Supplementary Material for

DUNE: Distilling a Universal Encoder from Heterogeneous 2D and 3D Teachers

Mert Bülent Sarıyıldız Philippe Weinzaepfel Thomas Lucas
Pau de Jorge Diane Larlus Yannis Kalantidis
NAVER LABS Europe
https://europe.naverlabs.com/dune

Contents

A. Training protocol details

A.1. Datasets

The 19 datasets that we use for co-distillation are listed in Tab. 1, and a few examples per dataset are provided in Figs. 8 and 9. As can be seen from Tab. 1, the datasets are quite unbalanced in size. During training, we construct a batch such that it contains an equal amount of randomly sampled images from the datasets associated with each teacher, *i.e*. DINO-v2, Multi-HMR and MASt3R.

Access to teacher training data. For presentation clarity and without loss of generality, in the main paper we assume that all the data used to train all the teachers is also available for distillation. This is in practice impossible at times, either because a subset of the dataset might not be public, or because of their size. In such cases, one can use only the subset of the datasets that is available, or source alternative data across all domains. This extends beyond distillation to the data used for finetuning.

Name	Size	Nature	Teacher
ImageNet-19K [11, 34]	13,153,480	Real	DINO-v2
Mapillary [32]	1,205,907	Real	
Google Landmarks v2 [33]	4,132,914	Real	
Habitat [19]	284,968	Rendered	MAST3R
ARKitScenes [10]	456,108	Rendered	
Blended MVS [35]	98,937	Rendered	
MegaDepth [16]	36,949	Real	
ScanNet++ [36]	60,188	Rendered	
CO3D-v2 [23]	185,100	Real	
Map-free [3]	41,300	Real	
WildRgb [1]	224,400	Real	
VirtualKitti [7]	1,200	Synthetic	
Unreal4K [28]	14,386	Synthetic	
TartanAir [31]	136,225	Real	
DL3DV [18]	208,800	Rendered	
BEDLAM [6]	353,118	Synthetic	Multi-HMR
AGORA [21]	14,314	Synthetic	
CUFFS [4]	54,944	Synthetic	
UBody [17]	54,234	Real	
Total size:	**20,717,472**		

Table 1. **Datasets used for training DUNE models.** The teacher column groups the datasets which are associated with each teacher.

Sample images from all datasets. In Figs. 8 and 9, we visualize 10 randomly sampled images from each dataset listed in Tab. 1.

A.2. Table of hyper-parameters

A table with the set of hyper-parameters we use for training our DUNE models are given in Tab. 2. Further details can be found on github.

Distillation loss. Following [24], given an image x, we minimize the combination of the cosine and smooth-ℓ_1 losses between the outputs of student $\boldsymbol{s}_i = h_i(f(x))$ and each teacher $\boldsymbol{t}_i = t_i(x)$:

$$\mathcal{L}_{\text{distil}} = \sum_{i=1}^{N} \mathcal{L}_{cos}(\boldsymbol{s}_i, \boldsymbol{t}_i) + \mathcal{L}_{s\ell_1}(\boldsymbol{s}_i, \boldsymbol{t}_i), \qquad (1)$$

Hyper-parameter	Value
Encoder	Architecture: ViT-Base Patch size: 14 Num. registers: 0 QKV bias: True LayerScale: True Path drop rate: 0
Projector	Architecture: TP Num. blocks: 1 Block configuration follows encoder
Image resolution	Initial: 336×336 Fine-tuned: 448×448
Batch size	128 per GPU
Num. GPUs	4
Optimizer	Type: AdamW Weight decay: $3e-2$ (β_1, β_2): $(0.9, 0.99)$
Learning rate	Min: $1e-6$ Max: $3e-4 \times$ batch-size$/256$ Schedule: Cosine
Data type	AMP with bloat16
Training data (Tab 1 in the main paper)	All (DUNE-20.7M, see Tab. 1)
Data sharing (Tab 2 in the main paper)	Full data sharing
Training budget	$1,281,167 \times 100$ images

Table 2. **Hyper-parameters used for training DUNE models.**

where

$$\mathcal{L}_{cos}(\boldsymbol{s}, \boldsymbol{t}) = 1 - \frac{\boldsymbol{s} \cdot \boldsymbol{t}}{||\boldsymbol{s}||_2 \times ||\boldsymbol{t}||_2}, \tag{2}$$

$$\mathcal{L}_{sl1}(\boldsymbol{s}, \boldsymbol{t}) = \begin{cases} 0.5 \times ||\boldsymbol{s} - \boldsymbol{t}||_2^2, & \text{for } ||\boldsymbol{s} - \boldsymbol{t}||_1 < 1, \\ ||\boldsymbol{s} - \boldsymbol{t}||_1 - 0.5, & \text{otherwise.} \end{cases} \tag{3}$$

B. Details on decoder fine-tuning

MASt3R. MASt3R relies on a binocular architecture with a Siamese ViT-encoder to encode the input images, followed by binocular decoders and prediction head. When finetuning this model, we simply replace the encoder and keep it frozen using the publicly available code of MASt3R [15]. Given the size of the decoders and heads, we initialize them with the released models, except for weights that have a mismatch of size, namely the fully-connected layer between the encoder and decoder, as our ViT-Base encoder has a smaller feature dimension than their ViT-Large one (768 *vs*. 1024) as well as the output layers that outputs a pixelwise prediction due the mismatch of patch sizes (14 *vs*. 16). We finetune the model on 6.5M image pairs with AdamW on images at different resolutions. For backbone distilled on 336×336 images, we use $\{448 \times 448$, 448×336, 448×294, 448×252, 448×224, $448 \times 140\}$, which corresponds to the same number of patches as MASt3R's setting. For backbone further distilled on 448×448 images, we use $\{518 \times 518$, 518×392, 518×336, 518×294, 518×252, $518 \times 168\}$ which corresponds to the resolutions close to the ones from MASt3R but that are multiple of 14.

