LoftUp: Learning a Coordinate-Based Feature Upsampler for Vision Foundation Models

Supplementary Material

This supplementary material to the main paper **"LoftUp: earning a Coordinate-Based Feature Upsampler for Vision Foundation Models"** is structured as follows:

- In Appendix A, we explain more implementation details of LoftUp training and the downstream tasks.
- In Appendix B, we show more quantitative results of LoftUp with CLIP and RADIO backbones and ablate the architecture and dataset size choices of training LoftUp. We also demonstrate additional results on the image matting task.
- In Appendix C, we provide more visualization of upsampled features, prediction results on various tasks, pseudo-GT, and cross-attention regions.

A. More Implementation Details

A.1. Training details of LoftUp

Our LoftUp upsampler is a 2-block cross-attention transformer that incorporates high-res image inputs with coordinates using an additional convolutional layer and low-res VFM features as keys and values in the cross-attention layers. Each transformer block consists of 1 cross-attention layer and 1 feedforward layer as in ViT [3]. To train LoftUp, we use a batch size of 8 and AdamW [17] optimizer with a learning rate of 1e-3 in Stage 1 and 1e-4 in Stage 2 for more stable improvement during self-distillation. In Stage 2, we take 2 random crops per image to construct crop(I_{HR}), and update the teacher upsampler's weights every 10 steps using the EMA of the student upsampler with a decay factor of 0.99. In both stages, we use $\alpha = 0.8$ for mask refinement when constructing pseudo-GT to balance sharp boundaries from masks and the fine-grained details from high-resolution features within each mask region. For all upsamplers, including our compared ones, we train for 1 epoch on a 1M-image subset of SA1B dataset [13].

A.2. Task setups

Semantic segmentation. Following [7, 8], we perform semantic segmentation on coarse classes in COCO-Stuff [1] (27 classes) and Cityscapes [2] (19 classes) and report mean Intersection-over-Union (mIoU) for each dataset. We train a linear decoder layer on upsampled features with a batch size of 8 and AdamW optimizer [17] with a learning rate of 1e-4 for 10 epochs.

Depth and normal estimation. Following [5], we evaluate depth and normal estimation using NAVI dataset [11] and train a DPT decoder head with 7 convolutional layers on top of the VFM features. We use a batch size of 8 and AdamW optimizer [17] with a learning rate of 5e-4 for 10 epochs. Following prior works [4–6], we report the root-mean-squared prediction error (RMSE) for both tasks and recall at δ_3 for scale-invariant depth estimation and at 30° for normal estimation. Here δ_3 is computed as the number of pixels whose ratio of depth prediction to groundtruth is less than 1.25^3:

$$\delta_3(d^{pr}, d^{gt}) = \frac{1}{N} \sum_{j \in N} \max\left(\frac{d_j^{pr}}{d_j^{gt}}, \frac{d_j^{gt}}{d_j^{pr}}\right) < 1.25^3,$$

where d^{pr} is predicted depth and d^{gt} is groundtruth depth.

Video object segmentation. This task involves propagating an object segmentation mask across video frames, given the ground truth mask for the first frame. Following prior evaluation protocols [10, 28], we compute dense feature affinity maps between frames to track objects. Performance is assessed using three metrics: J Mean, F Mean, and J & F Mean. Specifically, J Mean denotes the average Intersection-over-Union (IoU) between predicted segmentations and groundtruth masks, while F Mean represents the average F-score, measuring contour accuracy via precision and recall against groundtruth boundaries.We evaluate our method on the DAVIS validation set [22], a popular benchmark for video object segmentation. The dataset comprises 30 videos of varying lengths, each containing between 1 and 4 objects.

Zero-shot Open-Vocabulary Segmentation. We incorporate upsampled VFM features into ProxyCLIP [14], a state-of-the-art method for zero-shot open-vocabulary segmentation (OVSeg), and evaluate on three popular OVSeg benchmarks: COCO [15], Cityscapes [2], and ADE20K [31]. ProxyCLIP enhances CLIP features by leveraging spatial feature correspondence from VFMs as proxy attention, effectively inheriting the strong local consistency of VFMs while retaining CLIP's remarkable zero-shot transferability. Due to the high computational cost of proxy attention, we perform upsampling to 8× for all upsampling methods. We use CLIP ViT-B/16 [23] as the CLIP backbone, DINOv2-S/14 [20] as the proxy VFM, and set the input resolution to 336px, matching the resolution of the CLIP backbone.

