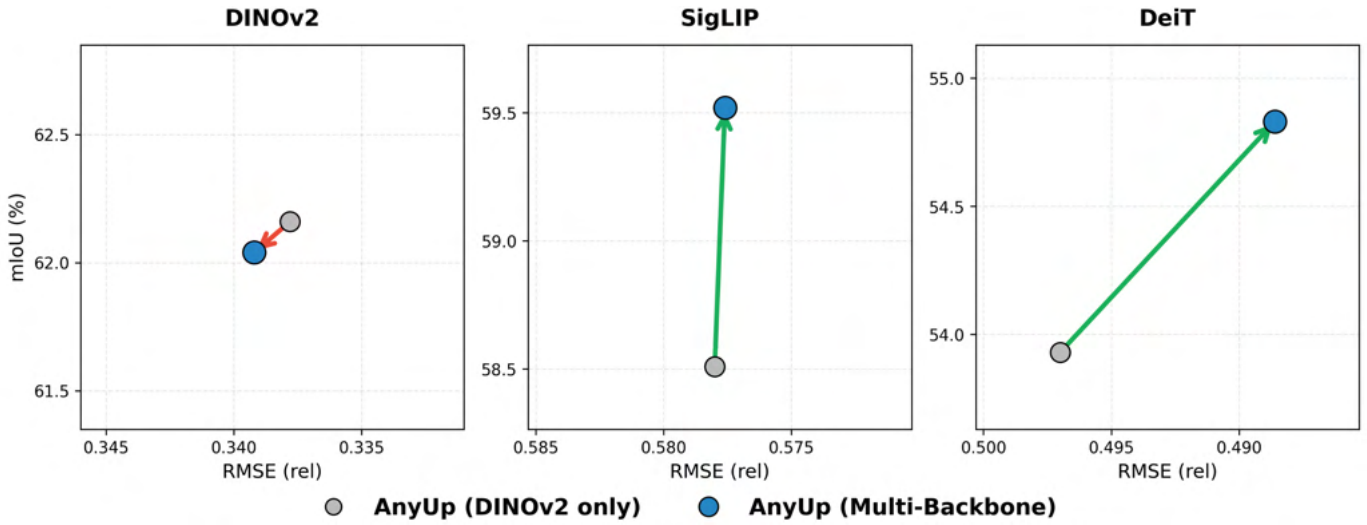


	Train Model	Test Model	mIoU (Acc)	RMSE (abs / rel)
AnyUp	DINOv2 (ViT-S)	SigLIP 2^{LoftUp}	51.68 (73.35)	0.59 / 0.48
AnyUp	SigLIP 2^{LoftUp}	SigLIP 2^{LoftUp}	**54.45 (75.49)**	**0.57 / 0.46**
LoftUp	SigLIP 2^{LoftUp}	SigLIP 2^{LoftUp}	40.73 (64.87)	0.72 / 0.60
AnyUp	DINOv2 (ViT-S)	SigLIP 2^{JAFAR}	58.51 (78.36)	0.91 / 0.58
AnyUp	SigLIP 2^{JAFAR}	SigLIP 2^{JAFAR}	**60.32 (79.57)**	**0.90 / 0.57**
JAFAR	SigLIP 2^{JAFAR}	SigLIP 2^{JAFAR}	60.10 (79.40)	0.93 / 0.60
AnyUp	DINOv2 (ViT-S)	DINOv3 (ViT-S$^+$)	62.96 (81.82)	0.51 / 0.37
AnyUp	DINOv3 (ViT-S$^+$)	DINOv3 (ViT-S$^+$)	62.99 (81.84)	0.51 / 0.37

AnyUp **generalizes well to unseen feature extractors** at test time.
Training on multiple feature extractors further improves generalization.

Generalization





AnyUp **generalizes well to unseen feature extractors** at test time.
Training on multiple feature extractors further improves generalization.

Generalization

	Semantic Segmentation						Depth Estimation					
	16 → 112		32 → 224		32 → 112		16 → 112		32 → 224		32 → 112	
	mIoU (↑)	Acc. (↑)	mIoU (↑)	Acc. (↑)	mIoU (↑)	Acc. (↑)	RMSE (abs) (↓)	RMSE (rel) (↓)	RMSE (abs) (↓)	RMSE (rel) (↓)	RMSE (abs) (↓)	RMSE (rel) (↓)
Bilinear	56.38	77.17	59.42	79.28	59.40	79.27	0.4927	0.3586	0.4606	0.3274	0.4600	0.3273
FeatUp	58.88	79.15	61.92	81.10	61.76	80.99	**0.4357**	**0.3231**	0.4507	0.3145	0.4513	0.3160
LoftUp	58.97	79.37	61.68	81.06	61.20	80.69	0.4896	0.3533	0.4591	0.3264	0.4636	0.3296
JAFAR	**59.79**	**79.87**	61.91	81.14	61.66	80.94	0.4871	0.3458	0.4825	0.3489	0.4812	0.3498
AnyUp	59.63	79.75	**62.25**	**81.41**	**62.07**	**81.26**	0.4746	0.3364	**0.4441**	**0.3079**	**0.4455**	**0.3073**