Multi-HMR. To evaluate our model on the task of Human Mesh Recovery (HMR), we use the training framework and public code of Multi-HMR [4]. We discard the projector modules and freeze the weights of the distilled student model. The Human Perception Head (HPH) proposed in Multi-HMR is used to predict HMR from the outputs of the backbone, with two transformer blocks prepended to it. This head is trained from scratch on the BEDLAM dataset, using images at a resolution of 672×672. Training is done with a learning rate of $4e-5$, a batch size of 16, and a cosine decay schedule over 200k iterations. After training, evaluation is performed on the BEDLAM validation set with a non-maximum suppression (NMS) kernel of size 3 and a detection threshold of 0.3, following the Multi-HMR protocol.

Notably, this evaluation procedure favors the teacher model, as its native resolution is 672×672, whereas the student model is distilled on images of resolution 448×448 only due to computational constraints.

Semantic segmentation and depth estimation evaluations. Semantic segmentation and depth estimation are dense prediction tasks, both formulated as classification tasks in this work, and solved following the simple setup proposed in [20], also followed by the most recent related works [22, 24]. We extract the tokens from the last output layer of the student model and use as input to a linear prediction head. For semantic segmentation, we additionally use the Transformer Projector of the DINO-v2 teacher as part of the frozen encoder, and train a linear head on top of the projector. to predict class logits from a patch token. This yields a 32×32 logit map that is upsampled via bilinear interpolation to the original image resolution of 512×512.

For depth estimation, we first upsample patch features by a factor of 4 via bilinear interpolation, concatenate them along the feature dimension with the CLS token, and use these vectors as input to a linear layer. Depth prediction is treated as a soft classification task following [5]; we use 256 uniformly distributed bins.

C. Attention map visualizations

In Fig. 2 of the main paper, we present a visualization of the encoder outputs from the teacher models and our student model using principal component analysis (PCA). This analysis is conducted on three randomly selected images from the Map-free and BEDLAM datasets. The visualization reveals that patch similarity patterns differ across the teacher models, while our student model appears to simultaneously attempt to capture and integrate multiple patterns from the different teachers.

To further investigate this phenomenon, we visualize in Fig. 2 the attention probabilities obtained at the last en-

coder layer of the student model, as well as those of the three teacher-specific Transformer Projectors (TP) attached on top during distillation. More concretely, given an image of size 448×448, we extract the 32×32 attention map for all the 1024 patches (the patch size for the student model is 14). In order to see the most prototypical attention patterns, we flatten all patch attentions and cluster them via k-Medoids ($k = 9$), with the version available in Scikit-Learn.[1]

We indeed observe different attention patterns for the last encoder block and the Transformer projectors. For instance, the projector for MASt3R yields much more localized attentions regardless of the input image compared to the projector for DINO-v2, whose attentions have much wider spatial extent. We also notice that the projector for Multi-HMR focuses mainly on the human, when there is one in the image (see Fig. 2).

Looking at the attentions of the last layer of the encoder, however, we observe once again that it seems to try to capture a mixture of the attentions of the three projectors: They exhibit a strong locality as in MASt3R, a spatial extent similar to DINO-v2, and also a strong preference for humans.

D. Additional Results

In this section, we provide additional evaluations for DUNE models. We report results for MASt3R with a DUNE encoder on multi-view depth estimation and camera pose regression tasks, as well as semantic segmentation performance on additional datasets and comparisons to 2D-to-3D distillation methods. Furthermore, we evaluate our models on Feat2GS, a recently proposed benchmark for assessing models' 3D awareness in geometry and texture via novel view synthesis, and present extended qualitative results.

D.1. Multi-view depth evaluation

We follow the protocol of [26] and evaluate multi-depth stereo depth evaluation on KITTI [12], DTU [2], ETH3D [25], Tanks And Temples [14] and ScanNet [9]. We report the Absolute Relative Error (rel) and the Inlier Ratio (τ) with a threshold of 1.03 on each test set, as well as the averages over all test sets. To extract depth prediction of one image, we follow DUSt3R [30] and extract depthmaps as the z-coordinate of the predicted pointmaps; and when multiple pointmaps are available for one image from different image pairs, we simply rescale the predicted depthmaps and average them with weights given by the predicted confidence values. Results are reported in Table 3. DUNE performs similarly to MASt3R and DUSt3R on this task overall, while using a smaller ViT-Base image encoder.

[1] https://scikit-learn-extra.readthedocs.io/

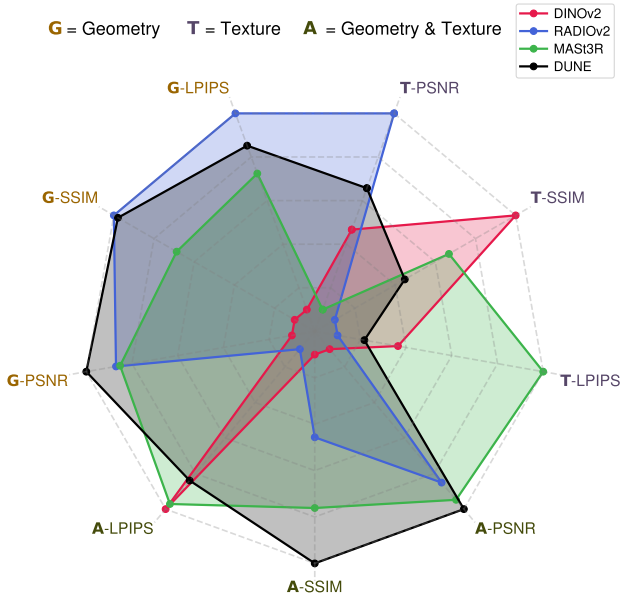

Figure 1. **Evaluating DUNE on the Feat2GS benchmark.** The spider plot shows comparisson of different encoder models. The Feat2GS benchmark [8] evaluates Novel View Synthesis as a proxy for 3D awareness. In all metrics, larger distance to the center indicates better performance. Note that models vary in size: RADIOv2 is a ViT-H, MASt3R a ViT-L and DINO-v2 and DUNE a ViT-B.

