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

Abstract

*Recent multi-teacher distillation methods have unified the encoders of multiple foundation models into a single encoder, achieving competitive performance on core vision tasks like classification, segmentation, and depth estimation. This led us to ask: Could similar success be achieved when the pool of teachers also includes vision models specialized in diverse tasks across both 2D and 3D perception? In this paper, we define and investigate the problem of heterogeneous teacher distillation, or co-distillation—a challenging multi-teacher distillation scenario where teacher models vary significantly in both (a) their design objectives and (b) the data they were trained on. We explore data-sharing strategies and teacher-specific encoding, and introduce **DUNE**, a single encoder excelling in 2D vision, 3D understanding, and 3D human perception. Our model achieves performance comparable to that of its larger teachers, sometimes even outperforming them, on their respective tasks. Notably, DUNE surpasses MASt3R in Map-free Visual Relocalization with a much smaller encoder.*

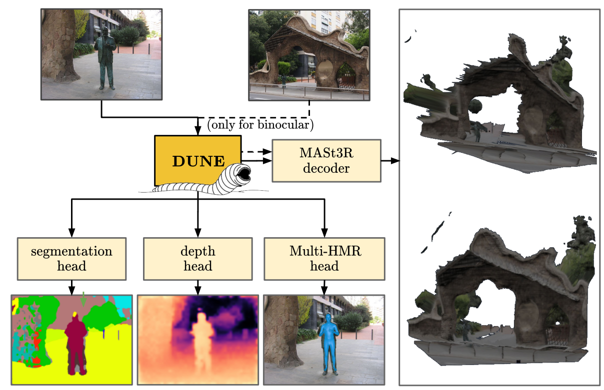


Figure 1. **DUNE** is a universal encoder for 2D and 3D tasks distilled from heterogeneous teachers. It enables multi-task inference with a single encoder. Teachers are DINO-v2 [33], MASt3R [25], and Multi-HMR [3] (see Fig. 3 for distillation details).

1. Introduction

Computer vision has seen the rise of foundation models [6] such as DINO-v2 [33] or SAM [51]. Trained on massive web-crawled datasets, they provide representations useful for multiple downstream tasks. A recent body of work including AM-RADIO [37], Theia [44] or UNIC [42] has successfully unified the encoders of several of these foundation models into a single compact encoder via *multi-teacher distillation*. Although an impressive feat, this line of research has so far only distilled models all trained on data of a similar nature, composed of "generic" web-crawled images. In fact, in all recent works [30, 37, 42], distilling using the ImageNet-1K [14] dataset suffices to match teacher performance for tasks like classification, segmentation and monocular depth.

In this work we study *heterogeneous teacher distillation*, or *co-distillation*.[1] We consider a set of teacher models as *heterogeneous* if they vary significantly with respect to both (a) the *tasks* these teachers are trained for and (b) the visual domains of their training data. We seek to answer the following question: *Can we train a single encoder that excels at widely diverse tasks by distilling from state-of-the-art heterogeneous models?*

To that end, we distinguish between task-agnostic teachers, aimed at producing representations that generalize across several tasks, and specialized teachers that achieve state-of-the-art performance on one specific task. Given the heterogeneity in training data across teachers, we examine which data should be used for distillation. Finally, we study the use of projectors, modules used in multi-teacher distillation to ensure compatibility across teachers and capture teacher-specific information. More specifically we question the projector design when distilling heterogeneous teachers with abundant specialized information.

[1] In chemistry, co-distillation refers to distillation performed on mixtures in which the compounds are not miscible.

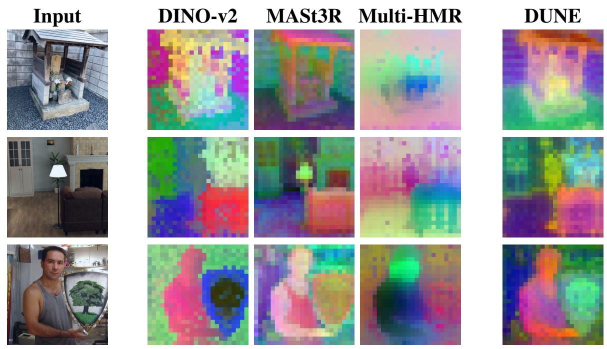


Figure 2. **PCA visualization of encoder outputs.** Given an image, we extract patch embeddings from the encoders of the teacher models and our student, and reduce their dimension to 3 via PCA.

To evaluate our framework, we choose three state-of-the-art models. Two are highly task-specific: MASt3R [25] solves 3D scene reconstruction and matching, while Multi-HMR [3] solves 3D human perception. As a third teacher, we add DINO-v2 [33], a popular visual foundation model that generalizes to various visual downstream tasks including semantic segmentation or monocular depth estimation. This leads to **DUNE**, obtained by **D**istilling a **UN**iversal **E**ncoder from heterogeneous 2D and 3D Teachers.

The selected teachers vary with respect to both the tasks they are tackling and the datasets they were trained on. Their heterogeneity is clearly visible when inspecting the patch features they produce. In Fig. 2, we plot the top three components (obtained via principal component analysis) for the features of each teacher, as well as the features of our co-distilled encoder, DUNE. We observe that each teacher has distinct and complementary features. We also see that our model captures properties present across all three teachers.

Our study yields several insights into the selection of distillation data and the impact of projector design for co-distillation. More importantly, it results in a powerful universal encoder that matches top models in binocular 3D reconstruction, 3D human perception, and classical 2D vision tasks while retaining much of DINO-v2's generalization performance. Additionally, it sets a new state-of-the-art on the Map-free Visual Relocalization Challenge[2] using a ViT-Base encoder.