Interactive Segmentation. We adapt the SimpleClick [16] architecture to evaluate upsampled features. Specifically, we use a frozen VFM backbone and train a single-layer click encoder that directly adds to the image patch embedding, along with a three-layer convolutional decoder head on top of the upsampled features for interactive segmentation. For training, we follow prior works [16, 27] and use the SBD dataset [9] to train for 20 epochs with the normalized focal loss [26, 27]. We employ the Adam optimizer [12] with a learning rate of 5e-5 and a batch size of 8. For evaluation, following common practice [13, 16, 27], we sample the first click point as the farthest point from the object boundary, and report the mean IoU of the predicted segmentation masks with the groundtruth, denoted as IoU@1 Click. We report results on three popular interactive segmentation benchmarks: GrabCut [25], Berkeley [18], and DAVIS [21].

B. More Quantitative Results

Resolution	Upsampler	COCO	Cityscapes
224	NA	40.30	30.79
	Bilinear	47.12	39.84
	FeatUp	52.08	33.50
	LoftUp	**52.58**	**44.66**
448	NA	42.14	37.36
	Bilinear	48.32	45.58
	FeatUp	52.55	40.00
	LoftUp	**53.87**	**50.14**

Table B.1. **Comparison of feature upsampers when VFM is CLIP-B/16 [23].**

Resolution	Upsampler	COCO	Cityscapes
224	NA	51.00	34.42
	Bilinear	56.77	43.09
	FeatUp	56.59	42.23
	LoftUp	**58.36**	**46.98**
448	NA	58.94	49.30
	Bilinear	62.29	57.42
	FeatUp	62.18	56.58
	LoftUp	**63.55**	**60.83**

Table B.2. **Comparison of feature upsampers when VFM is RADIOv2.5-B [24].**

Upsampling CLIP and RADIO. In Tab. B.1 and Tab. B.2, we compare LoftUp with FeatUp and a bilinear upsampling baseline using CLIP [23] and RADIO [24] as VFM backbones. The upsamplers are trained following the same procedure as on the DINOv2 backbone and evaluated at resolutions 224 and 448. As with DINOv2, LoftUp consistently outperforms all baselines when using CLIP and RADIO as VFM backbone, demonstrating the general applicability of our approach across different VFMs.

	Feat PE	Image conv	# blocks	# Train data	COCO	Cityscapes
(a)	no	1x1	2	50k	56.46	43.13
	learnable	1x1	2	50k	57.89	46.06
	learnable	3x3	2	50k	58.40	47.35
	Sine	1x1	2	50k	58.10	47.24
	Sine	3x3	2	50k	**58.65**	**48.63**
	RoPE	3x3	2	50k	58.56	48.50
(b)	Sine	3x3	1	50k	57.94	48.13
	Sine	3x3	2	50k	**58.65**	**48.63**
	Sine	3x3	3	50k	58.35	48.32
(c)	Sine	3x3	2	50k	58.65	48.63
	Sine	3x3	2	200k	59.40	49.84
	Sine	3x3	2	1M	**59.87**	**50.43**

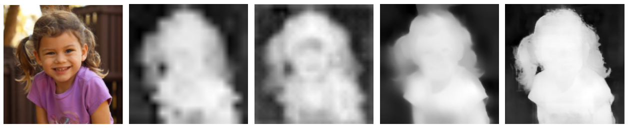
Table B.3. **Ablation of architecture and training data size choices.** Upsamplers are trained using only Stage 1 loss for convenience.

