From Pixels to Physical Intelligence: 3D Data Generation at Internet Scale

Mosam Dabhi

CMU-RI-TR-XX-XX

May, 2025

School of Computer Science
The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania

Thesis Committee:
Laszlo Jeni, *chair*
Simon Lucey, *co-chair*
Katerina Fragkiadaki
Jason Saragih (Meta AI)

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Abstract

Modern AI won't achieve physical intelligence until it can extract rich, semantic spatial knowledge from the wild ocean of internet video—not just curated motion-capture datasets or expensive 3D scans. This thesis proposes a self-bootstrapping pipeline for converting raw pixels into large-scale 3D and 4D spatial understanding.

It begins with multi-view bootstrapping: using just two handheld videos and 1% 2D keypoints to produce dense 2D, 3D keypoints—no calibration, no 3D ground truth required. This sets the stage for geometry-only supervision at scale. Next, category-agnostic 3D lifting transformer model generalizes from a single RGB frame or keypoint set to full 3D shape and pose—across dozens of object classes, zero-shot to unseen categories. Then, label-free mixers, a lightweight MLP architecture, rivals transformer accuracy in unsupervised 2D→3D lifting—proving that with the right inductive biases, 3D supervision becomes obsolete. Finally, template-free 4D rigging reanimates articulated objects with dynamic dense meshes and skeletal motion, removing the need for SMPL-style priors.

Building on these validated components, this thesis will contribute: (1) a unified framework integrating these approaches into a continuous learning pipeline, (2) extensive evaluation across diverse domains, and (3) demonstration of physical intelligence capabilities in downstream robotics and AR/VR applications. Preliminary results show this integrated approach expands articulated data distribution coverage while reducing annotation costs compared to traditional methods. The completed thesis will provide a scalable, geometry-grounded foundation for embodied AI, enabling robots, AR/VR agents, and multimodal systems to perceive, reason, and act robustly in the complex spatial world.

Contents

When this dissertation is viewed as a PDF, the page header is a link to this Table of Contents.

List of Figures

List of Tables

Chapter 1

Introduction

In robotics and embodied AI, the ability to infer 3D structure from visual data "in the wild" – i.e. in unconstrained, real-world settings – is a crucial yet unsolved challenge (MBW: Multi-view Bootstrapping in the Wild — OpenReview). Traditional methods for 3D vision often rely on controlled conditions, such as multi-camera rigs or rigid scenes, to obtain ground-truth 3D data. Unfortunately, such setups are expensive, labor-intensive, and impractical in real-world scenarios (MBW: Multi-view Bootstrapping in the Wild — OpenReview) (MBW: Multi-view Bootstrapping in the Wild — OpenReview). This poses a major bottleneck for scaling robot perception: how can we enable machines to understand arbitrary 3D objects and scenes without relying on specialized sensors or extensive manual annotation? The central theme of this proposal is "3D Data Generation in the Wild." By this, we refer to techniques that can generate or supervise 3D structural data from unstructured inputs (like casual videos or single images) captured in everyday environments – with minimal human intervention. Solving this problem has far-reaching importance. Autonomous agents navigating the real world need 3D awareness for safe interaction, while augmented reality and graphics applications demand accurate 3D models of diverse objects. However, unlike common categories such as humans, most objects (especially in the long-tail of category distributions) lack large labeled 3D datasets (MBW: Multi-view Bootstrapping in the Wild — OpenReview) (MBW: Multi-view Bootstrapping in the Wild — OpenReview). The few labeled examples available for such categories make classical training of 3D detectors or reconstructors unreliable. Key Challenges:

Obtaining 3D supervision for arbitrary deformable objects is inherently difficult. Many objects are non-rigid, exhibiting varying shapes and articulations over time ((PDF) Multi-view NRSfM: Affordable Setup for High-Fidelity 3D Reconstruction). Multi-view geometry tells us that triangulating 3D points ideally requires two views, but in practice even two views are insufficient for high-fidelity reconstruction under non-rigid motion ((PDF) Multi-view NRSfM: Affordable Setup for High-Fidelity 3D Reconstruction). Classic Structure-from-Motion (SfM) assumes rigidity; breaking that assumption leads to the field of Non-Rigid Structure-from-Motion (NRSfM). NRSfM is ill-posed: without strong priors, many 3D shape interpretations can explain the same 2D observations (3D-LFM: Lifting Foundation Model) (3D-LFM: Lifting Foundation Model). Meanwhile, single-view 3D lifting (inferring 3D from one image frame) is an even more under-constrained problem (3D-LFM: Lifting Foundation Model), historically requiring object-specific models or extensive training data per object category (3D-LFM: Lifting Foundation Model). These limitations result in poor scalability to new object types or scenes. Despite these hurdles, progress in data-driven approaches suggests a path forward. Recent advances show that with neural networks and self-supervised learning, it's possible to lift 2D keypoints or images into 3D without direct 3D labels (MBW: Multi-view Bootstrapping in the Wild — OpenReview) (3D-LFM: Lifting Foundation Model). The motivation of this thesis work is to tap into multi-view cues, powerful neural architectures, and semantic priors to develop a pipeline that can automatically generate 3D annotations and models for "in-the-wild" data. By dramatically reducing the need for expensive 3D capture or manual annotation, such a pipeline would democratize 3D data collection and unlock new frontiers in robotics and computer vision (GitHub - mosamdabhi/MBW: This work generates 2D and 3D landmark labels from videos with only two or three uncalibrated, handheld cameras moving in the wild. NeurIPS 2022.). Importance to Robotics and AI: The ability to "understand 3D from 2D" in unseen scenarios will directly impact robot autonomy. For example, a home assistant robot should grasp the 3D shape of a new object it sees (for safe manipulation), even if that object category was never seen during training. A drone or self-driving car should infer the 3D layout of a novel environment on the fly for navigation. These capabilities require generalizable 3D perception. Our proposed research aims to provide that by learning 3D structure in a category-agnostic way. In doing so, it contributes to the vision of embodied

AI systems that can learn in any environment. As recent work on 3D foundation models suggests, achieving cross-category 3D understanding can be "revolutionary" for applications in robotics and AR (3D-LFM: Lifting Foundation Model - Extended Reality Technology Center - Carnegie Mellon University). Indeed, the 3D Lifting Foundation Model (3D-LFM) developed in our prior work (described later) has demonstrated the feasibility of reconstructing objects never seen during training, showcasing a level of adaptability and intelligence beyond traditional models (3D-LFM: Lifting Foundation Model - Extended Reality Technology Center - Carnegie Mellon University). In summary, this thesis proposal is driven by the following question: How can we scalably obtain 3D structural knowledge of the world from limited supervision, enabling machines to perceive any object or scene in 3D? The subsequent sections outline our approach to answering this question, through a combination of multi-view bootstrapping, self-supervised learning, and novel neural architectures that together form a unified trajectory towards scalable 3D data generation in the wild.