Contributions. (a) We define the problem of heterogeneous teacher distillation, where a single model is distilled from teacher models that vary significantly in training tasks and image domains. (b) We investigate suitable distillation strategies for this setting, focusing on distillation data and projector design. (c) We introduce **DUNE**, a strong ViT-Base encoder capable of excelling in 3D scene understanding, 3D human perception, and 2D vision tasks.

[2] https://research.nianticlabs.com/mapfree-reloc-benchmark/leaderboard

2. Related work

Combining multiple models. Various approaches to combining models have been explored. One method involves extracting features from each model and either concatenating or fusing them for downstream tasks, a strategy used in early works [46] as well as more recently in [8, 24]. While effective, this approach is often impractical due to memory and compute requirements. Other studies focus on merging models together [12, 15, 23, 41, 45, 52], typically under the assumption that models have the same architecture and size.

Multi-teacher distillation is another way to combine multiple models into one. It replaces the encoder of each teacher by a single student encoder. This encoder is obtained by distilling the outputs of the teacher encoders on well-chosen data. This offers great flexibility: there are a-priori no constraints on the student architecture or size. Recent examples include generic methods like AM-RADIO [19, 37], UNIC [42], PHI-S [36], UNIT [65], robotics-specific models such as Theia [44], and methods using sequential distillation such as in [39]. By distilling from three or more teachers, these teacher-agnostic strategies sometimes even yield students that outperform their teachers on some tasks [37, 42]. However, all of these methods focus on distilling teachers of the same nature, foundation models trained on massive web-crawled image collections with self- or weak supervision. Common choices include DINO-v2 [33], CLIP [35], and SAM [51]. The students obtained are then applied to common benchmark tasks like classification, segmentation, or monocular depth.

In this paper, we explore distillation from heterogeneous teachers, *i.e.* trained on distinct domains and solving diverse tasks. Specifically, we distill from MASt3R [25] for binocular 3D tasks, Multi-HMR [3] for human mesh recovery, and DINO-v2 [33] for 2D tasks like monocular depth and semantic segmentation. This results in an encoder that excels across many of these tasks out of the box.

Training using heterogeneous data domains. All aforementioned multi-teacher distillation methods have another thing in common: they distill only from generic data suitable to all teachers: DataComb-1B [18] for AM-RADIO, ImageNet-1K [14] for UNIC [42] and Theia [44]. Heterogeneous data has recently been used for distillation of different datasets for classification [22, 62] or domain adaptation [48]. To the best of our knowledge, no existing work has looked into distillation using data that contains natural images as well as synthetic data from 3D engines, CAD models, simulators, and rendered from structure from-motion reconstructions. Our teachers' training data even vary in terms of content, from small groups of people for human mesh recovery to empty indoor rooms and outdoor buildings for 3D. We hypothesize that leveraging such data is necessary to accurately distill highly specialized teacher models such as MASt3R [25] and Multi-HMR [3], and attempt to do so in this paper.

Relation to multi-task learning. As in multi-task training, our heterogeneous teacher distillation tackles specialized tasks such as segmentation, human pose estimation, and 3D reconstruction with a single encoder. However, it also preserves the generic nature of some teachers, ensuring strong generalization. Unlike multi-task training, distillation relies on teacher outputs instead of ground-truth labels and does not require access to the original training data.

Relation to 3D-to-2D distillation. Prior work [20, 63] has used 3D-aware models to improve results on 2D tasks. While these methods are related, we show that DUNE excels at *both* 2D and 3D.

3. Problem definition and challenges

First, we describe the problem of **heterogeneous teacher distillation** or **co-distillation**, a challenging multi-teacher distillation setup where teacher models significantly vary with respect to (a) the goals behind their design, and (b) the data they were trained on. We would like to jointly distill from a set of teachers that satisfy the following properties:

1. **The teachers cover an heterogeneous set of tasks.** We want to jointly distill teachers that vary in the *design objectives* they are trained on. This includes *task-agnostic* teachers (models with strong generalization properties, typically self-supervised and trained on pretext tasks) together with *specialized* models tailored to specific tasks.
2. **Their individual training sets consist of heterogeneous data.** We want to jointly distill teachers trained on huge generic image datasets crawled from the web, together with teachers trained on highly curated and potentially carefully annotated datasets composed of natural or synthetic images.

This leads to the question driving our research: *Can we distill from such an heterogeneous set of teachers and get a universal visual encoder that retain strong generalization abilities and at the same time* excels at multiple diverse tasks?

3.1. Task-agnostic *vs*. specialized teachers

Following our problem definition, we differentiate between teacher models trained on proxy tasks that capture inherent image priors and those specializing in a specific task or set of tasks. While the former might consist solely of a visual encoder and be referred to as a foundation model, the latter typically include task-specific decoder heads and are trained using task-specific data and supervision. Our goal is not only to match the performance of specialized models, but also retain the generalization capability of the representations learned from the task-agnostic teachers.

Task-agnostic teachers refers to models trained on proxy tasks such as self-supervised objectives like context prediction or photometric invariance. They aim to capture broadly used visual priors that are useful on a wide range of tasks. This includes DINO-v2 [33] that we consider as one of the teachers for DUNE.

Such models are typically evaluated on tasks for which the encoder representations can be used directly, such as k-NN or zero-shot [35] classification, or by training linear classifiers for various classification tasks at the image or pixel level (*e.g*., semantic segmentation or monocular depth estimation using a linear head). Notably, these models are recognized for the *generalization* strength of their representations. Those have been shown to be beneficial to a wide range of tasks [57, 58]. However, on their own, they usually underperform compared to specialized models trained with supervised or privileged information.