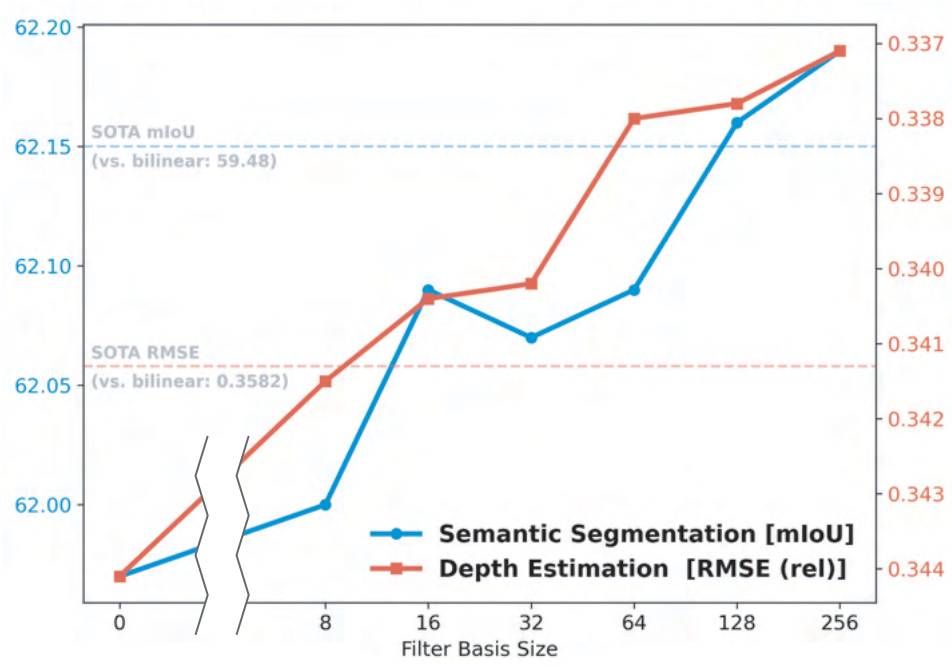
AnyUp **generalizes well to unseen upsampling factors** at test time.

Ablations

(a) Effects of removing specific model or training components

	mIoU (Acc.) (↑)	RMSE (abs / rel) (↓)
AnyUp	**62.16 (81.37)**	**0.4755** / 0.3378
w/o window attn. (4.2)	62.12 (81.34)	0.4854 / 0.3449
w/o our data sampling (4.3.1)	62.03 (81.28)	0.4773 / 0.3387
w/o $L_{\text{self-consistency}}$ (4.3.2)	62.09 (81.33)	0.4763 / **0.3363**
w/o any regularization (4.3.2)	61.90 (81.23)	0.4786 / 0.3401

(b) Impact of filter basis size in feature-agnostic layer



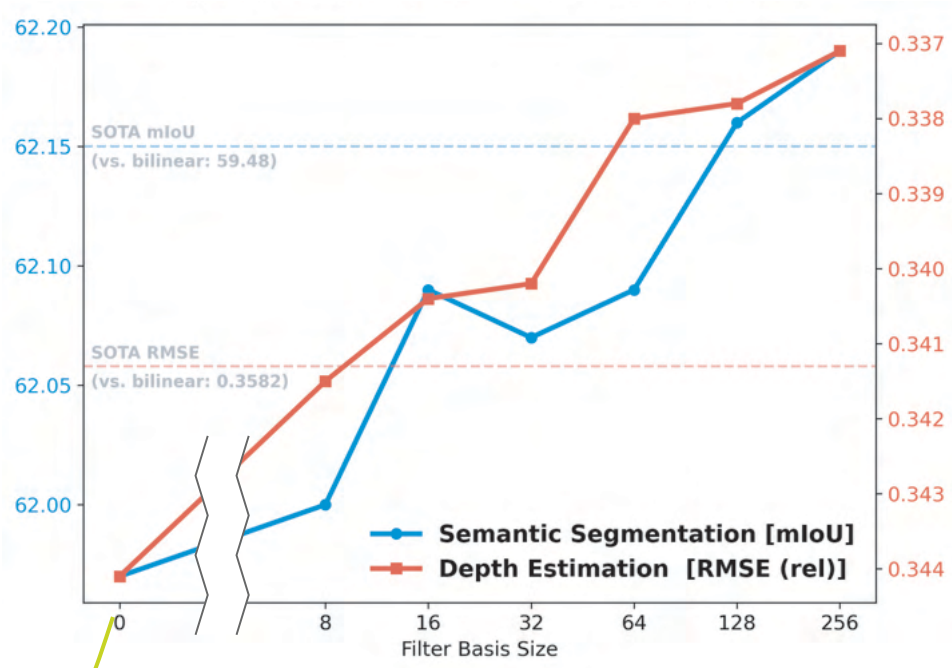
All proposed components lead to notable impact on downstream performance.
Upsampling **performance increases with larger filter basis** sizes in the feature-agnostic layer.