D.2. Multi-view camera pose regression evaluation

Following the protocol of [15, 29], we evaluate on the task of multi-view pose estimation on the CO3Dv2 [23] and RealEstate10K [38] datasets using sequences of 10 images. Matches obtained as output of the MASt3R decoder and head for an image pair are used to estimate Essential Matrices and relative pose. We report the Relative Rotation Accuracy (RRA) and Relative Translation Accuracy (RTA) on image pairs at a threshold of 15°, as well as the mean Average Accuracy (mAA30), *i.e.*, the area under the accuracy curve of the angular differences (RRA@30, RTA@30). Results are reported in Table 4. DUNE performs on par with DUSt3R and MASt3R on the object-centric Co3Dv2 dataset, while it outperforms them on the more challenging RealEstate10K dataset. Once again, DUNE uses a ViT-Base encoder while DUSt3R and MASt3R are based on a ViT-Large encoder.

D.3. Evaluating DUNE on the Feat2GS benchmark

In Fig. 1 we compare different encoder models in the Feat2GS benchmark [8]. The Feat2GS benchmark has three modalities, *i)* **G**eometry: When only geometry parameters are predicted from features and texture is free-optimized for Novel View Synthesis. *ii)* **T**exture: When only the texture is predicted from encoder features and the geometry is free-

Method	Encoder	KITTI		ScanNet		ETH3D		DTU		T&T		Average	
		rel. ↓	τ ↑	rel. ↓	τ ↑	rel. ↓	τ ↑	rel. ↓	τ ↑	rel. ↓	τ ↑	rel. ↓	τ ↑
DeepV2D [27]	Hourglass	10.00	36.20	4.40	54.80	11.80	29.30	7.70	33.00	8.90	46.40	8.60	39.90
DUSt3R [30]	ViT-Large	5.88	47.67	**3.01**	**72.54**	3.04	75.17	2.92	73.94	2.93	78.51	3.56	69.56
MASt3R [15]	ViT-Large	**3.54**	**65.68**	4.17	65.22	**2.44**	**82.77**	3.46	66.89	**2.04**	**87.88**	**3.13**	**73.69**
DUNE	ViT-Base	4.88	50.76	4.24	59.68	2.48	77.97	**2.69**	**75.63**	2.60	79.19	3.38	68.65

Table 3. **Multi-view depth evaluation** with the absolute relative error (rel) and the inlier ratio (τ) on several test sets, and the average across all test sets in the last column. DeepV2D uses ScanNet in the training set, explaining its better performance on this dataset. DUNE uses a ViT-Base encoder while DUSt3R and MASt3R a ViT-Large encoder.

Method	Encoder	Co3Dv2 ↑			RealEstate10K ↑
		RRA@15	RTA@15	mAA(30)	mAA(30)
DUSt3R [30]	ViT-Large	93.3	88.4	77.2	61.2
MASt3R [15]	ViT-Large	**94.6**	**91.9**	**81.8**	76.4
DUNE	ViT-Base	92.2	90.7	78.8	**79.9**

Table 4. **Multi-view pose regression evaluation** on the CO3Dv2 [23] and RealEstate10K [38] datasets with 10 random frames. DUNE uses a ViT-Base encoder while DUSt3R and MASt3R a ViT-Large encoder.

Model	Cityscapes (mIoU ↑)	NYUv2 (mIoU ↑)	ScanNet (mIoU ↑)	Avg. (mIoU ↑)
Pri3D [13]	56.3	54.8	61.7	57.6
MASt3R [15]	58.9	60.2	57.0	58.7
DUNE (no proj.)	65.6	66.1	61.2	64.3
DUNE	**70.6**	**68.2**	**65.2**	**68.0**

Table 5. **Additional semantic segmentation evaluations.** As described in the paper, for improved segmentation performance we can use the DINO teacher projector as part of the frozen encoder, and learn a linear classifier on top.

optimized. And *iii)* **A**ll: When both geometry and texture are predicted from features. Our DUNE encoder leads to the best performance when **A**ll the parameters are predicted from features (to our understanding the most challenging setting) and leads to the largest area over all settings and metrics. For more detailed results, we also present Tab. 7 with per-dataset evaluations of all metrics and modalities. While all encoders use a ViT architecture, they vary significantly in size, mainly due to the absence of ViT-B models for certain methods. Namely, RADIOv2 only has a ViT-H model open-sourced and MASt3R a ViT-L, DINO-v2 and our model DUNE are ViT-B. Thus, the fact DUNE is obtaining the overall best performance compared to much larger models is even more remarkable.