Chapter 2

Background

Achieving general 3D understanding in unconstrained settings touches on several research threads. Here we review key background areas: (a) Non-Rigid Structure from Motion (NRSfM) and 2D-to-3D lifting, (b) Multi-view self-supervision and bootstrapping, (c) Transformer and MLP-Mixer architectures in vision, and (d) Category-level modeling and semantic priors. 2.1 NRSfM and Single-View 3D Lifting Non-Rigid Structure-from-Motion (NRSfM) addresses recovering 3D shape from 2D observations of a deformable object. Early NRSfM methods (e.g. Bregler et al. 2000) relied on low-rank shape models and often required long video sequences of a single object with point tracks. These classical approaches had limited success on complex real-world deformations. Modern deep learning techniques have breathed new life into NRSfM by learning powerful shape priors from data (C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion) (C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion). For example, the Canonical 3D Pose Networks (C3DPO) of Novotny et al. introduced a deep factorization model that can reconstruct a deformable object from single-view keypoints, by separating shape and viewpoint effects (C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion) (C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion). C3DPO demonstrated state-of-the-art results on unsupervised 3D reconstruction without using any ground-truth 3D, reconstructing categories like human poses from 2D keypoints alone (C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion) (C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From

Motion). Despite such progress, a fundamental limitation remained: most deep NRSfM or lifting models had to be trained per object category, assuming consistent keypoint correspondences and structure within that category (3D-LFM: Lifting Foundation Model). For instance, C3DPO's approach needed all training examples to have the same semantic keypoints (e.g. all are human bodies) to learn a common shape basis. Similarly, the PAUL method (Procrustean Autoencoder for Unsupervised Lifting by Wang Lucey, 2021) achieved impressive results for category-specific 3D keypoint learning, but it was "intimately wedded to the object class" and not easily extensible to new classes (3D-LFM: Lifting Foundation Model). In other words, category-specific models lack scalability – training a new model for each object class is infeasible if we want to cover the long tail of the visual world. To overcome this, recent research has started exploring category-agnostic or multi-category 3D lifting. Our work on 3D-LFM (Lifting Foundation Model) is a direct response to this need. By harnessing transformers' permutation-invariance, 3D-LFM can lift 2D keypoints to 3D for over 30 object categories within a single unified model (3D-LFM: Lifting Foundation Model) (3D-LFM: Lifting Foundation Model). This eliminates the requirement of fixed correspondences across the training data – a significant departure from prior art (3D-LFM: Lifting Foundation Model). The transformer-based design allows the model to handle varying numbers and configurations of keypoints, making it possible to train on diverse categories jointly (3D-LFM: Lifting Foundation Model) (3D-LFM: Lifting Foundation Model). As a result, 3D-LFM achieves object-agnostic single-frame 2D-3D lifting, and it was shown to generalize to entirely unseen categories (even ones with different keypoint layouts) (3D-LFM: Lifting Foundation Model). This capability – reconstructing novel object types without explicit training data – was previously out of reach for traditional models (3D-LFM: Lifting Foundation Model - Extended Reality Technology Center - Carnegie Mellon University). There is also a lineage of works that combine ideas from NRSfM with more explicit structural priors. For example, some methods assume a skeletal or part-based representation for certain categories (like human or animal bodies) to constrain the 3D reconstruction problem. Approaches such as SMAL for animal bodies or Jointformer for human pose incorporate the connectivity of keypoints (e.g. kinematic chains) to regularize lifting. A recent model, JointFormer, applied transformer-based attention to human 2D-to-3D pose lifting, treating joints as tokens, and showed improved accuracy

by learning relationships between keypoints (3D-LFM: Lifting Foundation Model). Likewise, classical human pose estimators (Martinez et al.'s SimpleBaseline, 2017) demonstrated that even a simple MLP can lift 2D skeletons to 3D when trained on large MoCap datasets, but those models are again tied to a specific skeleton topology. Our proposed work will leverage insights from these efforts – in particular, the idea that incorporating structural knowledge (like joint connectivity or symmetry) can enhance generalization. Later in this proposal, we discuss extending our models with semantic part-based reasoning to further boost performance across categories (Section 5). In summary, the literature shows a clear evolution: from category-specific NRSfM solutions that work in limited settings, toward more general frameworks (like 3D-LFM) that aim to cover "anything from 2D to 3D." Our work builds on these foundations, pushing towards the holy grail of few-shot or zero-shot 3D reconstruction for arbitrary objects. 2.2 Multi-View Self-Supervision and Bootstrapping While single-view 3D lifting is extremely challenging, using multiple views of an object can provide vital geometric cues. Classic multi-view geometry (Hartley Zisserman) underpins rigid Structure-from-Motion and SLAM, which have succeeded in reconstructing static scenes from image collections. However, rigid SfM breaks down for deformable targets. One strategy to handle non-rigid objects is to capture them with multiple cameras simultaneously – e.g., motion capture studios or multi-view rigs (like CMU's Panoptic Studio (C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion)). With many synchronized views, one can reconstruct instantaneous 3D shape even if the object moves, by essentially treating short time windows as multi-view snapshots. The drawback is obvious: multi-camera systems are expensive and confined to labs, not usable "in the wild" (C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion). Our research shows that we can achieve similar multi-view benefits with far fewer cameras, by introducing learned priors and intelligent data augmentation. In our early work on Multi-View NRSfM (MV-NRSfM), we argued that two uncalibrated cameras can suffice for high-fidelity 3D reconstruction of non-rigid objects, if we drop the rigidity assumption and leverage a neural shape prior ((PDF) Multi-view NRSfM: Affordable Setup for High-Fidelity 3D Reconstruction). The idea is that even though two views alone are insufficient to triangulate complex deforming shapes classically, a learned prior (trained on other similar shapes) can fill in the missing information. This insight laid the groundwork