Ablation on the architecture and training data size. In Tab. B.3, we conduct an ablation study on both the architecture components of LoftUp and the training data size. For convenience, the upsamplers are trained using only the Stage 1 training objective. Specifically, in experiment (a), we demonstrate that employing a sinusoidal positional encoding for the low-resolution features—combined with a 3x3 convolutional layer to process the high-resolution coordinates and image inputs—yields improved performance. This result is in line with prior work showing that sinusoidal positional encodings excel in coordinate-based methods [19, 29, 30] and that stronger image processing layers help better integrate high-resolution information. In experiment (b), we observe that two blocks of the cross-attention transformer are sufficient for optimal feature upsampling, with performance saturating at greater depths. Finally, in experiment (c), we find that training with a larger dataset improves performance, although the benefits begin to diminish as the dataset size increases. Consequently, we select a 1M-subset of the SA1B dataset [13] to achieve the best balance among data diversity, model performance, and training time.

Additional experiments on image matting. To further validate LoftUp's task-agnostic design, we evaluated on image matting using a linear probing layer. LoftUp consistently outperforms all baselines, demonstrating its ability to generalize to semantic-sensitive scenarios - for example, accurately matting the intricate hair strands of the subject. See Tab. B.4 and Fig. B.1.

Datasets	Bilinear	LiFT	FeatUp	LoftUp
Matting Human	0.0312	0.0380	0.0217	**0.0143**
COCO Matting	0.1578	0.1586	0.1264	**0.1080**

Table B.4. **Image matting comparison (MSE)**



(a) Img (b) Bilinear (c) LiFT (d) FeatUp (e) LoftUp

Figure B.1. **Qualitative comparison on image matting.**

C. More Visualization

We further provide more visualization examples of upsampled features of various methods in Fig. C.1, more prediction examples in semantic segmentation, depth estimation, and video object segmentation in Fig. C.2 and Fig. C.3, more examples of different pseduo-GT in Fig. C.4, and more examples of the attended regions of a high-resolution pixel in Fig. C.5.