D.4. Comparison to 3D-to-2D distillation and 3D-uplifting methods

In Tab. 5, we report semantic segmentation evaluations on three datasets, comparing DUNE to Pri3D [13] (a 3D-to-2D distillation method) and MASt3R. For DUNE, we present results using the encoder outputs directly, DUNE (no proj.), and with the DINO-v2 projector applied after the encoder, DUNE. In all cases, only a linear layer is trained to predict patch labels. DUNE significantly outperforms both Pri3D and MASt3R.

Model	NYUv2 (RMSE ↓)
FiT-3D [37]	0.380
DUNE	0.358

Table 6. **Comparisson to Fit-3D** on monocular depth.

In Tab. 6, we evaluate depth estimation performance on NYUv2, comparing DUNE to FiT-3D [37], a recent method that enhances DINO-v2 features for 3D tasks. DUNE achieves substantially better performance than FiT-3D.

D.5. Qualitative comparisons of teacher outputs to DUNE

MASt3R. In Figs. 3 to 5 we present qualitative results for MASt3R and our student side-by-side. We see that the student clearly improves over the teacher in some cases.

Multi-HMR. In Figs. 6 and 7 we present qualitative results for images randomly sampled from the bedlam validation set, comparing the outputs of the student and teacher. Both models achieve results of comparable visual quality.

References

[1] RGBD objects in the wild: Scaling real-world 3D object learning from RGB-D videos. *arXiv:2401.12592*, 2024. 1

[2] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *IJCV*, 2016. 3

	LLFF									DL3DV									Casual								
	Geometry			Texture			All			Geometry			Texture			All			Geometry			Texture			All		
Feature	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DINO-v2	19.69	.7405	.2148	18.79	.7173	.2179	19.90	.7257	.2521	18.24	.7042	.3427	17.00	.6605	.3382	18.15	.7138	.3556	19.28	.6557	.3613	17.86	.5871	.3571	19.15	.6654	.3877
RADIOv2	19.66	.7454	.2121	18.85	.7137	.2215	19.83	.7139	.3048	18.27	.7092	.3296	17.04	.6582	.3400	18.00	.7159	.3687	19.50	.6646	.3372	17.76	.5826	.3580	19.43	.6698	.4034
MASt3R	19.74	.7477	.2061	18.85	.7169	.2181	19.84	.7269	.2588	18.30	.7102	.3347	17.04	.6602	.3373	18.05	.7161	.3538	19.65	.6594	.3459	17.71	.5968	.3369	19.60	.6691	.3882
DUNE	19.69	.7499	.2041	18.75	.7147	.2200	19.70	.7261	.2689	18.33	.7088	.3337	17.02	.6595	.3392	18.18	.7185	.3571	19.51	.6665	.3445	17.81	.5835	.3569	19.56	.6728	.3894
	MipNeRF 360									MVImgNet									Tanks and Temples								
	Geometry			Texture			All			Geometry			Texture			All			Geometry			Texture			All		
Feature	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DINO-v2	21.15	.5154	.3794	19.60	.4746	.3625	21.12	.5136	.4533	19.44	.5973	.3152	16.84	.5362	.3323	19.41	.5956	.3651	18.29	.6334	.3818	18.18	.6368	.3202	18.82	.6503	.3976
RADIOv2	21.21	.5341	.3438	19.71	.4760	.3656	21.21	.5228	.4930	19.55	.6121	.2934	16.97	.5327	.3348	19.57	.5952	.3940	19.43	.6695	.3422	18.13	.6311	.3220	19.07	.6609	.4067
MASt3R	21.27	.5272	.3568	19.55	.4722	.3633	21.26	.5217	.4572	19.50	.6055	.2971	16.92	.5354	.3323	19.53	.5998	.3654	19.11	.6542	.3596	18.02	.6385	.3094	18.90	.6569	.3930
DUNE	21.38	.5340	.3527	19.72	.4791	.3636	21.27	.5254	.4609	19.59	.6115	.2912	16.93	.5346	.3342	19.48	.5980	.3685	19.36	.6621	.3546	18.11	.6300	.3219	19.02	.6592	.3951

Table 7. **Per-dataset results of Novel View Synthesis metrics in the Feat2GS benchmark [8]**. Note that models vary in size: RADIOv2 is a ViT-H, MASt3R a ViT-L and DINO-v2 and DUNE a ViT-B.

[3] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Áron Monszpart, Victor Adrian Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *Proc. ECCV*, 2022. 1

[4] Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-HMR: Multi-person whole-body human mesh recovery in a single shot. In *Proc. ECCV*, 2024. 1, 2

[5] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proc. CVPR*, 2021. 2

[6] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proc. CVPR*, 2023. 1

[7] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *CoRR*, 2020. 1

[8] Yue Chen, Xingyu Chen, Anpei Chen, Gerard Pons-Moll, and Yuliang Xiu. Feat2GS: Probing visual foundation models with gaussian splatting. 2025. 3, 5

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. CVPR*, 2017. 3

[10] Afshin Dehghan, Gilad Baruch, Zhuoyuan Chen, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *NeurIPS Datasets and Benchmarks*, 2021. 1

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009. 1

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 3

[13] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *Proc. ICCV*, 2021. 4

[14] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graphics*, 2017. 3

[15] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *Proc. ECCV*, 2024. 2, 3, 4

[16] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *Proc. CVPR*, 2018. 1

[17] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3D whole-body mesh recovery with component aware transformer. In *Proc. CVPR*, 2023. 1

[18] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *Proc. CVPR*, 2024. 1

[19] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proc. ICCV*, 2019. 1

[20] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 2

[21] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proc. CVPR*, 2021. 1

[22] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. AM-RADIO: Agglomerative vision foundation model reduce all domains into one. In *Proc. CVPR*, 2024. 2

[23] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotný. Common objects in 3D: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proc. ICCV*, 2021. 1, 3, 4