Ablations

(a) Effects of removing specific model or training components

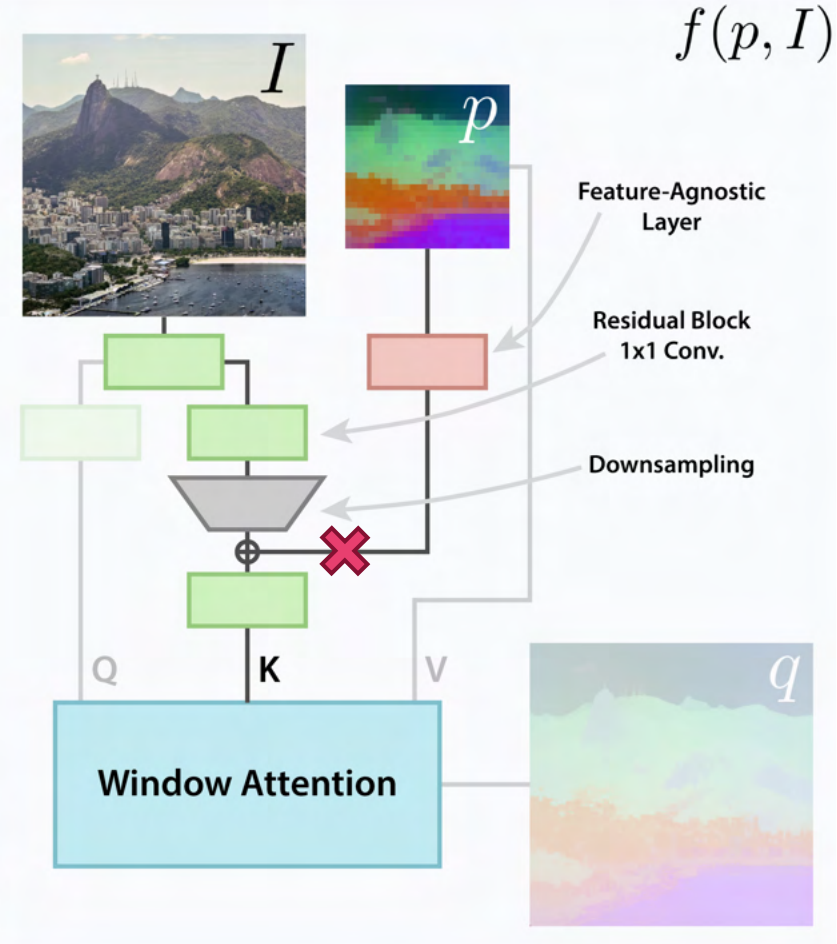
	mIoU (Acc.) (↑)	RMSE (abs / rel) (↓)
AnyUp	**62.16 (81.37)**	**0.4755** / 0.3378
w/o window attn. (4.2)	62.12 (81.34)	0.4854 / 0.3449
w/o our data sampling (4.3.1)	62.03 (81.28)	0.4773 / 0.3387
w/o $L_{\text{self-consistency}}$ (4.3.2)	62.09 (81.33)	0.4763 / **0.3363**
w/o any regularization (4.3.2)	61.90 (81.23)	0.4786 / 0.3401
w/o feature path for key computation	61.97 (81.23)	0.4791 / 0.3441

(b) Impact of filter basis size in feature-agnostic layer

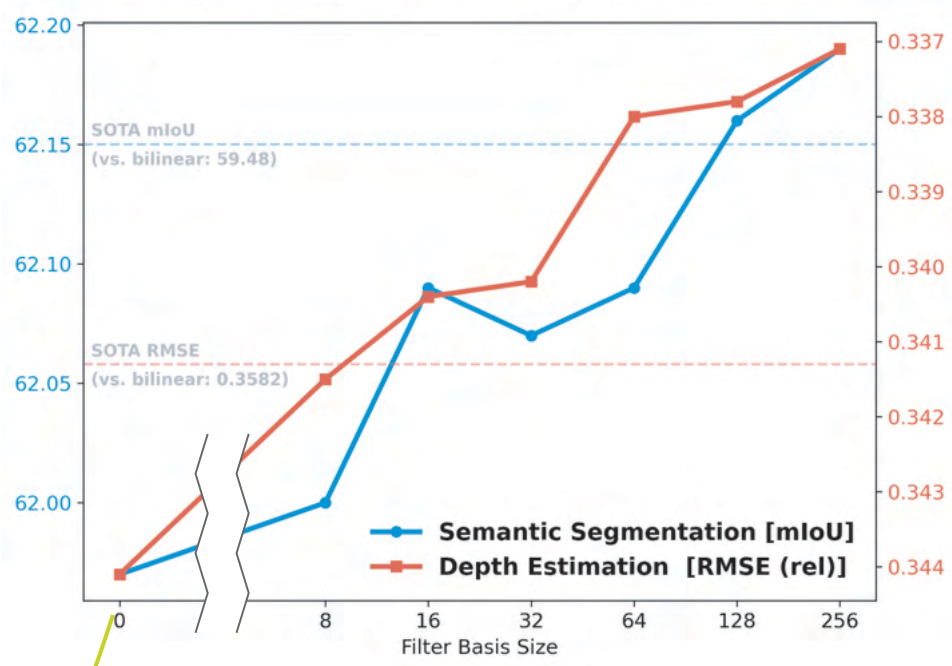


Upsampling based on position and color matching is a strong baseline!

Ablations



(b) Impact of filter basis size in feature-agnostic layer



Upsampling based on position and color matching is a strong baseline!

Limitations

Upsampled features are a *linear combination* of low-resolution input features.

→ A larger and more complex upsampling model could likely extract sub-patch-level information.

Figure 3 | **Qualitative reconstruction comparison.** ♠ denotes trained only on ImageNet. RAEv2 despite only being trained on Imagenet performs competitively with proprietary VAEs. Training on more data (e.g., text) can further help reconstruction [51] (see Fig. 10). Results use DINOv3-L (K=23) for RAEv2.

Singh, Jaskirat, et al. "Improved Baselines with Representation Autoencoders." arXiv preprint arXiv:2605.18324 (2026).

Limitations

Feature Upsampling in its current form is a *post-hoc operation*.

→ Future work could focus on adaptively “allocating compute” on more important image parts during the feature extraction through e.g. adaptive patch sizes.