Specialized teachers [3] focus on specific perception domains, such as 3D for MASt3R [25] or human pose understanding for Multi-HMR [3]. They are typically trained with weak or strong supervision, and, although all use ViT encoders, they leverage a domain-specific parameterization and require varying levels and types of annotation. The encoders of specialized teachers may differ in the *nature* of the information they encode (SMPL body model [29] parameters *vs*. dense matches) and *how* they encode it (*e.g*., Multi-HMR captures the full 3D human pose in a single patch token).

Generalization *vs*. specificity trade-off. The distinction above highlights the trade-off between generalization to novel tasks, afforded by task-agnostic teachers, and performance on specialized tasks coming from specialized teachers. By incorporating this distinction into our formulation, we can interpret the proposed distillation setup in two ways: (a) as a means to enhance the performance of specialized teachers on certain novel tasks by leveraging task-agnostic models, or (b) as a way to improve the performance of self-supervised foundation models on the set of specialized tasks.

3.2. Distillation using heterogeneous data

The diversity of teacher training domains has not been a significant concern for existing multi-teacher distillation approaches [30, 37, 42]. Foundation models like DINO-v2 [33], SAM [51], and CLIP [21, 35] are all trained on data of a similar nature, *i.e*. "generic" datasets like LAION [43] or DataComp1B [18], and recent works [30, 37, 42] show that even using ImageNet-1K [14] is enough to match teacher-level performance with a unified encoder.

One of the main challenges of co-distillation is the teachers' training data that spans *multiple visual domains*. Alongside visual encoders trained on natural images, *i.e*. DINO-v2 [33], we aim to jointly distill encoders trained on diverse types of data (*e.g*. synthetic data from 3D engines,

[3] We use *specialized* rather than *task-specialized* to account for teachers solving multiple tasks within an area (human understanding, 3D vision).

CAD models, simulators, *etc.*) and with diverse content (*e.g.* data focused on small groups of people in the case of Multi-HMR [3], or empty indoor rooms and outdoor buildings for MASt3R [25]). When jointly distilling teachers trained on such heterogeneous data, the choice of data for distillation becomes less straightforward: *Do we need to incorporate data from all teacher domains, or is generic data sufficient?*

4. Framework for co-distillation

In this section, we introduce our co-distillation framework for heterogeneous teacher distillation, as shown in Fig. 3. We begin with the fundamentals of multi-teacher distillation, then discuss teacher-specific projectors and our choices towards an efficient inference.

4.1. Background on multi-teacher distillation

We focus on visual encoders based on the ViT [16] architecture. These models take an image $x \in \mathcal{I}$ as input and produce a set $Z \in \mathcal{Z}$ of feature vectors, where $\mathcal{Z} \in \mathbb{R}^{(HW+1)\times d}$. This feature set includes HW features for the $H \times W$ patches, along with an optional global feature corresponding to a CLS token. We assume each feature vector has a dimensionality of d. These feature vectors are typically used as an input to one or more decoder heads to tackle specific computer vision tasks.

Let $\mathcal{T} = \{\mathcal{T}_1, \dots, \mathcal{T}_N\}$ represent the set of N teachers we want to distill, each parametrized by an encoder $t_i(x)$. Our objective is to learn the parameters of a student model f that produces outputs closely aligned with those of *all* teachers simultaneously. We train this student encoder by applying a distillation loss on both the global and patch token features. Let $f(x) : \mathcal{I} \to \mathcal{Z}$ denote the student's encoder, and h_i represent a teacher-specific projector for each teacher $\mathcal{T}_i, i = 1,..,N$. We minimize the cosine-similarity and smooth-ℓ_1 losses combined among all teachers:

$$\mathcal{L}_{\text{distil}} = \sum_{i=1}^{N} \mathcal{L}_{cos}\big(f_i(x), t_i(x)\big) + \mathcal{L}_{s\ell_1}\big(f_i(x), t_i(x)\big), \quad (1)$$

where $f_i = h_i(f(x))$, and $\mathcal{L}_{cos}$ and $\mathcal{L}_{s\ell 1}$ denote the cosine and smooth-ℓ_1 losses, respectively, as defined in [42].

The process described above imposes no restriction on the types of teachers used and, with minor modifications, forms the foundation of several recent works [37, 42, 44].

4.2. Teacher-specific projector design

Existing multi-teacher distillation approaches [30, 37, 42] address the issue of teacher-specific (or "complementary") information derived from the different training objectives of teachers using teacher-specific modules also known as *projectors*. In our problem formulation (Sec. 3), such modules become even more important: The patch representations from the encoder and their interactions can differ significantly between teachers. The Multi-HMR model [3], for example, requires pose information for the whole human body to be captured in the representation of the patch where the head and nose are detected, since the corresponding decoder only takes such patches as input.

A simple projector design is a two-layer MLP appended to the top of the encoder for each teacher [37]. More complex designs have been explored [30, 42, 44]. UNIC [42] attaches multiple additional MLPs to intermediate layers of the student. Its "ladder of projectors" (LP) improves information flow to the teacher-specific parameters and was shown to help improve distillation across teachers and tasks.

For all the cases mentioned above, projectors operate *per-patch*, *i.e.* teacher-specific parameters cannot explicitly capture patch feature interactions. This requires inter-patch interactions specific to each teacher to be embedded in the attention layers of the shared encoder. In other words, it is the shared encoder that has to model all patch interactions relevant for any teacher. This motivated us to introduce attention-based projectors that we define next.