[24] Mert Bülent Sarıyıldız, Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. UNIC: Universal classification models via multi-teacher distillation. In *Proc. ECCV*, 2024. 1, 2

[25] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proc. CVPR*, 2017. 3

[26] Philipp Schröppel, Jan Bechtold, Artemij Amiranashvili, and Thomas Brox. A benchmark and a baseline for robust multi-view depth estimation. In *Proc. 3DV*, 2022. 3

[27] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. In *Proc. ICLR*, 2020. 4

[28] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. SMD-Nets: Stereo mixture density networks. In *Proc. CVPR*, 2021. 1

[29] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proc. ICCV*, 2023. 3

[30] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *Proc. CVPR*, 2024. 3, 4

[31] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. 2020. 1

[32] Frederik Warburg, Søren Hauberg, Manuel López-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proc. CVPR*, 2020. 1

[33] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 – A large-scale benchmark for instance-level recognition and retrieval. In *Proc. CVPR*, 2020. 1

[34] Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *Proc. CVPR*, 2020. 1

[35] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In *Proc. CVPR*, 2020. 1

[36] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proc. ICCV*, 2023. 1

[37] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2D Feature Representations by 3D-Aware Fine-Tuning. In *Proc. ECCV*, 2024. 4

[38] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *Proc. SIGGRAPH*, 2018. 3, 4

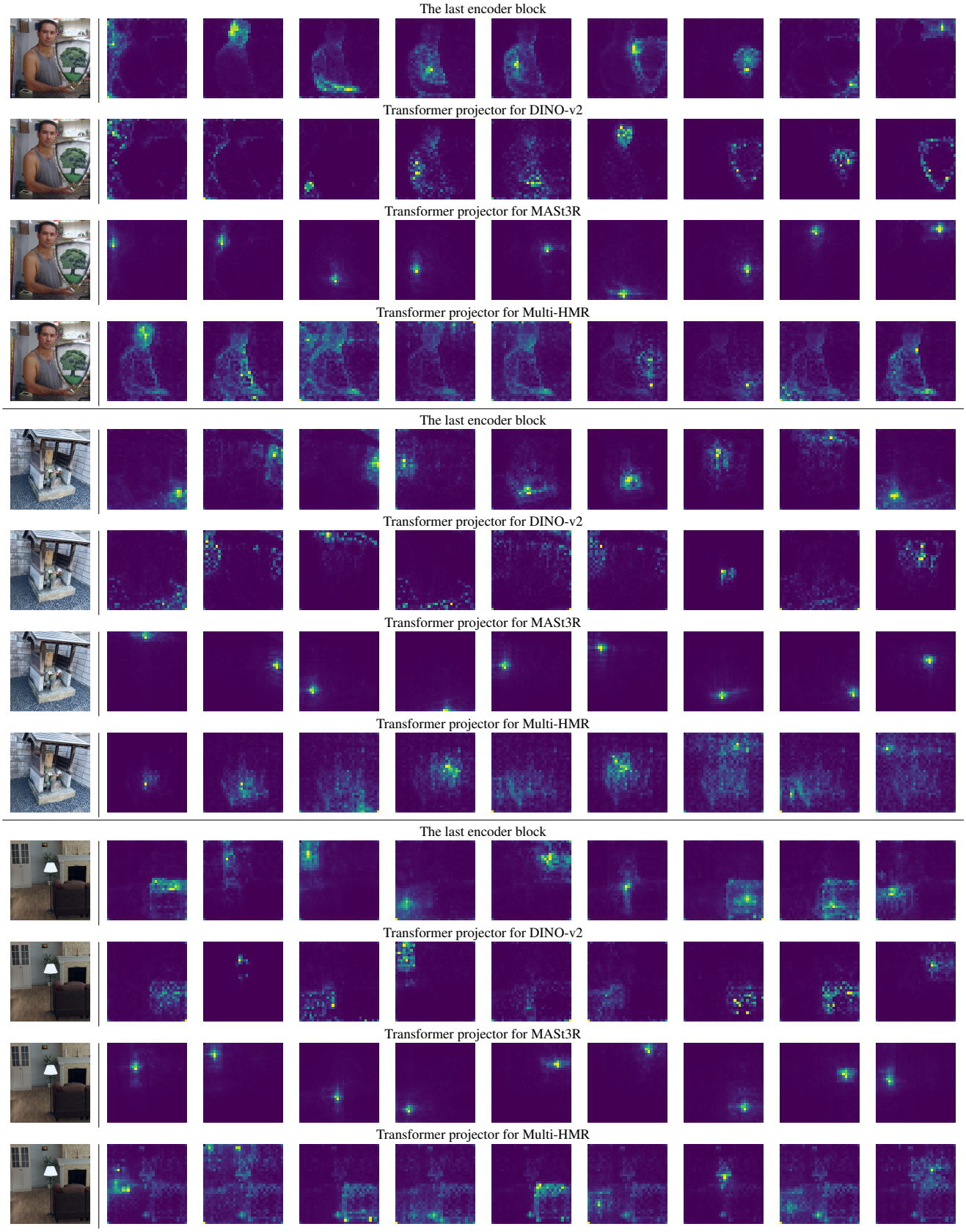

Figure 2. **Visualization of attention maps.** Given an image of resolution 448×448 (1st column), we extract using our student model the attention probability map (of size 32×32) for each patch from either the last encoder layer or the Transformer projector for each teacher. Then, we flatten each map and run k-medoids clustering with $k = 9$, and visualize centroids.