for the Multi-view Bootstrapping in the Wild (MBW) system. Bootstrapping with Minimal Supervision: MBW, which we presented at NeurIPS 2022 (MBW: Multiview-bootstrapping in the Wild), is a pipeline that automatically generates 2D and 3D keypoint labels from just a few annotated frames in a multi-view video (MBW: Multi-view Bootstrapping in the Wild — OpenReview) (MBW: Multi-view Bootstrapping in the Wild — OpenReview). The input to MBW is a set of casual, uncalibrated videos of an object or creature captured by 2–3 handheld cameras moving around – for example, visitors filming animals in a zoo (MBW: Multiview-bootstrapping in the Wild) (MBW: Multi-view Bootstrapping in the Wild — OpenReview). Only 1–2Results of MBW: With this approach, MBW was able to achieve 2D keypoint accuracy comparable to fully supervised methods (that use 100Other related efforts include classical structure-from-motion with learned priors, and recent work on unsupervised keypoint discovery. For instance, research on discovering category-specific keypoints (Thewlis et al. 2017; Zhang et al. 2018) use self-supervised objectives (like equivariance under transformations) to learn keypoints without labels. However, those typically yield 2D points that lack explicit 3D understanding. MBW, in contrast, explicitly recovers 3D structure by integrating multi-view geometry into the learning loop. In the domain of human pose, self-supervised approaches (e.g. Rhodin et al. 2018) have used multiple camera views and consistency checks to train pose estimators without 3D labels – an idea philosophically similar to MBW, though applied to a single known skeleton (human) with calibrated cameras. Our work generalizes the multi-view self-supervision paradigm to uncalibrated cameras and entirely new object categories, guided by a neural NRSfM prior. This body of literature underlines a crucial point: multi-view redundancy can compensate for limited annotations if used cleverly. Even two views can detect errors and supervise each other when coupled with a learned deformable model ((PDF) Multi-view NRSfM: Affordable Setup for High-Fidelity 3D Reconstruction) ((PDF) Multi-view NRSfM: Affordable Setup for High-Fidelity 3D Reconstruction). We will carry this principle forward in the proposed research, particularly in exploring egocentric or cross-domain scenarios (Section 5) where multiple viewpoints (even if not simultaneous) may be available over time. 2.3 Transformers and MLP-Mixers in Vision Neural network architecture design has played a pivotal role in pushing the boundaries of 3D understanding. Two recent trends are especially relevant: the rise of Vision Transformers and the emergence of

MLP-Mixer networks as alternatives to both CNNs and Transformers. Transformers for 3D data: Transformers, with their attention mechanism, have proven extremely effective for a variety of vision tasks by modeling long-range relationships and set-based inputs. In our 3D-LFM work, a transformer encoder is central to handling unordered keypoints and different category shapes (3D-LFM: Lifting Foundation Model) (3D-LFM: Lifting Foundation Model). The transformer's permutation equivariance means that it doesn't matter in which order the keypoints are provided – an advantage when each object category might have its own keypoint ordering or when some keypoints are occluded/missing (3D-LFM: Lifting Foundation Model) (3D-LFM: Lifting Foundation Model). This property was exploited by us to create a truly category-agnostic 3D lifting model. Additionally, transformers allow flexible token mixing – they can learn interactions between any pair of keypoints, which is useful to capture both local structure (e.g. points on a limb) and global shape consistency. Recent innovations like NRSfM-Former have also applied transformers specifically for NRSfM, indicating the growing interest in using attention to solve 3D reconstruction problems (3D-LFM: Lifting Foundation Model). NRSfM-Former (2022) introduced an attention-based model for non-rigid factorization, showing improved ability to handle missing data and complex deformations by attending over point trajectories. These advances hint that sequence modeling of point sets (either spatial sequences or temporal sequences) is a powerful paradigm for 3D. However, transformers are not without drawbacks. They tend to be heavy in terms of parameters and computational cost, and require careful training to converge, especially with limited data. In the context of 3D-LFM, while the model achieves remarkable generalization, it might be overkill for certain simpler or more structured scenarios, and it could potentially obscure straightforward inductive biases (like rigid part rotations) in a sea of learned attention weights. This motivates exploring simpler architectures like MLP-based models for our problem. MLP-Mixer Networks: The MLP-Mixer (Tolstikhin et al., NeurIPS 2021) showed that even pure multi-layer perceptron architectures (no convolution, no attention) can attain competitive performance on image classification ([2105.01601] MLP-Mixer: An all-MLP Architecture for Vision - arXiv). The key is to alternate between mixing along the spatial dimension (pixels or patches) and mixing along the feature dimension, using only MLPs for both. This effectively allows communication across the image, similar to what self-attention would do, but with a simpler operation.

In our context, an MLP-Mixer style model could be applied to sets of keypoints or tokens representing parts of an object. One appeal of the Mixer is that it can be more lightweight and easier to train, since it's essentially a sequence of matrix multiplications and nonlinearities. Our ICCV submission (under review) explores a novel architecture that uses an MLP-Mixer for category-specific token mixing without bottlenecks. The motivation is as follows: In a transformer like 3D-LFM, one often uses a bottleneck latent (e.g. a learned latent token or a fixed-size embedding) that aggregates information from all keypoints before decoding to 3D. This bottleneck can sometimes limit the detail or impose a need for correspondences (if, say, one had a fixed position in the latent for each keypoint). We propose an alternative where each category or object-type is associated with a learned token-mixing scheme, implemented via MLPs, that directly maps 2D keypoints to 3D structure without a restrictive bottleneck. In simpler terms, rather than compressing everything into one latent vector, the network mixes the input coordinates through layered MLPs that exchange information across keypoints in a flexible way. Early experiments indicate that this Mixer-based model can achieve similar accuracy to the transformer-based 3D-LFM on seen categories, while being easier to extend to new categories: we can adjust the token mixing per category with minimal re-training, avoiding the need to learn an entirely new attention map from scratch. This architecture is still in development (hence in the Completed Work we describe it as "ICCV submission"), but it represents a promising direction to simplify 3D lifting models. Moreover, MLP-Mixers might synergize with part-based representations. For instance, one can imagine an MLP-mixer where each "token" corresponds to a specific semantic part (like a limb, or a component of an object), and the mixing layers learn how those parts relate (potentially differently for different object classes). This approach naturally aligns with category-specific parameter sharing – parameters can be shared for common parts across categories, but also specialized where needed. Literature on universal representations vs. category-specific models suggests that a mixture-of-experts or adapter-style architecture can often yield the best of both worlds: generalization to new categories and specialization to known ones. Our Mixer model can be seen as a step in that direction, eliminating a single bottleneck and instead distributing the capacity across many tiny "mixing" networks. 2.4 Category-Level Models and Semantic Priors Finally, we consider work in semantic 3D understanding – that is,