Transformer projectors. For projectors to model interactions across patches, a sensible and efficient design would be an attention-based projector composed of a single transformer [50] block:

$$a = f(x) + \text{SA}(\text{LN}(f(x))), \quad (2)$$

$$m = a + \text{MLP}(\text{LN}(a)), \quad (3)$$

$$h = \text{Linear}(m), \quad (4)$$

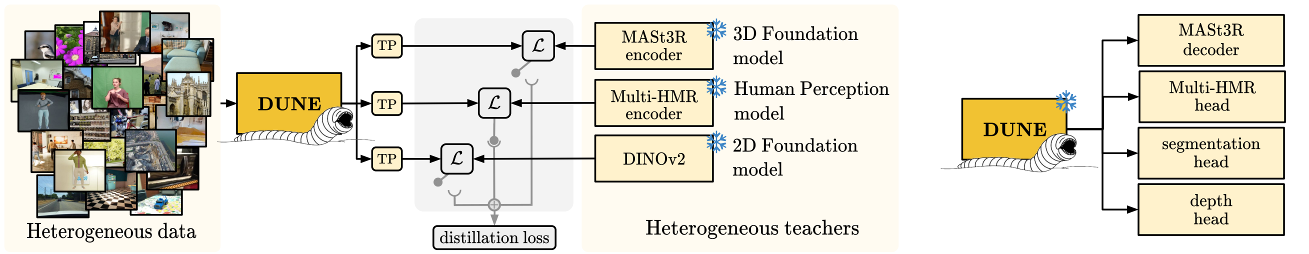
where LN denotes layer normalization, SA a multi-head self-attention layer, MLP a two-layer perceptron and Linear a fully-connected layer. We refer to this projector design as a **transformer projector** or **TP**. Sec. 5 compares it with more standard projectors.

4.3. Co-distillation with heterogeneous data

In co-distillation, the optimal distribution of data to distill from is not trivial: Data associated with a specific teacher can be irrelevant or even harmful to others. It therefore makes sense to control which data gets forwarded to each teacher-specific projector.

Let $\mathcal{D}_i$ denote the data associated with teacher $\mathcal{T}_i \in \mathcal{T}$ with $i = 1..N$. We assume that $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset$ for all specialized teachers i, j, *i.e.* the specialized teachers are all trained on different datasets. We also assume that all task-agnostic teachers are trained with the same generic data $\mathcal{D}_g$. Let h_i denote the projector associated with teacher i. We explore three simple ways of sharing datasets across teachers:

- **No data sharing**: Projector h_i receives only $\mathcal{D}_i$, *i.e.* the data associated with its teacher $\mathcal{T}_i$.
- **Full data sharing**: Projector h_i receives *all* data, *i.e.* $\cup\mathcal{D}_i, i = 1..N$, as well as generic data $\mathcal{D}_g$.
- **Generic data sharing**: Projector h_i receives only $\mathcal{D}_i$ (the



(a) Distilling **DUNE**, a **universal encoder** from heterogeneous 2D and 3D teachers. (b) Fine-tuning task heads independently.

Figure 3. **Overview of the DUNE encoder training process.** (a) DUNE is trained via distillation from *heterogeneous* teachers across 2D vision, 3D vision, and 3D human perception, leveraging diverse data from multiple visual domains. We use *teacher dropping* regularization from [42]. (b) Task-specific heads are then fine-tuned independently for each task, with the DUNE encoder kept frozen.

data associated with $\mathcal{T}_i$) and generic data $\mathcal{D}_g$.

In the next section, we evaluate these different ways of sharing data across teachers and study their impact on area-specific tasks as well as representation generalization tasks.

4.4. Fine-tuning task heads for efficient inference

Most related works [19, 30, 37, 44] require the teacher-specific projectors learned during training to be used during inference. This allows for a plug-and-play reuse of task-specific decoders, yet this results in more parameters, not only for the student encoder itself but also for all the teacher-specific projectors. The number of those additional parameters scales linearly with the number of teachers.

Projectors would become irrelevant if the decoder modules were jointly fine-tuned for the tasks to solve, once the student is trained. Such an approach offers several advantages: It introduces no additional modules during inference, it keeps the encoder size and memory constant regardless of the number of teachers, and, most importantly, it does not impact inference time. For DUNE, we opt for that second option and fine-tune the different heads and decoders, a one-time cost that enables a more efficient inference.

5. Experiments

In this section, we experimentally evaluate the co-distillation framework from Sec. 4. After introducing the evaluation protocol (Sec. 5.1), we experimentally validate the projector design, the use of heterogeneous data and data sharing across projectors, and we present results on five very different tasks (Sec. 5.2). We then present some feature and loss analyses, as well as some qualitative results comparing the teacher and student model outputs (Sec. 5.3).

5.1. Evaluation protocol

Teachers. We select a representative set of heterogeneous teachers for our experimental validation. The first teacher is DINO-v2 [33] with registers [11], a self-supervised task-agnostic model whose strong representations have been used in many computer vision tasks, and a common teacher choice in multi-teacher distillation [37, 42]. Then, we select two state-of-the-art domain-specialized models: Multi-HMR [3] a model for human mesh recovery, the winner of the Robin Challenge at CVPR'24,[4] and MASt3R [25] a 3D foundation model, the winner of the Map-free Visual Relocalization at ECCV'24.[5] The selected teachers are designed to span across both 2D and 3D tasks, with the two 3D-oriented teachers differing in the *nature* of the 3D information they encode (SMPL parameters *vs*. dense matches) and *how* they encode it (*e.g.* Multi-HMR captures the full 3D human pose in a single patch token). We use the publicly available ViT-Large models for all teachers.