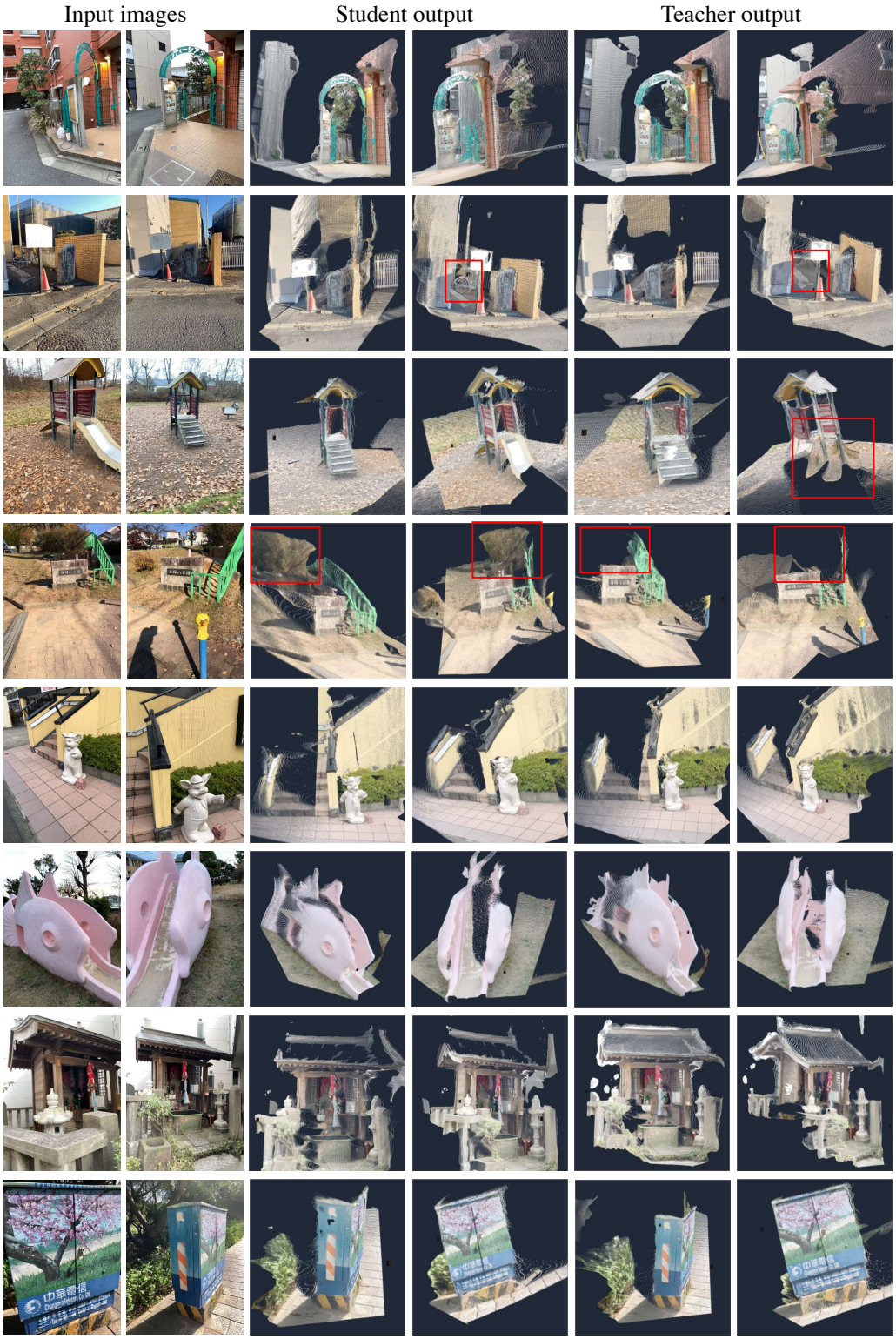

Figure 3. **Qualitative results for the MASt3R teacher and our student.** Each row presents two input images and corresponding 3D reconstructions. Images were sampled from the Niantic dataset. With a red square, we highlight regions where our student seems to outperform the teacher.