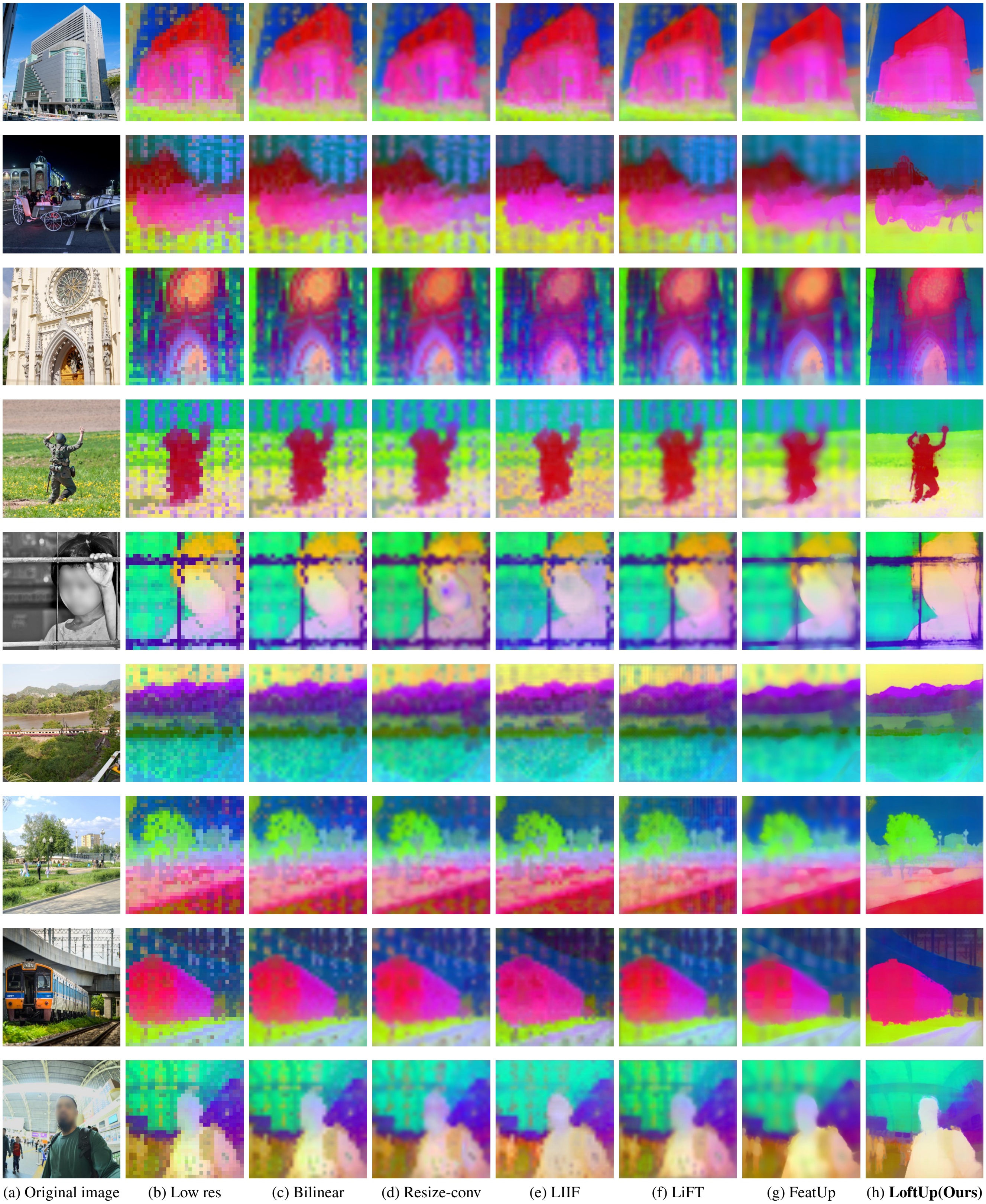