Datasets. We use 19 publicly available datasets from the training sets of the three teachers as distillation data, leading to around 20.7M images in total: ImageNet-19K [14] (2021 release [59]), Mapillary [55] and Google Landmarks v2 [56] from DINO-v2, AGORA [34], BEDLAM [5], UBody [17, 27] and CUFFS [3] from Multi-HMR; Habitat [31], ARKitScenes [13], Blended MVS [60], MegaDepth [26], ScanNet++ [61], CO3D-v2 [38], Map-free [2], WildRgb [1], VirtualKitti [7], Unreal4K [49], TartanAir [54] and DL3DV [28] from MASt3R. We only use the images of those datasets and discard all annotations.

Implementation details. During distillation, our student is composed of a ViT-Base encoder and three projector heads, one for each teacher. In our initial experiments, we used the Ladder of Projectors (LP) design from UNIC [42], which was shown beneficial for dense prediction tasks. We also evaluate the Transformer Projector we introduced in Sec. 4.2. Unless otherwise specified, *projectors are discarded* after distillation and are not used for evaluations or inference. By default, we fix the compute budget for all the distillation variants to process $100 \times 1,281,167$ images, which amounts to 100-epoch training on ImageNet-1K [40]. We follow the data augmentation and optimization practices of [42], and also use their teacher dropping regularization.

[4] https://rhobin-challenge.github.io/

[5] https://nianticlabs.github.io/map-free-workshop/2024/

Distil. Data	Proj. Design	ADE20K (mIoU ↑)	NYUd (RMSE ↓)	MapFree (AUC ↑)	BEDLAM (PA-PVE ↓)
IN-19K	LP	42.4	0.446	91.4	83.9
IN-19K	TP	**44.9**	0.433	93.6	73.5
All	SP	42.3	0.413	92.2	73.1
All	LP	44.7	0.384	91.5	78.2
All	TP	**44.9**	**0.377**	**93.7**	**68.3**

Table 1. **Distillation data and projector design**. We distill models using either ImageNet-19K as a generic dataset or all the 19 teacher datasets combined with full data sharing (Sec. 4.3), employing either a Simple Projector (SP) [37], a Ladder of Projectors (LP) [42], or the Transformer Projector (TP) presented in Sec. 4.2.

Unless otherwise stated, we distill our encoders at image resolution 336×336. We further distill some of our models at resolution 448×448 for a couple of additional epochs.

Evaluation tasks and metrics. We evaluate our encoder on the tasks that our selected teachers excel at. For domain-specialized teachers, we choose tasks for which they are the state of the art: multi-person human mesh recovery for Multi-HMR and map-free visual relocalization for MASt3R. For Multi-HMR, we report results on the BEDLAM validation set, using F1-Score to evaluate detection performance and PA-PVE to measure mesh reconstruction errors. For MASt3R, we report the Area under the Curve (AUC) for samples with Virtual Correspondence Reprojection Error (VCRE) below 90px on the validation set of the Map-free Visual Relocalization dataset [2]. We also provide results for multi-view depth estimation and multi-view camera pose regression in the supplementary material. We evaluate the generalization performance of our encoder on semantic segmentation with ADE20K [64] (mIoU) and depth estimation on NYUdv2 [32] (RMSE), following related work [37, 42]. We also report comparisons to 3D-to-2D distillation methods [20, 63] as well as segmentation results on Cityscapes [9], NYUdv2 [32] and ScanNet [10] in the supplementary material. Further details on the evaluation protocol are provided there too.

Fine-tuning task heads. For specialized tasks, we attach the corresponding teacher's decoder module to the frozen encoder and fine-tune the decoder for that task. For semantic segmentation and depth estimation, we train a linear head from scratch, appending it to the frozen encoder, as in [33]. For segmentation, re-using the frozen transformer projector for the DINO-v2 teacher before training the linear layer significantly improved performance. For this case only, we report results *using* the projector, similar to [37]. We compare results with and without projectors in the supplementary material.

5.2. Results

Is "generic" data enough? We start our experiments by checking whether or not a large, generic dataset like ImageNet is enough for distilling heterogeneous teachers. We train two student models using a sparsely-connected Ladder of Projectors (LP), *i.e.* a multi-layer Perceptron attached after every 3 encoder blocks. The first model only uses ImageNet-19K as distillation data whereas the second one uses all distillation datasets mentioned above. From the results presented in Tab. 1, we observe that using additional specific data improves the performance of the students on all the tasks by a decent margin.

Data Sharing	ADE20K (mIoU ↑)	NYUd (RMSE ↓)	MapFree (AUC ↑)	BEDLAM (PA-PVE ↓)
No data sharing	41.6	0.426	93.2	68.7
Generic data sharing	40.1	0.416	92.7	71.7
Full data sharing	**44.9**	**0.377**	**93.7**	**68.3**

Table 2. **Data sharing among teachers.** We train three student models on all 19 datasets, using the data-sharing strategies outlined in Sec. 4.3: *No data sharing*: Teachers do not share any data. *Generic data sharing*: ImageNet-19K is shared across all teachers. *Full data sharing*: All images are shared among all teachers.

Transformer Projectors. Looking at Fig. 2, we notice that feature similarity patterns significantly vary from one teacher's encoder to another. Visualizing the attention maps from the final layer encoder of each teacher (provided in the supplementary material) also shows clear differences across models: MASt3R produces highly localized attentions while the DINO-v2 attention span is much larger. Attention maps for Multi-HMR seem more focused to the persons' head, ignoring the remaining content in many cases.