models that incorporate higher-level knowledge about object categories or parts into the 3D reconstruction process. As noted, one limitation of earlier methods like PAUL or C3DPO was the need for consistent keypoint semantics during training (3D-LFM: Lifting Foundation Model). They essentially baked in the correspondence by training on aligned keypoints (e.g. always the "nose" keypoint of an animal corresponds to the same index in the shape basis). Our work tries to remove this need, but at some stage, leveraging semantics explicitly could be beneficial, especially for interpretability and for tasks beyond pure geometry (like understanding object affordances or part properties). There is a rich history of part-based models in computer vision – from Pictorial Structures for 2D pose (Felzenszwalb Huttenlocher 2005) to Morphable Models for 3D faces or bodies (Blanz Vetter, SMPL, etc.). The idea is to represent an object as a collection of parts (which could be points, bones, mesh segments, etc.) with constraints between them. In the deep learning era, we see this in approaches such as SMAL for 3D animal shape (Zuffi et al. 2018), which uses a deformable model with learned shape components for different body parts of quadrupeds. Such models provide strong semantic priors – e.g. all four-legged animals have analogous limb structures – which can help generalize to new animals by deforming the known template. In context of our research, we initially avoided category-specific templates to maintain generality. But one direction (outlined in Proposed Work) is to incorporate semantic part reasoning on top of our foundation model. This could involve identifying common structures (like "limb" vs "torso" vs "head" keypoints) across categories and ensuring the model treats them in a consistent way. For example, a "wing" of a bird and the "arm" of a human are different, but a model that knows they are both limb structures might enforce that each has a certain degree of freedom relative to the body. Some recent unsupervised keypoint learning works have attempted to discover such groupings automatically (e.g. grouping keypoints by motion correlation or appearance). We may leverage techniques from graph neural networks or structured priors to encode part relationships. A relevant piece of literature is on unsupervised learning of category-specific symmetric 3D keypoints (Kulkarni et al., ECCV 2019), which discovered symmetric parts of objects by enforcing symmetry and consistency constraints ([PDF] Unsupervised Learning of Category-Specific Symmetric 3D ...). That work, and others like it, indicate that even without part annotations, a network can pick up on part structure if properly guided. Additionally, self-supervised learning

in vision provides tools like contrastive learning or autoencoding that could benefit 3D tasks. For instance, an autoencoder that projects an image to 3D and back could learn a latent 3D representation without explicit supervision – this is an "analysis-by-synthesis" approach seen in works like Deep NRSfM++ where a network predicts 3D and projects to 2D to match observations ([PDF] arXiv:2001.10090v1 [cs.CV] 27 Jan 2020). Deep NRSfM++ (2020) specifically showed that with proper network design, one can jointly recover 3D shapes and camera poses from 2D keypoints alone (a purely unsupervised approach) ([PDF] arXiv:2001.10090v1 [cs.CV] 27 Jan 2020). It achieved strong results on benchmark datasets, pointing to the potential of truly self-supervised 2D-to-3D lifting. Lastly, beyond single objects, there's the notion of scene-level 3D understanding in the wild, such as estimating 3D layouts of entire scenes or multiple objects from head-mounted cameras. While that is somewhat outside the primary scope of this thesis (which focuses on object-centric 3D generation), it's worth noting efforts like the EFM3D benchmark (2024) which is aimed at Egocentric 3D Foundation Models (EFM3D: A Benchmark for Measuring Progress Towards 3D ... - arXiv). EFM3D targets first-person video data (e.g. from AR glasses) and includes tasks like 3D object detection and reconstruction in everyday environments (EFM3D: A Benchmark for Measuring Progress Towards 3D ... - arXiv). The existence of such benchmarks underscores a growing demand for 3D perception that works outside of traditional datasets, aligning with our goals. Techniques that succeed on EFM3D likely need to combine all the above: multi-view cues (from head motion), category-level priors (to detect and reconstruct objects), and robust architectures that handle novel scenes. In summary, the literature provides a rich toolbox of ideas – from neural shape priors and transformers to part-based models and self-supervised losses – that we will draw from. Our work stands at the intersection of these areas, aiming to unify them into a coherent approach for scalable, generalizable 3D data generation.

Chapter 3

Foundations of Scalable 3D Labeling

3.1 Multi-View NRSfM: Affordable 3D Reconstruction Setup As a precursor to MBW, we investigated how far one can push Non-Rigid SfM with a minimal camera setup. This resulted in a Master's thesis (2021) titled "Multi-view NRSfM: Affordable Setup for High-Fidelity 3D Reconstruction." The core idea was that by using just two handheld smartphone cameras, one could reconstruct deformable objects at a quality comparable to much larger camera rigs ((PDF) Multi-view NRSfM: Affordable Setup for High-Fidelity 3D Reconstruction) ((PDF) Multi-view NRSfM: Affordable Setup for High-Fidelity 3D Reconstruction). Traditionally, getting high-fidelity 3D of a moving subject required dozens of cameras (e.g. Panoptic Studio, or multi-view dome setups) to capture all angles simultaneously (C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion) (C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion). Our insight was that relaxing the rigidity assumption (which those systems implicitly rely on for multi-view triangulation) opens the door to using far fewer viewpoints. We employed a neural shape prior approach: a deep network was trained on known 3D shapes of a category to learn a deformable shape basis, then at test time it optimized the shape coefficients and camera poses to fit the two-view observations ((PDF) Multi-view NRSfM: Affordable Setup for High-Fidelity 3D Reconstruction) ((PDF) Multi-view NRSfM: Affordable Setup for High-Fidelity 3D Reconstruction). Additionally, we enforced a form of multi-view equivariance –

the idea that the 3D shape predicted from one view's observation, when projected into the other view, should align with that second view's observation. This cross-view consistency acted as a powerful self-supervisory signal. The outcome of this project demonstrated, on categories like human bodies and certain animals, that two or three uncalibrated cameras can yield near motion-capture quality 3D reconstructions ((PDF) Multi-view NRSfM: Affordable Setup for High-Fidelity 3D Reconstruction) ((PDF) Multi-view NRSfM: Affordable Setup for High-Fidelity 3D Reconstruction). This was a foundational insight: it proved the feasibility of "high fidelity 3D with minimal setup". The limitations included needing an initial training on similar shapes (to learn the prior) and relatively short sequences (since solving for shape per frame without temporal continuity). Nonetheless, this work set the stage for MBW by showing that affordable, in-the-wild capture of 3D is possible. It also underlined the importance of breaking the rigidity tenet in multi-view geometry to tackle non-rigid targets ((PDF) Multi-view NRSfM: Affordable Setup for High-Fidelity 3D Reconstruction). 3.2 MBW: Multi-View Bootstrapping in the Wild (NeurIPS 2022) Building directly on the above, MBW was developed as a full pipeline to automatically label deformable objects in videos using a few multi-view examples. MBW stands for Multi-view Bootstrapping in the Wild, reflecting its aim to bootstrap learning on unconstrained data (MBW: Multiview-bootstrapping in the Wild) (MBW: Multiview-bootstrapping in the Wild). This work was published in the NeurIPS 2022 Datasets and Benchmarks track (MBW: Multiview-bootstrapping in the Wild) and also featured a public release of the MBW-Zoo dataset (MBW: Multi-view Bootstrapping in the Wild — OpenReview) (MBW: Multi-view Bootstrapping in the Wild — OpenReview). Technical Approach: As described earlier in the literature review, MBW combines neural 3D priors, deep optical flow, and an iterative self-training scheme. Concretely: A small subset of video frames (1

A neural lifting network (incorporating a learned NRSfM prior) is trained on those few frames to predict 3D keypoints from multi-view 2D inputs. Even with few examples, it captures a coarse shape model.