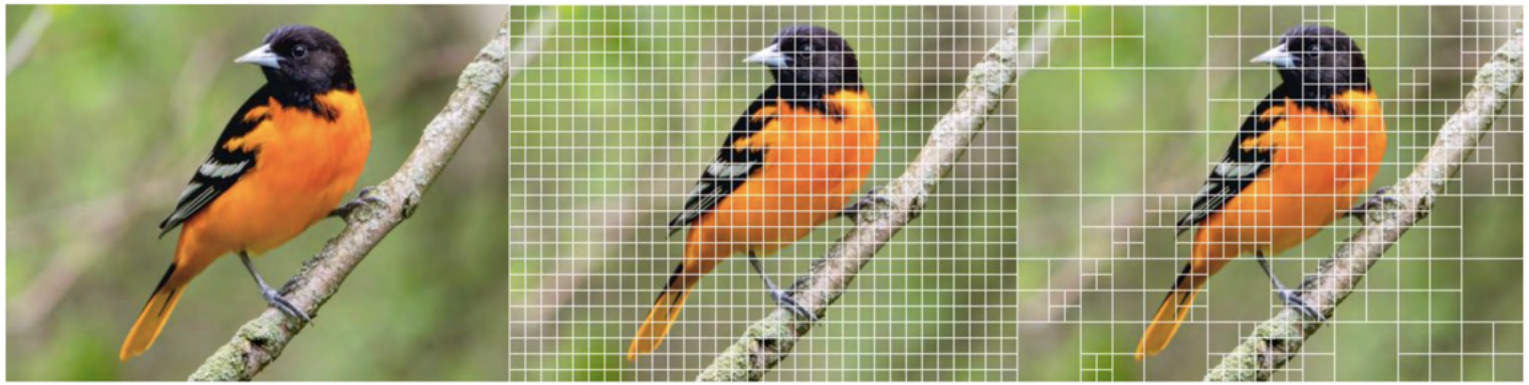


Figure 1: **Adaptive Patch Sizing.** We present APT, Adaptive Patch Transformers, which significantly accelerate vision transformer training and inference by patchifying images based on their content. Complex regions receive more, smaller tokens, while simpler, homogeneous regions receive fewer.

Choudhury, Rohan, et al. "Accelerating Vision Transformers with Adaptive Patch Sizes." arXiv preprint arXiv:2510.18091 (2025).

```python
# load the AnyUp upsampler model
upsampler   = torch.hub.load('wimmerth/anyup', 'anyup')

# upsampled high-resolution features (B, C, H, W)
hr_features = upsampler(hr_image, lr_features)
```

Website: https://wimmerth.github.io/anyup/
Code: https://github.com/wimmerth/anyup

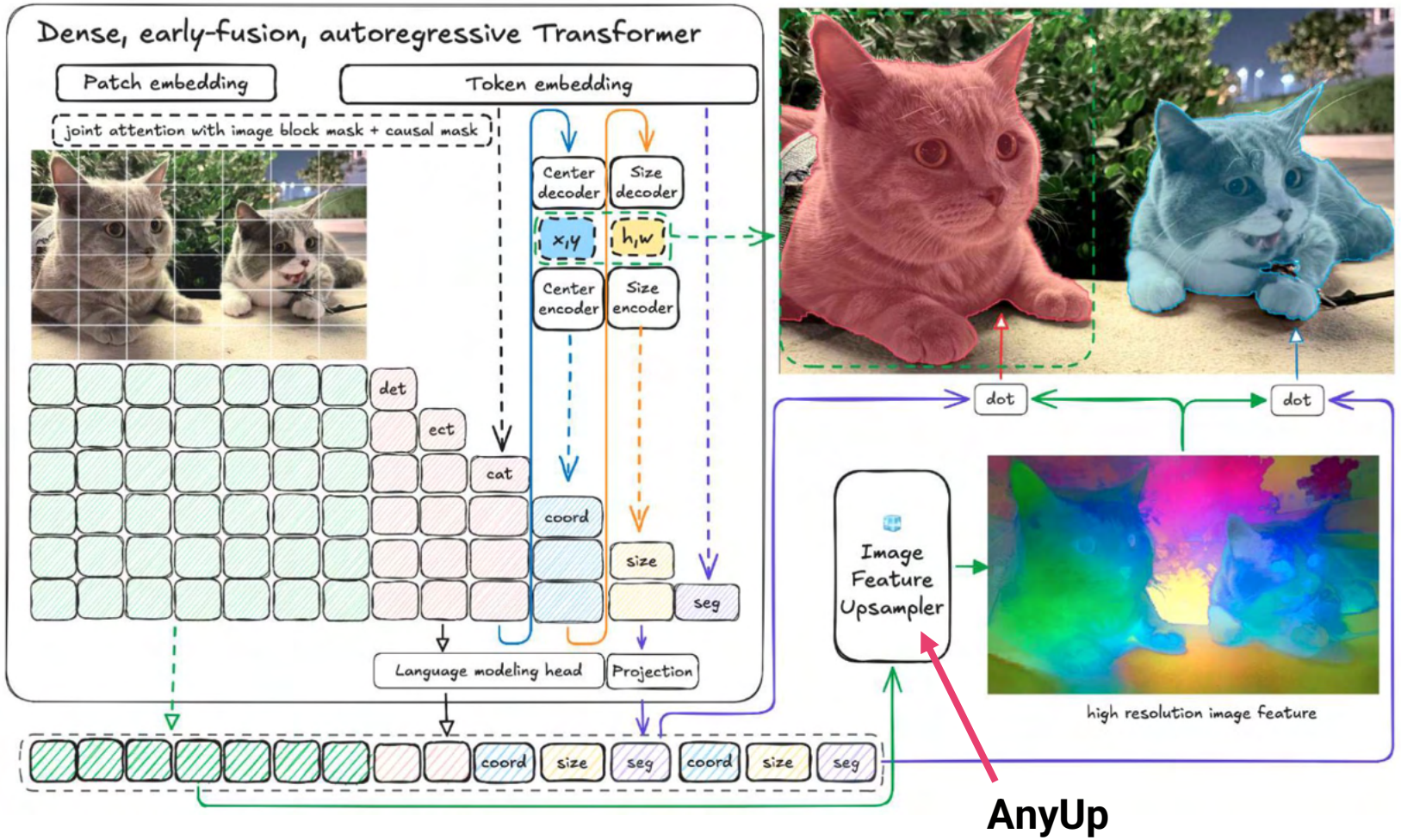
MAX PLANCK INSTITUTE FOR INFORMATICS · SIC Saarland Informatics Campus · ETH zürich · Google · TUM · JIA via RESEARCH CENTER · cls · ellis European Laboratory for Learning and Intelligent Systems

Where is AnyUp used in practice?