We argue that capturing feature similarities of such different spatial extents would be easier using the Transformer projector (TP) presented in Sec. 4.2. To test this hypothesis, we replace the LP-based projectors attached to multiple layers of the student encoder with a single TP projector after the last layer of the encoder. We compare Tab. 1 the LP and TP designs, as well as the simple MLP projector (SP) used in [37] (row 3). We observe that TP outperforms both LP and SP across all tasks.

Sharing data across teacher projectors. In Sec. 4.3, we have presented three different ways of sharing data across teachers. In Tab. 2, we evaluate all three and see that sharing all data among teachers generally yields the best performance. It suggests that the domain gap between the datasets might not be an issue and teachers still produce useful information for out-of-domain images. Interestingly, sharing only generic data is the best for semantic segmentation. This suggests that semantic information is better preserved by the encoder when ImageNet-19K is shared by MASt3R and Multi-HMR.

Comparing to state-of-the-art multi-teacher distillation. In the middle part of Tab. 3, we compare our distilled encoder to the strongest comparable ViT-Base encoders available, *i.e.* DINO-v2 [33] and AM-RADIO-v2.5 [19]. For

Model	Encoder Arch.	Training Data	Training Res.	ADE20k (mIoU ↑)	NYUd (RMSE ↓)	BEDLAM (F1-score ↑)	BEDLAM (PA-PVE ↓)	MapFree (AUC ↑)
Teacher models								
DINO-v2	ViT-Large	LVD-142M	518	47.7	0.384	-	-	-
Multi-HMR	ViT-Large	HMR-500K	672	-	-	95	36.9	-
MASt3R	ViT-Large	MASt3R-1.7M	512	-	-	-	-	91.2
State-of-the-art ViT-Base encoders								
DINO-v2	ViT-Base	LVD-142M	518	47.3	0.399	86	76.5	89.6
AM-RADIO-v2.5	ViT-Base	DataComp-1B	512	**50.0**	0.718	89	83.2	93.1
DUNE	ViT-Base	DUNE-20.7M	336	44.9	0.377	91	68.3	93.7
DUNE	ViT-Base	DUNE-20.7M	448	45.6	**0.358**	**94**	**56.0**	**94.7**

Table 3. **Performance across 2D vision, 3D human understanding and 3D vison tasks with a *universal* encoder.** The top section shows the performance of the teacher models we use, all of size ViT-Large. The middle section compares our encoder to two state-of-the-art ViT-Base encoders: DINO-v2 and the latest AM-RADIO-v2.5 model. *DUNE-20.7M* is the heterogeneous collection of 19 public datasets we use for co-distillation. **Colored** results highlight cases where our model outperforms *the best ViT-Large teacher*.

both models, we fine-tune the decoder heads for each task using the same procedure as for our encoder. DUNE outperforms both models on all evaluations except semantic segmentation, where AM-RADIO-v2.5 achieves the best results by a significant margin. This is expected, as AM-RADIO-v2.5 is distilled from two semantically rich teachers, CLIP and OpenCLIP, along with the strong segmentation model SAM [51].

Improving Map-free Visual Relocalization. In Tab. 4, we report results from the official leaderboard of the Map-free visual relocalization dataset.[6] On this benchmark, MASt3R [25] was shown to significantly outperform the state of the art [4, 47, 53], where all recent approaches build upon a ViT-Large backbone. Surprisingly, when replacing the ViT-Large encoder of MASt3R by the frozen ViT-Base encoder of our student model, and then fine-tuning the MASt3R decoder, we obtain even better performance than MASt3R despite using a significantly smaller encoder.

5.3. Analysis and visualizations

In this section, we provide some analysis to better understand the distillation process and feature spaces learned by co-distillation. We then provide qualitative results on the tasks of the two specialized teachers.

Feature analysis. In Fig. 4, we plot the cumulative explained variance, *i.e*. the proportion of the dataset's variance that is cumulatively explained by each additional PCA component. We plot curves for three representative datasets, comparing the teacher encoder features to the corresponding student projector features. Interestingly, we observe that rather than seeing a significant change across datasets, the most consistent difference in terms of feature compactness is across teachers. The Multi-HMR teacher needs consistently fewer PCA components to explain its feature variance while DINO-v2 needs the most. We argue that this could be because the Multi-HMR model has a more specialized task and training data, while DINO-v2 has been trained with a more diverse dataset aiming to be a versatile encoder. MASt3R seems to be somewhere in the middle as a versatile encoder specialized in 3D tasks.

In the same figure, we also plot explained variance on the three datasets for the features of our learned encoder after the teacher-specific heads and observe that they follow the ranking of the corresponding teacher features. However, student representations are consistently more compact than the corresponding teachers.

Correlation of loss updates. In Fig. 5, we plot how often loss updates are correlated for three pairs of teachers, *i.e*. we look at the change in loss magnitudes after each weight update, and measure the correlation between the loss fluctuations for each teacher. If the teachers are well aligned, we expect a strong positive correlation (*e.g*., minimizing the loss for Multi-HMR also minimizes the loss for DINO-v2), while a low correlation would indicate the teacher feedback is less aligned and training might be more unstable.

The first four bars represent the correlation for two training data choices (ImageNet-19K or all 19 datasets) and two projector designs. We observe that with LP, teachers are always less aligned compared to TP, regardless of whether training is performed only on ImageNet-19K or on all datasets. This may explain LP's inferior performance, particularly on specialized tasks, as shown in Tab. 1. For TP, we additionally measure teacher alignment across the three data-sharing options presented in Sec. 4.3 (rightmost three bars). All in all, we observe that using all data for all teachers results in the highest possible correlation on all teacher pairs, something that is further reflected on performance in Tab. 2.