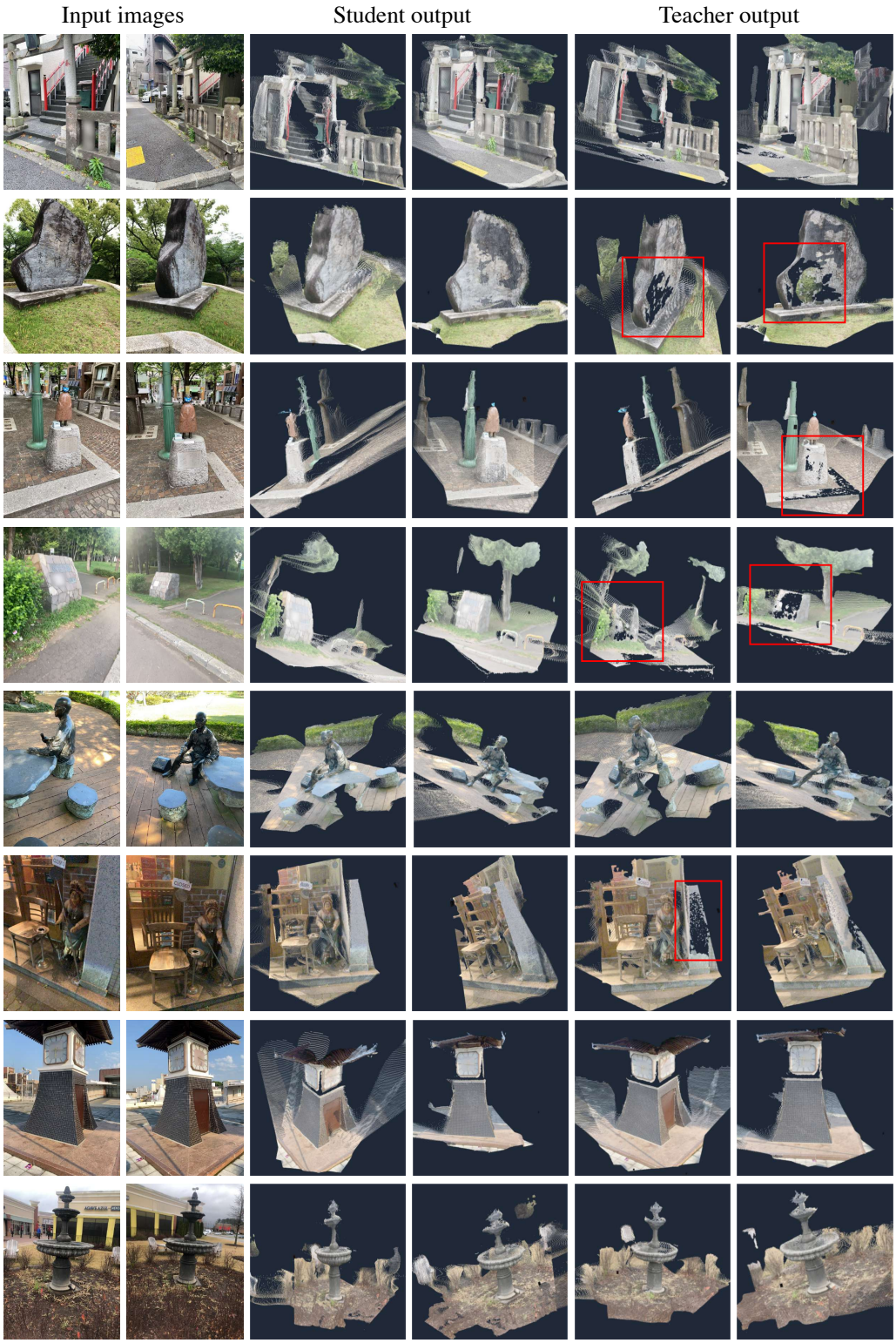

Figure 4. **Qualitative results for the MASt3R teacher and our student.** Each row presents two input images and corresponding 3D reconstructions. Images were sampled from the Niantic dataset. With a red square, we highlight regions where our student seems to outperform the teacher.

Figure 5. **Scene reconstructions from longer input sequences for the MASt3R teacher and our student.** With a red square, we highlight regions where our student seems to outperform the teacher.

Figure 6. **Qualitative Human Mesh Recovery results.** Qualitative comparison of outputs between teacher and student. Images sampled in the validation set and sorted by alphabetical order. The two models produce outputs of comparable visual quality.

Figure 7. **Qualitative Human Mesh Recovery results (continued).** Qualitative comparison of outputs between teacher and student. Images sampled in the validation set and sorted by alphabetical order. The two models produce outputs of comparable visual quality.