- Open-vocabulary segmentation (on edge devices)

Falcon Perception

Falcon Vision Team (Bevli et al.)
Preprint, 2026



Bevli, Aviraj, et al. "Falcon Perception." *arXiv preprint arXiv:2603.27365* (2026).

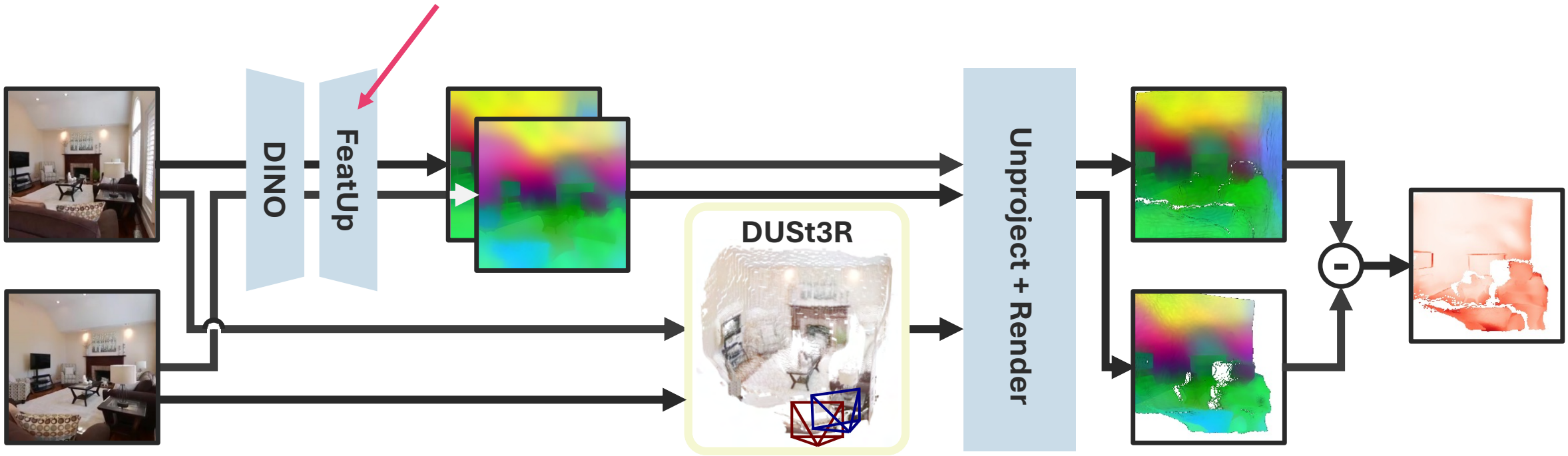
Where is AnyUp used in practice?

- Open-vocabulary segmentation (on edge devices)
- Consistency metrics

MEt3R: Measuring Multi-View Consistency in Generated Images

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, Jan Eric Lenssen
CVPR 2025

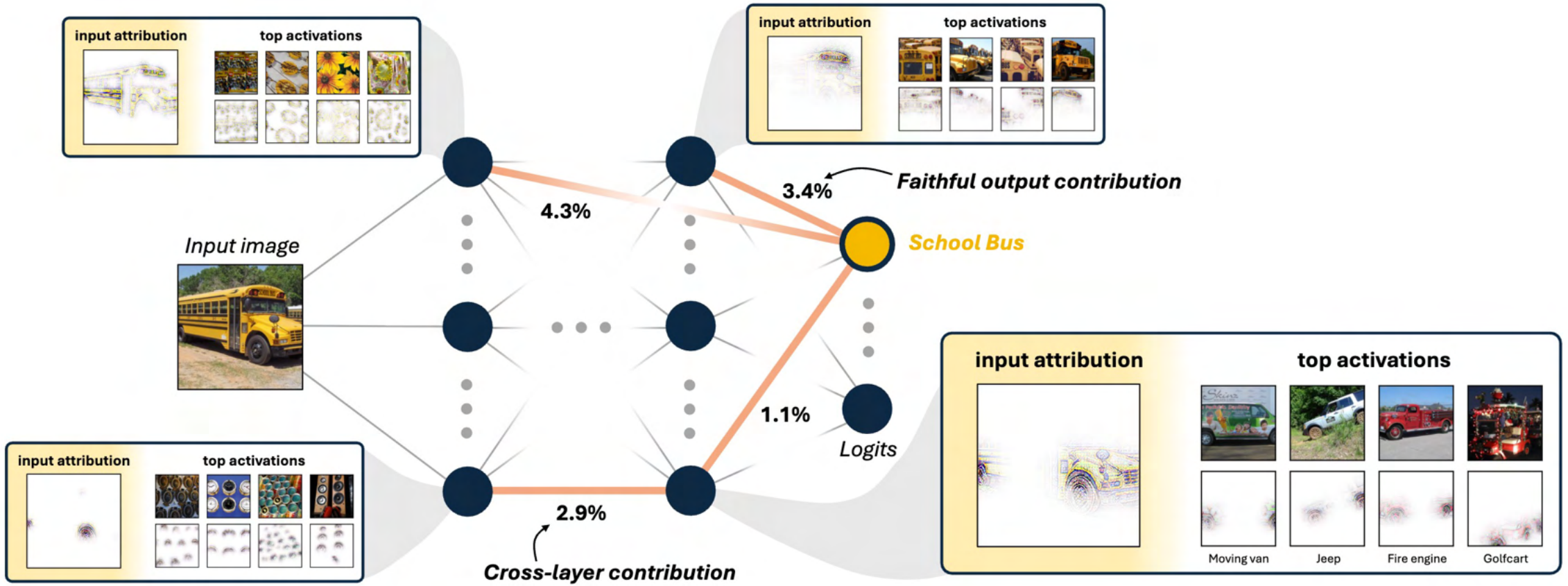
AnyUp now allows for using any source feature extractor