Using this network's predictions, MBW identifies outlier frames where the reprojection error is high – meaning the current model cannot explain the observations well (neurips$_p oster_d abhi$)($neurips_p oster_d abhi$).$These likely indicate novel poses or occlusions$.

A pre-trained deep optical flow (such as RAFT) propagates the known 2D keypoints

to neighboring frames along motion trajectories (neurips$_p oster_d abhi$)($neurips_p oster_d abhi$).$Thisgenerates$

The lifting network is then used to filter these guesses: any propagated point that leads to a large 3D reprojection error is rejected as an outlier (i.e. likely an incorrect propagation) (neurips$_p oster_d abhi$).$Theremaininghigh-confidencepropagatedpointsareacceptedaspseu$ $labels(\backslash inliers$").

The label set is updated with these new pseudo-labels, and the lifting network is retrained on the expanded set (plus a 2D detector is trained in parallel).

This loop repeats for a few iterations, each time expanding the training set and refining the model, until no more significant outliers are found.

This approach is essentially a self-supervised bootstrapping process guided by multi-view geometry. The multi-view aspect is crucial: with only a single view, there'd be no geometric way to verify propagated labels. But having two or three viewpoints, along with the learned shape prior, means the system can triangulate and detect physically implausible estimations. Key Results: MBW was tested on a variety of categories. For humans (a well-studied category), with just 10 labeled frames out of a video, MBW attained 2D keypoint accuracy close to that of fully supervised HRNet (a state-of-the-art human pose model) (MBW: Multi-view Bootstrapping in the Wild — OpenReview) (MBW: Multi-view Bootstrapping in the Wild — OpenReview). This is impressive given the tiny fraction of supervision. More strikingly, MBW produced full 3D poses for these humans without using any 3D ground truth or calibrated cameras – something impossible with traditional methods. On animals like big cats, monkeys, and birds, MBW similarly obtained realistic 3D skeletal reconstructions, effectively creating new labeled datasets for these creatures (MBW: Multiview-bootstrapping in the Wild) (MBW: Multiview-bootstrapping in the Wild). The MBW-Zoo dataset, released with the paper, contains image frames and their auto-generated 2D/3D keypoint labels for seven exotic animal categories (MBW: Multiview-bootstrapping in the Wild) (MBW: Multi-view Bootstrapping in the Wild — OpenReview). This dataset has been made public to spur research into less-studied categories. MBW's approach of using geometry for OOD (out-of-distribution) detection was particularly novel. Typically, in machine learning, OOD detection refers to identifying inputs that fall outside the training distribution (often using uncertainties or feature distances). Here, we introduced a geometric OOD test: any frame pose that our current shape model can't explain is considered an outlier that needs attention

$(\text{neurips}_{p}oster_{d}abhi).This iesthenotionof\backslash OOD"totheshapespaceoftheobject.Sinceourshapeprior$
$the-wild"variations,untilitstabilizes.Intermsoffoundationalcontribution,MBWdemonstrated$
$shot3Ddatalabelingatscale.Iteffectivelyreducesthehumaneffortfromlabelingthousandsoffram$

Chapter 4

Unifying Scalable 3D Reconstruction

3.3 3D-LFM: 3D Lifting Foundation Model (CVPR 2024) The next major piece of completed work is 3D-LFM, which represents a shift from multi-view scenarios to the single-view 2D-to-3D lifting problem – but with a broad, foundation-model perspective. 3D-LFM stands for 3D Lifting Foundation Model (3D-LFM: Lifting Foundation Model). It was accepted at CVPR 2024 and is a joint effort (Mosam Dabhi, Laszlo Jeni, Simon Lucey) to create a unified model for inferring 3D structure from a single image across many object categories. Problem Addressed: Prior to 3D-LFM, as discussed, most learned 3D reconstruction models were limited to one object category or a small handful of similar categories (3D-LFM: Lifting Foundation Model). The inability to handle category diversity was a glaring gap if we envision scalable 3D perception. Our goal was to design a single model that can be trained on a large variety of object types – humans, animals, vehicles, furniture, 30+ categories in total – and learn a generic lifting capability. We intentionally did not bake in any category-specific template; instead, we relied on the power of data (a mixture of many datasets) and a suitable architecture. Architecture: 3D-LFM uses a transformer-based encoder to process 2D keypoints (the 2D keypoints are assumed given by an off-the-shelf detector, or by manual annotation for training). Each keypoint (with its 2D coordinate and an embedded identity like "this is left eye" if known) is treated as a token. The transformer encoder computes an embedding for each token,

attending to all others. We incorporate a form of positional encoding called Token Positional Encoding (TPE) (3D-LFM: Lifting Foundation Model). Unlike a typical transformer that might use a fixed positional encoding or a learned embedding per keypoint type (which would impose correspondence), our TPE encodes the spatial arrangement of points in a camera-agnostic canonical frame. In essence, we give the model a hint about the relative positions or connectivity of the keypoints (for instance, feeding a graph of the expected connections as an adjacency matrix to the transformer (3D-LFM: Lifting Foundation Model)). This helps it understand, say, which points are neighbors on the object's surface. By not encoding the semantic label of each keypoint explicitly, the model remains agnostic to category – a key difference from correspondence-based methods (3D-LFM: Lifting Foundation Model). After the encoder, a regression head outputs the 3D coordinates of the points and the camera viewing angles. We integrate a Perspective-n-Point (PnP) optimization inside the model to ensure the predicted 3D and camera indeed project to the given 2D points (3D-LFM: Lifting Foundation Model) (3D-LFM: Lifting Foundation Model). This can be seen as a soft constraint or regularization guiding the network to find physically valid 3D shapes. It also anchors the scale and orientation of the prediction. Training data: We amassed a training set covering 30+ categories, drawing from public keypoint datasets (like human pose datasets, animal pose datasets, and even everyday object keypoint datasets when available). The training data is inherently imbalanced (e.g. thousands of human examples vs. maybe tens of examples for rarer categories). Yet, one finding in 3D-LFM is that the model preserves performance across individual categories despite imbalance, essentially behaving as a foundation model that does not forget the underrepresented classes (3D-LFM: Lifting Foundation Model) (3D-LFM: Lifting Foundation Model). Figure 1 of our CVPR paper (the "teaser") visually showed the model's reconstructions on a spectrum of objects, from human body to animal to object, with ground-truth overlay – illustrating its broad competence (3D-LFM: Lifting Foundation Model) (3D-LFM: Lifting Foundation Model). Results: 3D-LFM achieves state-of-the-art or competitive performance on multiple 2D–3D lifting benchmarks (3D-LFM: Lifting Foundation Model). For example, on the Human3.6M dataset (a standard human pose benchmark), our model – trained concurrently on many other categories – is on par with dedicated human pose models. On animal benchmarks like TigDog or MacaquePose, it similarly excels. The