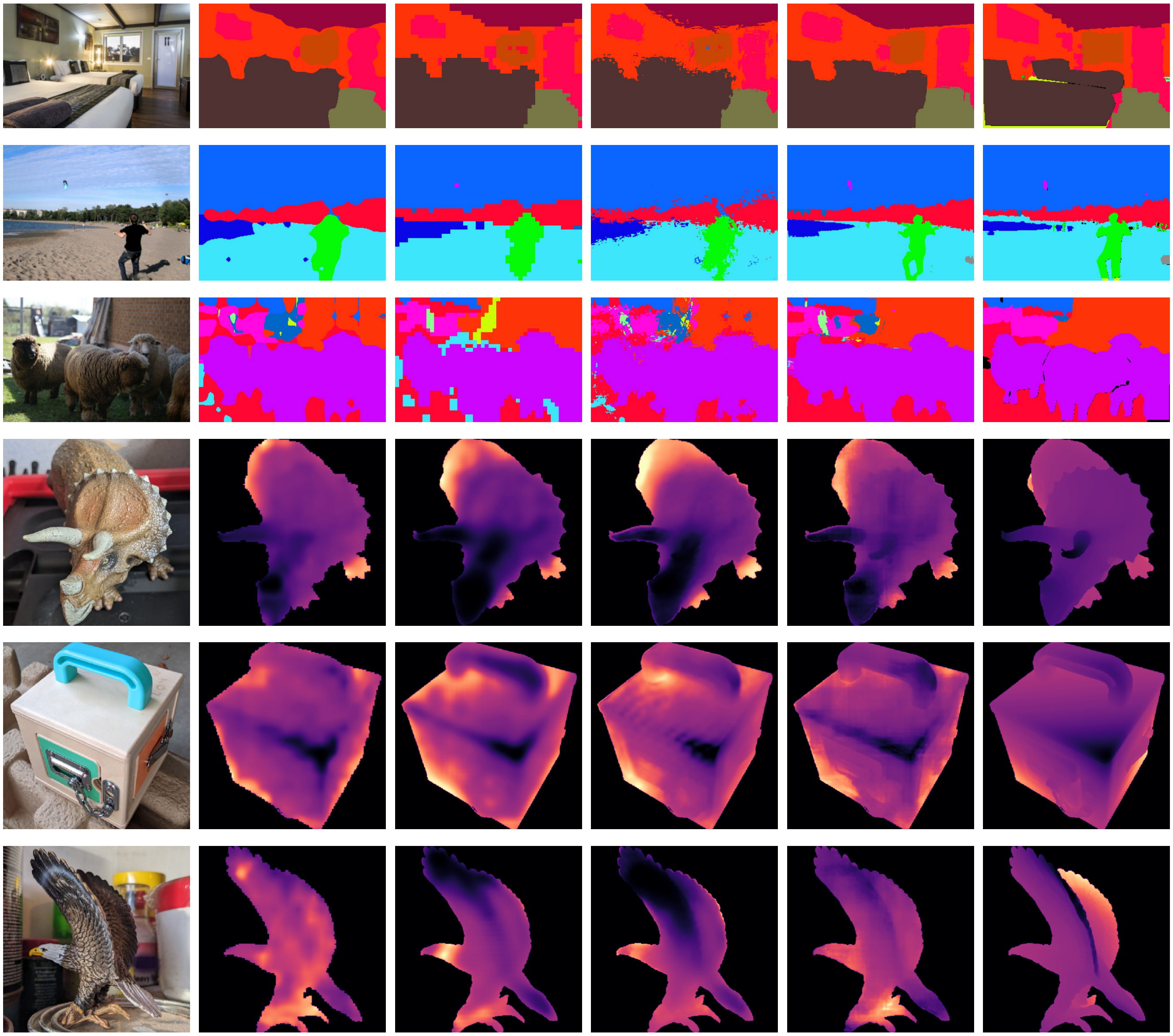