FaCT: Faithful Concept Traces for Explaining NN Decisions

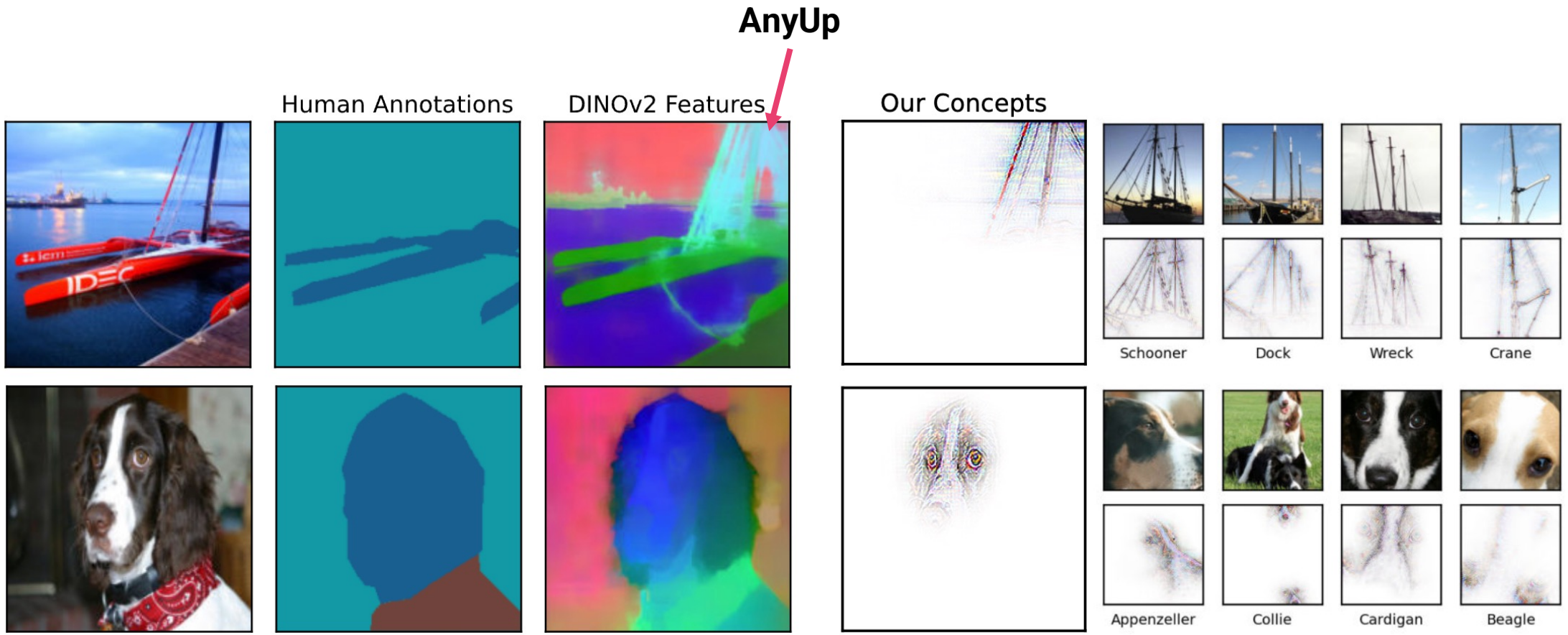
Amin Parchami-Araghi, Sukrut Rao, Jonas Fischer, Bernt Schiele