most exciting result was testing 3D-LFM on out-of-distribution (OOD) categories that it never saw during training. We held out some categories entirely (e.g. fish, or certain furniture) and then evaluated the trained model on those. The model could still produce reasonable 3D poses for these novel categories (3D-LFM: Lifting Foundation Model). One showcase was reconstructing the 3D pose of a fish from a single image, even though no fish data was in training – the model treated the fish's keypoints as just another set of points and inferred a plausible 3D structure. The ability to handle new keypoint configurations (different number of keypoints, different topology) stems from the transformer design and the fact that it learned a very general notion of 3D shape from all the other categories. In effect, 3D-LFM contains a universal implicit understanding of 3D shape for many structures, which it can apply in a zero-shot way. It is this zero-shot reconstruction capability that led us to term it a "foundation model." In the context of vision, foundation models (like CLIP for images, or large language models in NLP) capture broad knowledge that can be adapted to many tasks. 3D-LFM similarly captures broad 3D knowledge of shapes, which could be fine-tuned for specific categories or used as-is for novel ones. We even developed a chat interface ("Chat with 3D-LFM") to interpret model outputs, highlighting its potential as a resource for the community (3D-LFM: Lifting Foundation Model). Another result worth noting is the robustness to occlusion and variable point counts. By using mask tokens for missing keypoints and the transformer's flexibility, 3D-LFM can handle cases where an image might have occluded joints (e.g. one arm of an animal is hidden) – it will infer the 3D shape with those points masked out, without breaking. This was quantitatively verified by dropping random keypoints during testing and observing graceful degradation. Traditional methods that require full correspondences would have struggled here. In summary, 3D-LFM is a significant milestone: it provides a single-model solution to lifting 2D observations into 3D across a wide array of object types. It validates our thesis that a unified approach to 3D in the wild is viable. The lessons from 3D-LFM (and its limitations, e.g. being relatively heavy and requiring known 2D keypoints as input) directly inform the remaining work (Sections 3.4, 3.5 and proposed work).

Chapter 5

Towards Template-Free Semantic 4D Re-animation

3.5 RAT4D: Rotation-Aware Transformer for 4D Dynamics The final piece of completed (or in-progress) work to discuss is RAT4D: Rotation-Aware Transformer for 4D dynamics. This is a newer direction, aimed at addressing robustness and equivariance in 3D sequence modeling. Here "4D" refers to 3D space + time (i.e., a dynamic 3D sequence), such as an animation of a moving object. Motivation: In both MBW and 3D-LFM, we largely dealt with static frames or sets of keypoints. However, many applications, especially in robotics, involve temporal sequences – e.g. tracking an object in 3D over time, or understanding the motion dynamics of an agent. When adding the time dimension, one key challenge is handling rotations and orientations. An object might perform the same motion but oriented differently in space (imagine a person walking north vs. east – the underlying motion is the same, just rotated). A model that explicitly understands rotations could generalize the dynamics better and be more robust to viewpoint changes. What is RAT4D? It is essentially a Transformer model designed to be equivariant to rotations in SO(3). We incorporate rotational information in the attention mechanism such that if the input sequence is rotated (in 3D), the model's predictions rotate accordingly, or remain invariant if that's desired. This draws on the field of equivariant neural networks, where one designs layers that respect symmetries (here the symmetry group is 3D rotations) (Rotation invariance and equivariance in 3D deep learning: a survey — Artificial Intelligence Review)

(Rotation invariance and equivariance in 3D deep learning: a survey — Artificial Intelligence Review). There have been prior works on rotation-equivariant CNNs and point cloud networks (e.g. SO(3) equivariant models, spherical CNNs) (Rotation invariance and equivariance in 3D deep learning: a survey — Artificial Intelligence Review) (Rotation invariance and equivariance in 3D deep learning: a survey — Artificial Intelligence Review), as well as some on sequence models. Our RAT4D specifically targets sequences of keypoint sets. It extends a normal transformer by embedding each token (keypoint) not just with a position, but with an orientation feature (for instance, an estimated local coordinate frame or a directional vector). The self-attention computes relationships in a way that is invariant to global rotation – often achieved by using inner products or distances that don't change under rotation, rather than raw coordinate differences (Rotation-Invariant Transformer for Point Cloud Matching) (Rotation-Invariant Transformer for Point Cloud Matching). Application scenarios: One scenario we experimented with is 4D hands – imagine tracking two hands interacting (a dynamic sequence). A model should be able to infer the 3D motion of the hands regardless of how the person is oriented relative to the camera. RAT4D's rotation awareness can improve matching of hand poses across time by not being confused if the person turns around (which rotates the global coordinate frame of the hands) (Rotation-Invariant Transformer for Point Cloud Matching) (Rotation-Invariant Transformer for Point Cloud Matching). Another scenario is a drone observing an animal moving: the animal might rotate its body, but we care about its articulated motion (legs moving, etc.) – a rotation-aware model can separate the global rotation from the internal motion, improving consistency. Though RAT4D is in earlier stages, initial results on synthetic data show that it can significantly improve rotation robustness. We measured performance on sequences where we randomly rotated the input in 3D: a standard transformer's error would increase (because it has to re-learn patterns for each orientation), whereas RAT4D maintained low error, effectively generalizing the learned motion to any orientation. For example, point correspondence across frames was preserved much better with RAT4D, aligning with known benefits of enforcing pose-invariance in matching tasks (Rotation-Invariant Transformer for Point Cloud Matching) (Rotation-Invariant Transformer for Point Cloud Matching). From a broader perspective, RAT4D ties into our theme by tackling the "wildness" of data in terms of viewpoint variations.

When deploying 3D perception in the real world, one cannot control how an object is oriented or how a camera moves – thus equivariance is a powerful tool for robustness. By embedding these ideas into our models, we aim for a system that does not need to see every possible angle during training (which is impossible), but rather inherently knows that rotating an object doesn't change what the object is – it just changes coordinates. This concept will be expanded in Proposed Work, where we plan to integrate RAT4D modules to improve the robustness and consistency of our 3D generation pipeline.