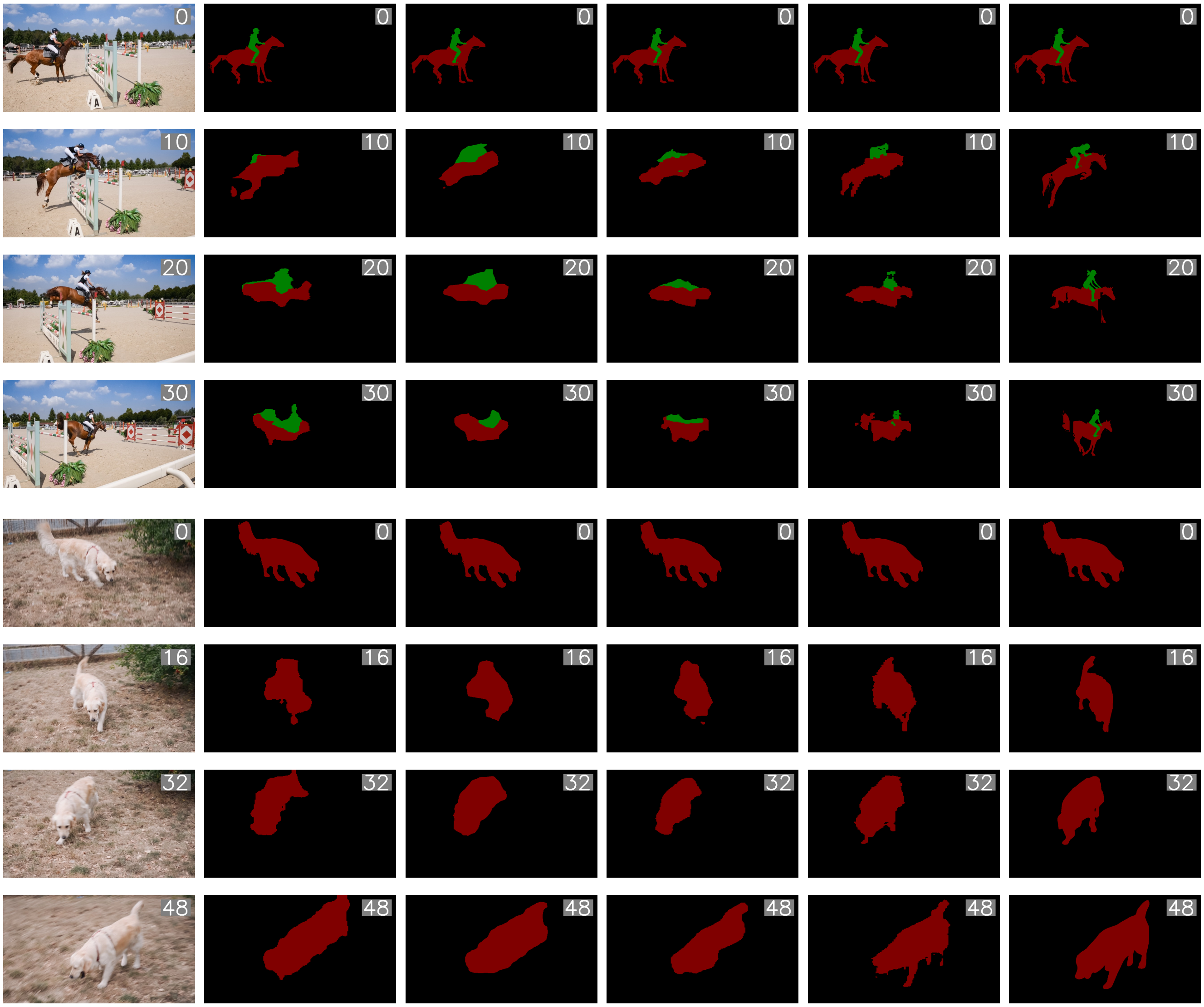