[6] https://research.nianticlabs.com/mapfree-reloc-benchmark

Method	Encoder	VCRE<45px		VCRE<90px		Median Reproj. Error (px)↓	Relative Pose Error		
		AUC↑	Prec.↑	AUC↑	Prec.↑		Median Error↓	Prec.↑	AUC↑
LoFTR [47]	CNN	39.7	18.2	61.8	33.5	166.8	2.31m 39.4°	26.9	9.8
DUSt3R [53]	ViT-Large	45.9	28.7	69.8	50.4	115.8	0.99m 7.1°	39.4	21.4
MicKey [4]	ViT-Large	57.2	31.2	74.8	49.3	129.5	1.59m 26.0°	28.3	12.0
MASt3R [25]	ViT-Large	81.7	63.0	93.3	79.3	48.8	**0.37m 2.2°**	74.0	54.7
DUNE	**ViT-Base**	**84.0**	**64.4**	**94.3**	**81.1**	**47.4**	0.39m 4.6°	**76.8**	**55.9**

Table 4. **Results on the map-free visual relocalization official leaderboard.** AUC and Precision (Prec.) are reported in percentages. MASt3R [25] on the leaderboard is a private version which performs slightly better than the released model that we use as teacher.

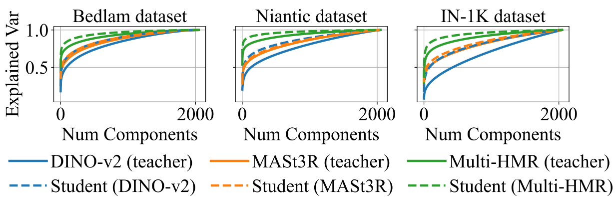


Figure 4. **Cumulative explained variance** computed over features from three representative datasets, for the three teacher encoders (solid lines) and student's projectors (dashed lines).

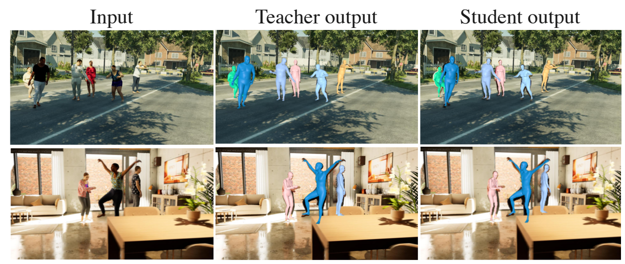


Figure 6. **Qualitative comparison on human mesh recovery** between Multi-HMR (the teacher) and DUNE (the student).

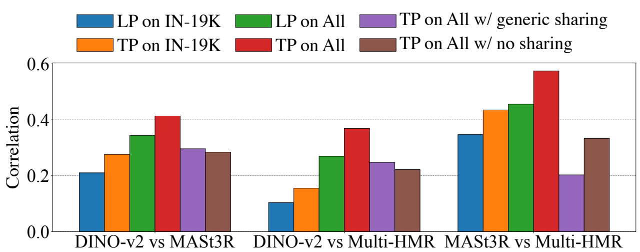


Figure 5. **Correlation of loss updates** during training for each pair of teachers when training with different strategies. Training with TP leads to more *alignment* between teachers regardless of the training data. On the other hand, using all data with all teachers seems to be the best data strategy to improve teacher alignment.

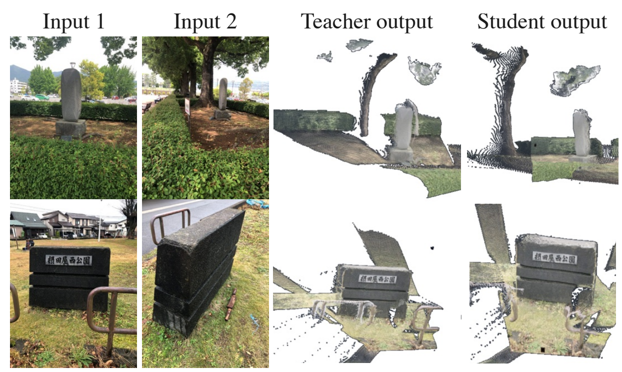


Figure 7. **Qualitative comparison on 3D reconstruction** between MASt3R (the teacher) and DUNE (the student).

Qualitative Results. We present side-by-side qualitative results for the specialized tasks our encoder jointly solves: human mesh recovery in Fig. 6 and 3D reconstruction in Fig. 7. DUNE, combined with the corresponding task decoder, produces visually similar outputs to those of the Multi-HMR and MASt3R teachers, respectively.

6. Conclusions

In this paper, we describe and tackle co-distillation, the challenging multi-teacher distillation task that arises when the set of teachers is composed of models of a very different nature, including generic teachers, whose features generalize well across tasks and domains by design, and teachers specialized for a certain task. We applied co-distillation to DINO-v2 [33], Multi-HMR [3], and MASt3R [25], whose training tasks, training data and properties are highly different. This distillation process produces a strong encoder, DUNE, which, when combined with each teacher's task-specific decoder, performs on par with or surpasses the teachers. Notably, our encoder outperforms MASt3R on the Map-free Visual Relocalization dataset while using significantly fewer parameters.

Acknowledgements. We thank the 3DH and GeoDe teams at NAVER LABS Europe who created Multi-HMR and MASt3R, respectively, for sharing their dataset, training and evaluation pipelines. We also thank César de Souza for developing a real-time multi-task demo interface for DUNE, as well as for naming our model.