NeurIPS 2025



FaCT: Faithful Concept Traces for Explaining NN Decisions

Concept Consistency Metric



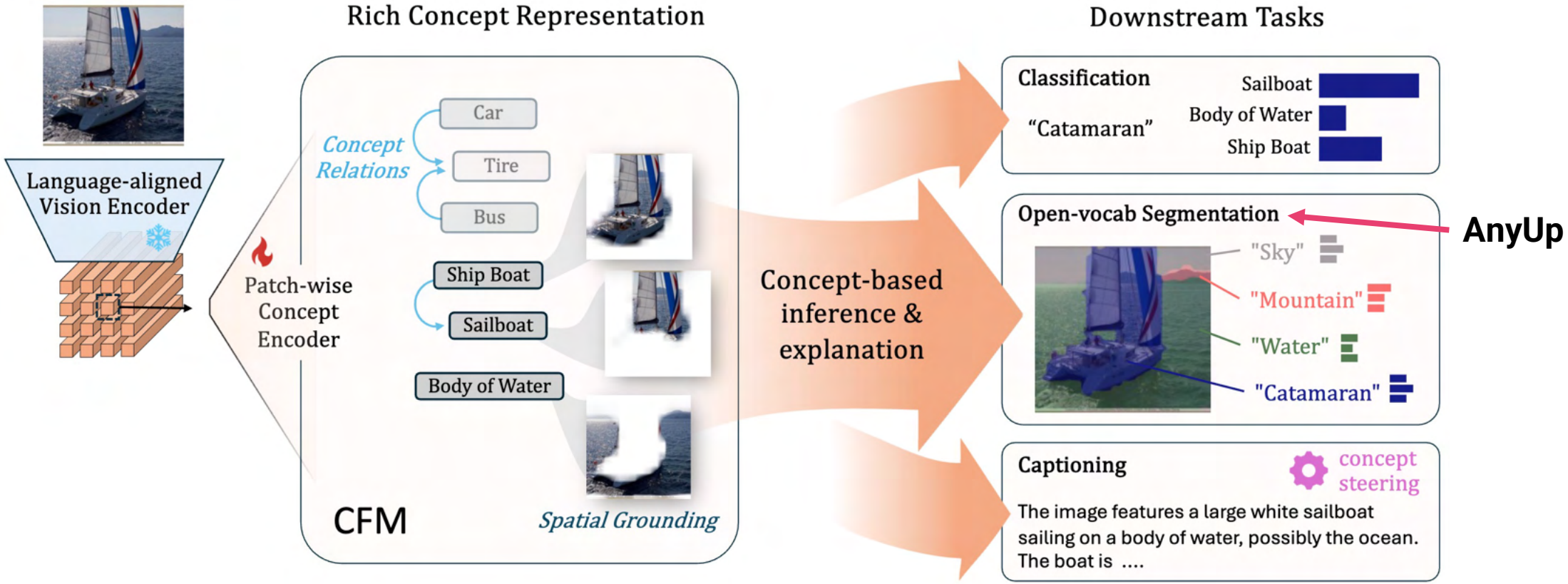
Parchami-Araghi, Amin, et al. "FaCT: Faithful Concept Traces for Explaining Neural Network Decisions." *NeurIPS*. 2026.

Where is AnyUp used in practice?

- Open-vocabulary segmentation (on edge devices)
- Consistency metrics
- Explainable AI

CFM: Language-aligned Concept Foundation Model for Vision

Kai Wittenmayer, Sukrut Rao, Amin Parchami-Araghi, Bernt Schiele, Jonas Fischer



Wittenmayer, Kai, et al. "CFM: Language-aligned Concept Foundation Model for Vision." *arXiv preprint arXiv:2601.13798* (2026).

Where is AnyUp used in practice?

- Open-vocabulary segmentation (on edge devices)
- Consistency metrics
- Explainable AI
- Other domains

Remote Sensing

DINO Soars: DINOv3 for Open-Vocabulary Semantic Segmentation of Remote Sensing Imagery

Ryan Faulkenberry
University of Houston
rfaulken@cougarnet.uh.edu

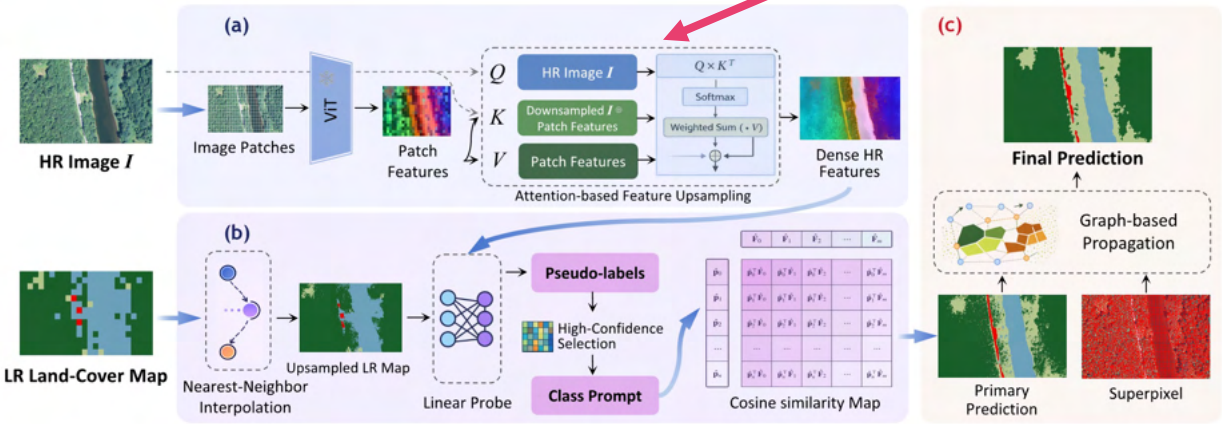
Saurabh Prasad
University of Houston
sprasad2@central.uh.edu

IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, VOL. 23, 2026 6011805

MapSR: Prompt-Driven Land-Cover Map Super-Resolution via Vision Foundation Models

AnyUp

Ruiqi Wang, Qi Yu, Jie Ma, *Member, IEEE*, and Hanlin Wu, *Member, IEEE*



Faulkenberry, Ryan, and Saurabh Prasad. "DINO Soars: DINOv3 for Open-Vocabulary Semantic Segmentation of Remote Sensing Imagery." *CVPR*. 2026.
Wang, Ruiqi, et al. "MapSR: Prompt-Driven Land Cover Map Super-Resolution via Vision Foundation Models." IEEE *Geoscience and Remote Sensing Letters* (2026).