Chapter 6

Overcoming Transformer Limitations

3.4 MLP Mixer for Category-Specific Token Mixing (ICCV Submission 2025) As part of our current efforts (with a manuscript submitted to ICCV 2025, now under review), we designed a new model that re-imagines the 2D-to-3D lifting task through the lens of MLP-Mixer architectures. The working title is "Category-Specific Token Mixing without Bottlenecks". While 3D-LFM proved the effectiveness of a transformer, we wanted to see if a simpler, bottleneck-free MLP approach could match or exceed its performance, especially in scenarios where categories have unique characteristics that a single transformer might not optimally capture. Concept: The model consists of a series of MLP-based mixing layers applied to the input keypoints (and their features). We partition the keypoint tokens by category – not hard partitioning during inference, but rather during training we allow the model to learn slightly different mixing dynamics for different category clusters. This is done via a mechanism akin to a conditional MLP: a small set of parameters (an "adapter") is activated based on the category identity, modulating the mixing layer. This way, the model can share most of its weights across all categories (ensuring generalization and efficiency) but still have category-specific pathways for fine-grained details. Importantly, there is no single latent bottleneck that all information must pass through (unlike a single vector embedding). Instead, the mixing layers gradually propagate information among keypoints. Why remove bottlenecks? In many neural architectures for structured

prediction, a bottleneck (like a fixed-length code) can become a point of information compression that might discard subtle details unique to certain inputs. For 3D shape, we hypothesize that a bottleneck-less design might preserve more variance – for example, a rare horn shape on a certain animal might be lost if the model is forced to compress shape into a generic latent. Our Mixer processes all keypoints in parallel and can maintain unique activation patterns for each structure through the layers. Preliminary findings: In experiments on a subset of categories, the MLP-Mixer model reached comparable accuracy to the transformer (3D-LFM) in reconstructing 3D keypoints, while using fewer parameters. One notable observation was that for categories with very different limb topology (say humans vs. birds), the Mixer's conditional layers adjusted in a way that resembled applying a different transformation to each – effectively it learned two slightly specialized models under one roof. Yet, because these share a base, if we introduce a new category (like a bat, which is somewhere between a quadruped and a bird in structure), the model can interpolate and adapt quickly, rather than starting from scratch. This hints at strong few-shot adaptability, a key goal of ours. Another advantage of this approach is training stability and speed. MLPs are easier to train than transformers on small datasets, generally speaking. We found that by removing self-attention, the training became more stable for rare categories (no issues of attention heads just focusing on dominant classes, etc.). Also, forward passes are faster which could be beneficial for deployment in robotics settings where real-time performance is needed. This work, if accepted, will be a nice complement to 3D-LFM: whereas 3D-LFM gave us broad coverage, the Mixer model gives us a pathway to refine and scale that coverage with potentially less computational overhead. In the narrative of the thesis, this represents exploring alternative neural architectures to improve scalability of 3D lifting.

Chapter 7

Applications, Impact, Extensions

Chapter 8

Proposed Research and Timeline

Connecting Completed Work: Each of the above projects addresses a facet of the overall problem: Multi-view NRSfM and MBW handle data generation and labeling with minimal supervision in multi-view setups.

3D-LFM and the Mixer model address generalizable 3D inference from minimal views (even a single view) across many categories.

RAT4D introduces robustness to transformations and temporal modeling, pushing towards real-time dynamic understanding.

Together, they form the basis of our unifying narrative, which we discuss next. 4. Unifying Narrative: Toward Scalable 3D Generation in the Wild The works completed so far may seem diverse – spanning multi-view geometry, single-view lifting, different network architectures, and even sequence modeling. However, they are unified by the common goal of enabling scalable 3D data supervision and generation in unconstrained settings. Here we articulate the connecting thread and how these pieces build upon one another to form a coherent pathway forward. From Data Bootstrapping to Foundation Models: We began by tackling the data scarcity problem (with MBW), showing that with a clever combination of multi-view cues and priors, one can generate labeled 3D data for new categories almost autonomously. This addresses the input side of the 3D learning loop – ensuring we can get supervision for even the most underrepresented object categories. MBW essentially turns a few manual annotations into a rich 3D dataset (MBW: Multi-view Bootstrapping in the Wild — OpenReview) (MBW: Multi-view Bootstrapping in the Wild — OpenReview). Once such data

is available (e.g., our MBW-Zoo dataset for animals), the next question is: how to learn a model that can make sense of it in a general way? This is where 3D-LFM comes in. We took on the challenge of learning across categories, moving towards a 3D foundation model that embodies generic knowledge of shape and pose. MBW gave us the means to curate diverse data; 3D-LFM provides the means to absorb that diversity into a single network. The unifying narrative here is one of scale and generalization: by training on many categories (some of which were made possible by MBW's auto-labeling), 3D-LFM learned to generalize to new categories without any explicit supervision for them (3D-LFM: Lifting Foundation Model - Extended Reality Technology Center - Carnegie Mellon University). This few-shot or even zero-shot adaptability is a cornerstone of scaling to "the wild" – we cannot possibly have annotated data for everything in the world, so the system must generalize from what it has seen to what it hasn't. The fact that 3D-LFM could reconstruct a completely unseen object (like a fish) is a validation of this principle (3D-LFM: Lifting Foundation Model). Multi-View vs. Single-View Synergy: Another thread connecting the pieces is the interplay between multi-view and single-view approaches. Intuitively, multi-view (as in MBW) is great for data acquisition and verification, because geometry can be cross-checked. Single-view (like 3D-LFM) is the ultimate test of generalization, because it forces the model to rely on learned shape priors. In a deployed scenario, these would work hand-in-hand: we might use multi-view setups opportunistically (e.g., a robot moving around an object to scan it) to bootstrap learning of new objects, and then use the refined single-view model to recognize or reconstruct those objects from any viewpoint later. This pipeline mimics how humans learn too – you may walk around a new object to understand it (multi-view), but later recognize it from just a glance (single view). The thesis narrative ties MBW and 3D-LFM together as sequential steps in a continuous learning system for 3D. Scalable Supervision: The phrase "3D data generation" in the title also refers to the generation of supervision signals. MBW generates labeled keypoints; we can imagine extending that to generating other forms of supervision like part segmentation or depth. The common idea is self-supervision via consistency – whether it's multi-view consistency (MBW) or model-based consistency (3D-LFM making sense of new points), each component reduces the need for explicit labels. All our contributions aim to reduce human supervision: MBW reduces it on the data labeling side, 3D-LFM reduces it on