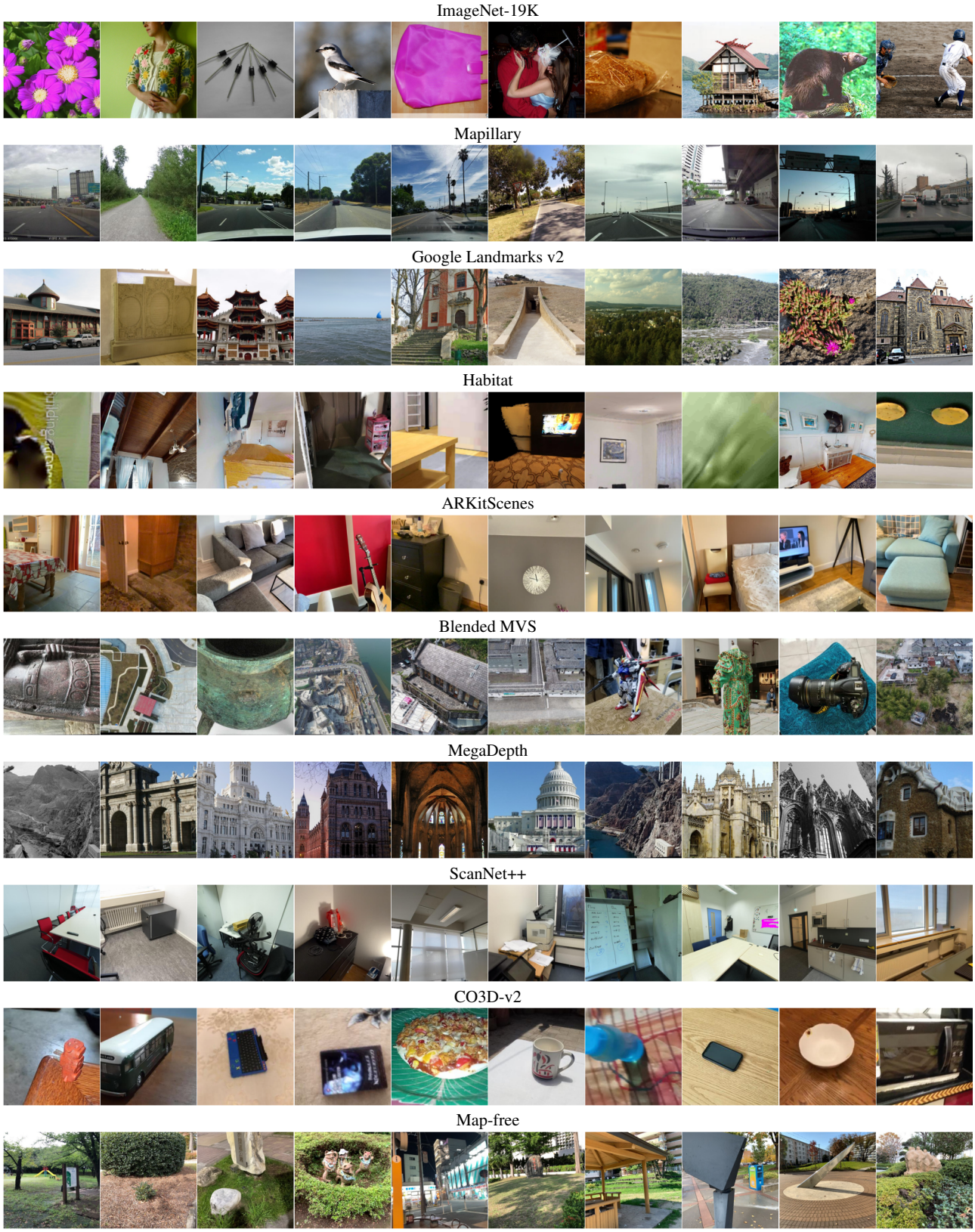

Figure 8. **Visualization of random samples from datasets.** We visualize 10 randomly sampled images from each dataset listed in Tab. 1. See Fig. 9 for the visualization of the remaining datasets.

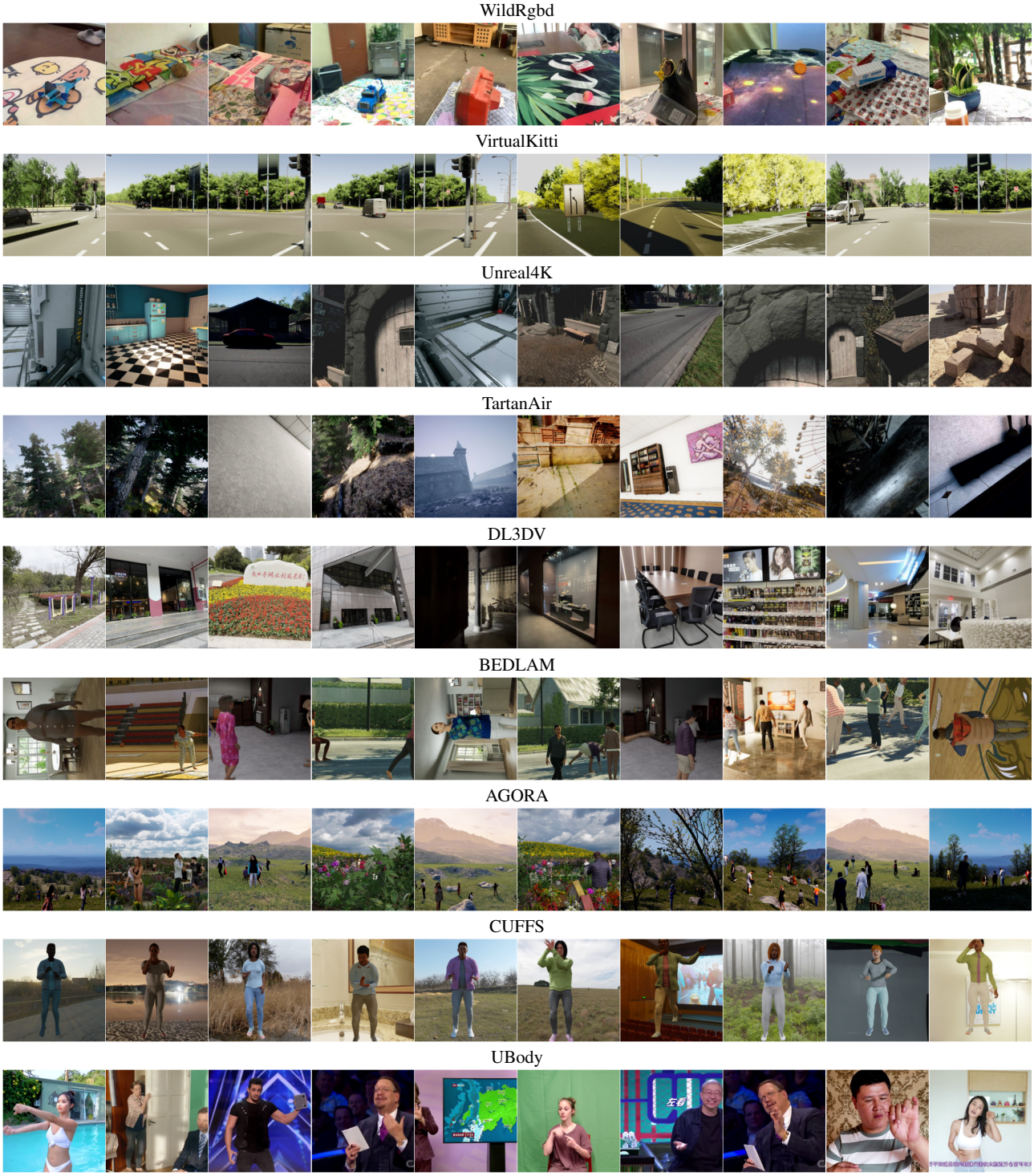

Figure 9. **Visualization of random samples from datasets (continuation of Fig. 8).** We visualize 10 randomly sampled images from each dataset listed in Tab. 1. See Fig. 9 for the visualization of the remaining datasets.