Microscopy Imaging

Dense Embeddings from Self-Supervision and Foundation Models Improve Cell Linking Performance

Constantin Dalinghaus[1] CONSTANTIN.DALINGHAUS@UNI-GOETTINGEN.DE
Anwai Archit[1] ANWAI.ARCHIT@UNI-GOETTINGEN.DE
Constantin Pape[1,2,3] CONSTANTIN.PAPE@INFORMATIK.UNI-GOETTINGEN.DE
[1] *Georg-August-University Göttingen, Institute of Computer Science*
[2] *CAIMed - Lower Saxony Center for AI & Causal Methods in Medicine, Göttingen*
[3] *Cluster of Excellence Multiscale Bioimaging (MBExC), Georg-August-University Göttingen*

Evaluating Vision Foundation Models for Pixel and Object Classification in Microscopy

Carolin Teuber[1], Anwai Archit[1], Tobias Boothe[2], Peter Ditte[2], Jochen Rink[2,3], and Constantin Pape[1,4,5]

[1] Georg-August-University Göttingen, Institute of Computer Science
[2] Department of Tissue Dynamics and Regeneration, Max Planck Institute for Multidisciplinary Sciences, Göttingen
[3] Georg-August-University Göttingen, Faculty of Biology and Psychology
[4] CAIMed - Lower Saxony Center for AI & Causal Methods in Medicine, Göttingen
[5] Cluster of Excellence Multiscale Bioimaging (MBExC), Georg-August-University Göttingen

Teuber, Carolin, et al. "Evaluating Vision Foundation Models for Pixel and Object Classification in Microscopy." *arXiv preprint arXiv:2603.19802* (2026).
Dalinghaus, Constantin, Anwai Archit, and Constantin Pape. "Dense Embeddings from Self-Supervision and Foundation Models Improve Cell Linking Performance." *Medical Imaging with Deep Learning-Short Papers*. 2026.

Robotics

Contents lists available at ScienceDirect

Smart Agricultural Technology

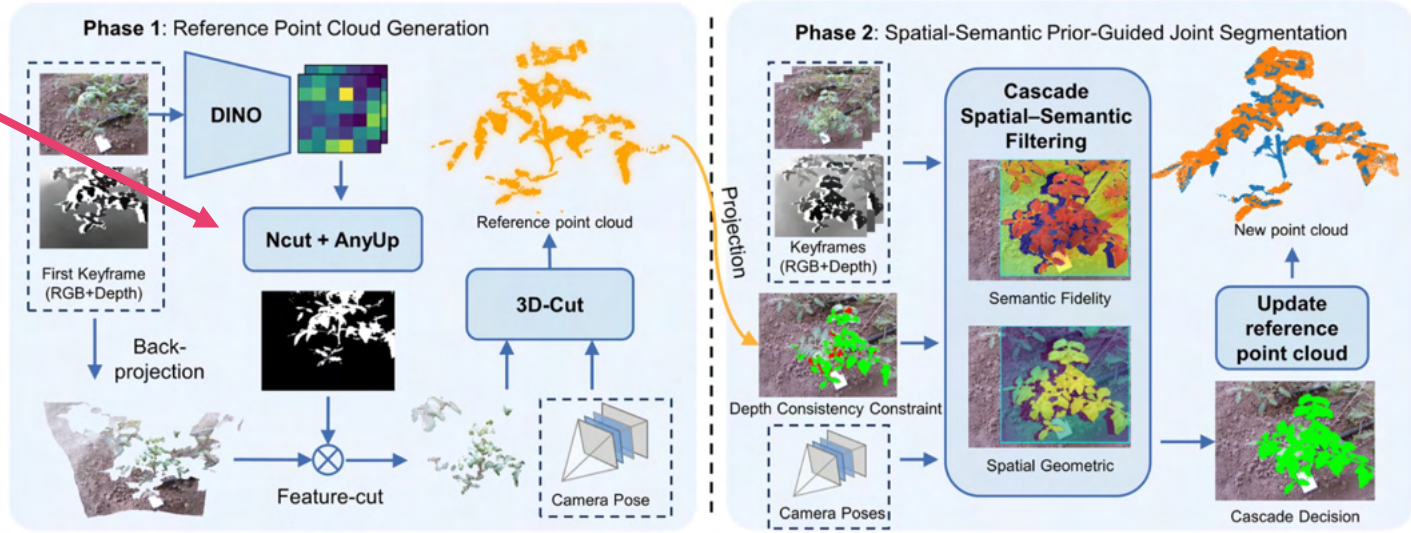
journal homepage: www.journals.elsevier.com/smart-agricultural-technology

RT-ZSDR: A real-time zero-shot segmentation and dense reconstruction framework for plant phenotyping robot

Yuanda Yang [a], Yuqi Zhang [b], Yida Li [b], Qiang Xu [c], Man Zhang [a], Han Li [a,*]

[a] Key Laboratory of Smart Agriculture System Integration, Ministry of Education, China Agricultural University, Beijing 100083, China
[b] Key Laboratory of Agricultural Information Acquisition Technology, Ministry of Agriculture and Rural Affairs, China Agricultural University, Beijing 100083, China
[c] Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou 450001, China

AnyUp



Yang, Yuanda, et al. "RT-ZSDR: A real-time zero-shot segmentation and dense reconstruction framework for plant phenotyping robot." *Smart Agricultural Technology* (2026): 101992.

Behind the scenes

You only ever see what went well, not the failures. Be aware of this bias!

Productive procrastination: “Side-questing” – being curious about new ideas, models and directions.

After reading a paper, **discuss it with peers**, and **verify** assumptions about new methods.

- Run small experiments with new models! Claude, Codex etc. can be really helpful with this.

ϵ**-progress**: Celebrate the small steps.

You are more than your research: Do something vastly different, a hobby, a community etc.

Thank you for your attention!

Time for Q&A