Figure C.1. **More visualization of features from various upsamplers.** Backbone is DINOv2-S/14 [20].



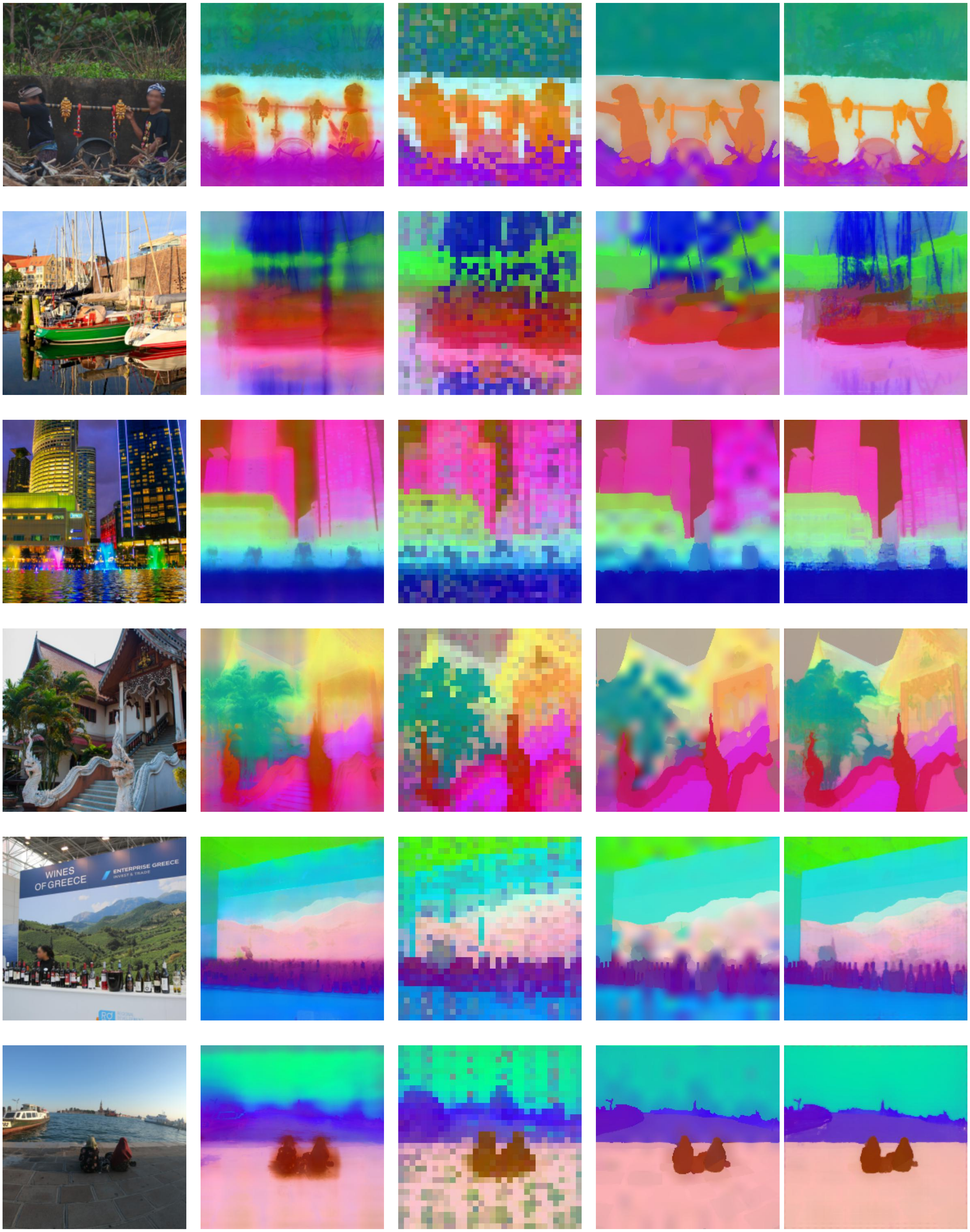
(a) Original image (b) Bilinear (c) LiFT (d) FeatUp (e) **LoftUp** (f) Groundtruth