References

[1] RGBD objects in the wild: Scaling real-world 3D object learning from RGB-D videos. *arXiv:2401.12592*, 2024. 5

[2] 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. 5, 6

[3] 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, 3, 4, 5, 8

[4] Axel Barroso-Laguna, Sowmya Munukutla, Victor Adrian Prisacariu, and Eric Brachmann. Matching 2D images in 3D: Metric relative pose from metric correspondences. In *Proc. CVPR*, 2024. 7, 8

[5] 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. 5

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021. 1

[7] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *CoRR*, 2020. 5

[8] Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. Can we edit multimodal large language models? In *Proc. EMNLP*, 2023. 2

[9] 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 *Proc. CVPR*, 2016. 6

[10] 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. 6

[11] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *Proc. ICLR*, 2024. 5

[12] MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. *arXiv:2312.06795*, 2023. 2

[13] 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. 5

[14] 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, 2, 3, 5

[15] Sombit Dey, Jan-Nico Zaech, Nikolay Nikolov, Luc Van Gool, and Danda Pani Paudel. ReVLA: Reverting visual domain limitation of robotic foundation models. *arXiv:2409.15250*, 2024. 2

[16] 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. In *Proc. ICLR*, 2021. 4

[17] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proc. CVPR*, 2021. 5

[18] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. In *Proc. NeurIPS*, 2024. 2, 3

[19] Greg Heinrich, Mike Ranzinger, Hongxu, Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2.5: Improved baselines for agglomerative vision foundation models. In *Proc. CVPR*, 2025. 2, 5, 6

[20] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *Proc. ICCV*, 2021. 3, 6

[21] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. 3

[22] Adrian Iordache, Bogdan Alexe, and Radu Tudor Ionescu. Multi-level feature distillation of joint teachers trained on distinct image datasets. *arXiv:2410.22184*, 2024. 2

[23] Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. Model stock: All we need is just a few fine-tuned models. In *Proc. ECCV*, 2024. 2

[24] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *Proc. ECCV*, 2024. 2

[25] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *Proc. ECCV*, 2024. 1, 2, 3, 4, 5, 7, 8

[26] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *Proc. CVPR*, 2018. 5

[27] 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. 5

[28] 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. 5

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: a skinned multi-person linear model. In *ACM Trans. Graphics*, 2015. 3

[30] Yuxiang Lu, Shengcao Cao, and Yu-Xiong Wang. Swiss army knife: Synergizing biases in knowledge from vision foundation models for multi-task learning. In *Proc. ICLR*, 2025. 1, 3, 4, 5

[31] 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. 5

[32] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. ECCV*, 2012. 6

[33] 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. 1, 2, 3, 5, 6, 8

[34] 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. 5

[35] 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. In *Proc. ICML*, 2021. 2, 3

[36] Mike Ranzinger, Jon Barker, Greg Heinrich, Pavlo Molchanov, Bryan Catanzaro, and Andrew Tao. Phi-s: Distribution balancing for label-free multi-teacher distillation. *arXiv:2410.01680*, 2024. 2

[37] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. AM-RADIO: Agglomerative vision foundation model reduce all domains into one. In *Proc. CVPR*, 2024. 1, 2, 3, 4, 5, 6

[38] 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. 5

[39] Karsten Roth, Lukas Thede, A. Sophia Koepke, Oriol Vinyals, Olivier J Henaff, and Zeynep Akata. Fantastic gains and where to find them: On the existence and prospect of general knowledge transfer between any pretrained model. In *Proc. ICLR*, 2024. 2

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3), 2015. 5

[41] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. Multitask prompted training enables zero-shot task generalization. In *Proc. ICLR*, 2022. 2

[42] 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, 3, 4, 5, 6

[43] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv:2210.08402*, 2022. 3

[44] Jinghuan Shang, Karl Schmeckpeper, Brandon B May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. 2024. 1, 2, 4, 5

[45] George Stoica, Daniel Bolya, Jakob Bjorner, Pratik Ramesh, Taylor Hearn, and Judy Hoffman. ZipIt! merging models from different tasks without training. *arXiv:2305.03053*, 2023. 2

[46] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *Proc. ICCV*, 2019. 2

[47] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proc. CVPR*, 2021. 7, 8

[48] Jialiang Tang, Shuo Chen, Gang Niu, Hongyuan Zhu, Joey Tianyi Zhou, Chen Gong, and Masashi Sugiyama. Direct distillation between different domains. In *Proc. ECCV*, 2024. 2

[49] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. SMD-Nets: Stereo mixture density networks. In *Proc. CVPR*, 2021. 5

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, 2017. 4

[51] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. SAM-CLIP: Merging vision foundation models towards semantic and spatial understanding. *arXiv:2310.15308*, 2023. 1, 2, 3, 7

[52] Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. Localizing task information for improved model merging and compression. *arXiv:2405.07813*, 2024. 2

[53] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *Proc. CVPR*, 2024. 7, 8

[54] 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. 5

[55] 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. 5

[56] 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. 5

[57] Monika Wysoczanska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzcinski, and Patrick Pérez. CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation. In *Proc. ECCV*, 2024. 3

[58] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proc. CVPR*, 2024. 3

[59] Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *Proc. CVPR*, 2020. 5

[60] 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. 5

[61] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proc. ICCV*, 2023. 5

[62] Nikolaos-Antonios Ypsilantis, Kaifeng Chen, André Araujo, and Ondřej Chum. Udon: Universal dynamic online distillation for generic image representations. *arXiv:2406.08332*, 2024. 2

[63] 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. 3, 6

[64] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20k dataset. *IJCV*, 2019. 6

[65] Yi Zhu, Yanpeng Zhou, Chunwei Wang, Yang Cao, Jianhua Han, Lu Hou, and Hang Xu. Unit: Unifying image and text recognition in one vision encoder. *arXiv:2409.04095*, 2024. 2