Figure C.2. **More visualization of predictions examples** on semantic segmentation on COCO-Stuff [1] and depth estimation on NAVI [11].



(a) Original image (b) Bilinear (c) LiFT (d) FeatUp (e) **LoftUp** (f) Groundtruth

Figure C.3. **Visualization of prediction examples** of video object segmentation on the DAVIS 2017 dataset [22]. Each image displays its corresponding frame number in the top right corner. The groundtruth segmentation for the 0-th frame is provided, and dense feature affinity maps are employed to propagate its segmentation labels to subsequent frames. We can see that LoftUp outperforms all the other baselines in accurately tracking the objects across the frames.



(a) Original image (b) Per-image optimized (c) 2x features (LiFT) (d) Mask-Bicubic (e) Self-Distilled

Figure C.4. **More visualization of different pseudo-GT.**

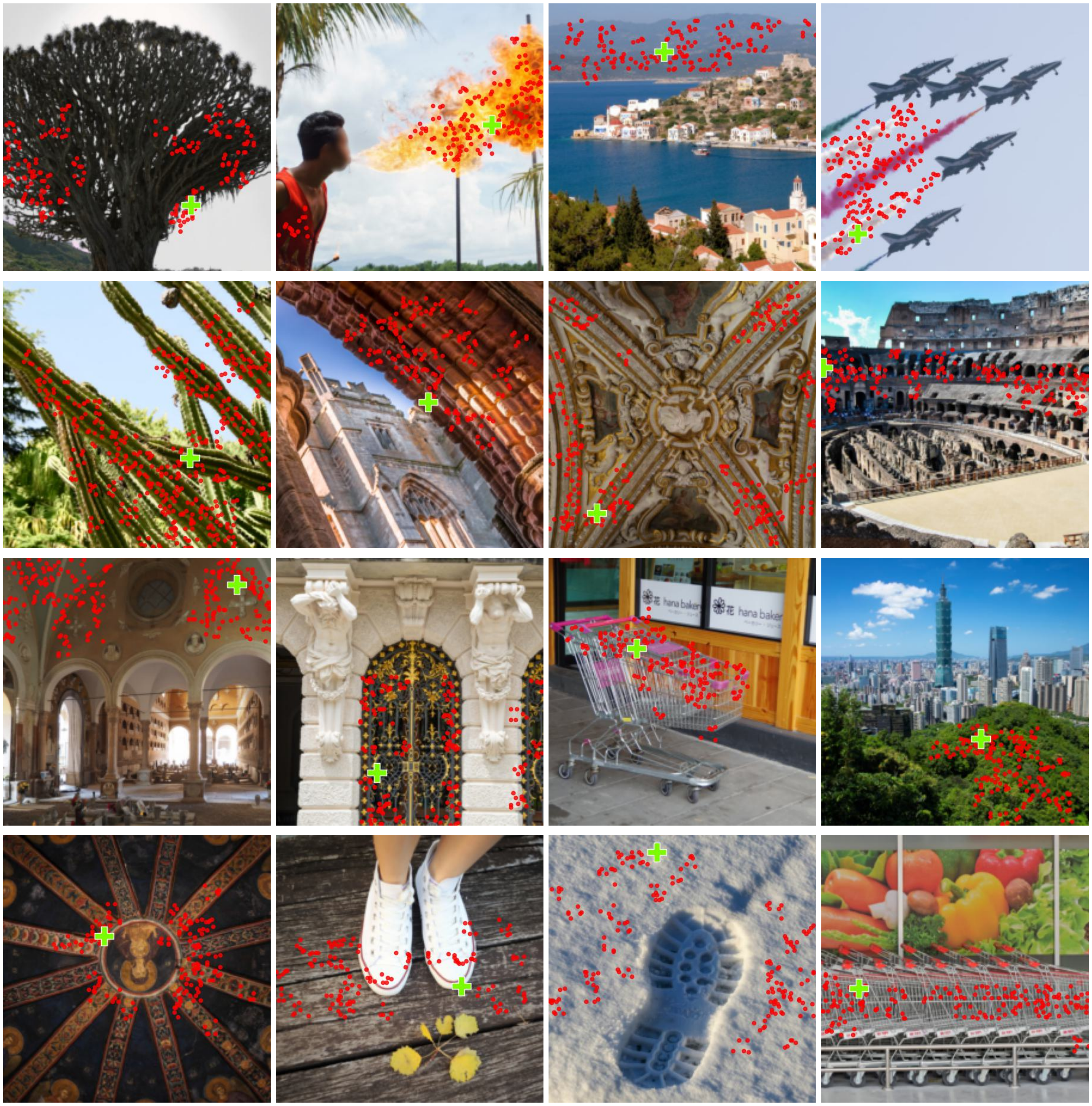


Figure C.5. **Visualization of attended region** (in dots) in the low-res features of a high-res pixel (in cross). The density of dots reflects the value of the attention map. LoftUp is able to use relevant information across the global feature map for upsampling features at each pixel.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 1, 5

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1

[4] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 27, 2014. 1

[5] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *CVPR*, pages 21795–21806, 2024. 1

[6] David F Fouhey, Wajahat Hussain, Abhinav Gupta, and Martial Hebert. Single image 3d without a single 3d image. In *ICCV*, pages 1053–1061, 2015. 1

[7] Stephanie Fu, Mark Hamilton, Laura E. Brandt, Axel Feldmann, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution. In *ICLR*, 2024. 1

[8] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022. 1

[9] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, pages 991–998. IEEE, 2011. 2

[10] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *NeurIPS*, 33: 19545–19560, 2020. 1

[11] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, Andre Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. NAVI: Category-agnostic image collections with high-quality 3d shape and pose annotations. In *NeurIPS*, 2023. 1, 5

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 1, 2

[14] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *ECCV*, 2024. 1

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1

[16] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *ICCV*, pages 22290–22300, 2023. 1, 2

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[18] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, pages 416–423. IEEE, 2001. 2

[19] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[20] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. 2024. Featured Certification. 1, 4

[21] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016. 2

[22] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 1, 6

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. pages 8748–8763. PmLR, 2021. 1, 2

[24] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *CVPR*, pages 12490–12500, 2024. 2

[25] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM TOG*, 23(3):309–314, 2004. 2

[26] Konstantin Sofiiuk, Olga Barinova, and Anton Konushin. Adaptis: Adaptive instance selection network. In *ICCV*, pages 7355–7363, 2019. 2

[27] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *ICIP*, pages 3141–3145. IEEE, 2022. 2

[28] Saksham Suri, Matthew Walmer, Kamal Gupta, and Abhinav Shrivastava. Lift: A surprisingly simple lightweight feature transform for dense vit descriptors. In *ECCV*, pages 110–128. Springer, 2024. 1

[29] Jianglong Ye, Naiyan Wang, and Xiaolong Wang. Featurenerf: Learning generalizable nerfs by distilling foundation models. In *ICCV*, pages 8962–8973, 2023. 2

[30] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. 2021 ieee. In *CVPR*, 2020. 2

[31] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 1