the model generalization side (fewer per-category models to train), and RAT4D aims to reduce the need for exhaustive viewpoint examples through built-in equivariance. Semantic Generalization: A key goal mentioned is semantic generalization and few-shot adaptability. Our current results show generalization in terms of geometry (keypoints). The next step is semantic – understanding what the parts are, not just where. By incorporating semantic part reasoning (Proposed Work), the system will be able to not only output 3D coordinates, but also attach meaning (e.g. identify this point as "tail" of an animal, or this region as a "wheel" of a car). This makes the 3D output far more useful for downstream tasks in robotics (like grasping the wheel vs. the car body). The works so far have laid the groundwork: MBW and 3D-LFM both rely on keypoints that inherently are semantic (they correspond to meaningful joints or landmarks). The unifying idea is to maintain semantic coherence while scaling up – e.g., ensure that when 3D-LFM handles 30 categories, it doesn't confuse fundamentally different parts (which it doesn't, due to how it's structured). The MLP-Mixer approach further allows category-specific handling of tokens, which implicitly respects that, for instance, "ears" only matter for certain categories while "wheels" matter for others. Towards a Complete Pipeline: Ultimately, we envision a full 3D understanding pipeline: start from raw video/images in the wild, apply a method like MBW to extract 2D/3D keypoint labels (and possibly other annotations like optical flow or silhouettes), then train or update a general model (like 3D-LFM or its Mixer-based successor) that can interpret new instances of those categories with minimal additional input. This model, enhanced with RAT4D, would be robust to viewpoint changes and could operate on streaming data (temporal sequences). If a new category appears that's too different, the system can perform a quick few-shot learning procedure (like MBW or a fine-tuning of the Mixer adapters) to incorporate that category – thus constantly expanding its repertoire. In other words, our contributions piece together a learning system that never stops at one category or one environment. It continuously grows: new data -¿ bootstrapped labels -¿ integrated into foundation model -¿ robust inference on future data. This addresses "in the wild" not as a static benchmark, but as a living process, akin to an embodied agent learning about its world progressively. All these efforts align with the thesis statement that scalable 3D data generation requires unifying multi-view geometry, learned priors, and architecture innovations. Each component tackled one aspect:

Geometry + small data: MBW.

Learned priors + big data: 3D-LFM.

Architecture innovations for efficiency: MLP-Mixer model.

Robustness to real-world conditions: RAT4D equivariance.

By unifying these, we approach the ideal system that can learn any object's 3D structure from minimal information and then recall or imagine that structure under any scenario. 5. Proposed Work (Next 6–12 Months) Having established a strong foundation, the proposed work for the next year will focus on consolidating these ideas and pushing them closer to a deployable, general 3D perception system. The major thrusts of proposed research are: 5.1 Finalizing Ongoing Experiments (ICCV Submission) – Complete the remaining experiments, ablations, and writing for the MLP-Mixer based approach, and respond to reviewer feedback if applicable. This includes benchmarking the Mixer model against 3D-LFM on more categories, testing few-shot learning explicitly (e.g., train on N categories, then fine-tune on a new category with few examples and compare with baseline), and quantifying the benefit of no bottleneck in terms of reconstruction detail. By Q2 2025, we aim to have this work accepted/published or at least in arXiv form.

5.2 Scaling to Out-of-Distribution Categories – Extend our framework to handle truly unseen categories at test time. This involves exploring techniques like knowledge distillation and model fine-tuning. For example, we can use 3D-LFM as a teacher to generate pseudo-3D labels for a new category (say, using a pre-trained vision model to guess keypoints on that new category, then lift with 3D-LFM's shape prior). Then train a student model that specializes in that category. Alternatively, develop an adapter module that can be learned quickly for a new category while reusing the backbone (similar to how NLP models use prompt tuning or adapters). The goal is that when confronted with a new object class, we require only a few images or a short video (which could be labeled via MBW) and can spin up a competent 3D predictor for that class. We will test this on categories that were not in our original training (e.g., training on mammals and then adapting to a reptile, or training on vehicles and adapting to a new type of machinery). Success will be measured by few-shot accuracy and how little performance drops compared to if the category had been in the original training.

5.3 Semantic Part-Based Reasoning – Integrate a part-level understanding into

our models. Concretely, we plan to augment the training data with part annotations or semantic keypoint labels where available (for instance, labels like “wing”, “leg” for animals, or “handle”, “blade” for tools). Using these, we will train a variant of our model that outputs not just 3D coordinates but also a part identity for each point (or perhaps group points into part clusters). We might introduce a graphical model or structured loss that ensures points belonging to the same part move coherently and maintain realistic relative geometry. One approach is to implement a two-stream network: one stream predicts 3D structure, another predicts a partitioning of the points into parts (clustering). The part predictions can be compared to ground-truth part segmentations if available, or enforced via symmetry (points symmetric across the object likely same part) or via kinematic constraints (adjacent in a kinematic tree). This part reasoning will be especially useful for articulated objects (like understanding a door as two parts: frame and panel) and for transfer learning (knowing that a horse’s leg and a cow’s leg are analogous parts, for example). By the end of this, we expect the model to output richer representations – perhaps a set of 3D keypoints grouped into semantically labeled parts.

5.4 Cross-Domain Generalization (Egocentric and Beyond) – Evaluate and adapt our methods to egocentric vision scenarios and other domain shifts. We will use datasets like EFM3D (Meta’s Project Aria data) (EFM3D: A Benchmark for Measuring Progress Towards 3D ... - arXiv) or the Ego4D keypoint dataset, if available, to test how our model performs when the camera perspective is from a first-person view (with unusual angles, motion blur, etc.). Likely, our current models will need domain adaptation because egocentric data has different characteristics (lots of motion, extreme close-ups, etc.). We will explore techniques such as domain adversarial training (to make features invariant to domain), or fine-tuning on a small set of labeled egocentric data. Another aspect is the presence of multiple objects in view – egocentric scenes are more cluttered. We may need to incorporate an object detection step to first isolate the object of interest. A potential extension is to integrate our 3D lifting with a detection model that can propose object bounding boxes and keypoints from raw images, making the pipeline end-to-end: from raw egocentric frames to 3D understanding of all objects in view. A successful outcome here is our system being able to take, for example, a GoPro video of a person interacting with tools and output the 3D poses of the person and objects, even if the video is very different from

our training data (domain gap).

5.5 Robustness via Equivariance (SO(3)-aware models) – Build on the RAT4D concept to incorporate rotation and possibly scale equivariance into our 3D models. This involves two parallel efforts: (a) integrating RAT4D modules into the inference pipeline, and (b) imposing data augmentation and equivariant loss during training. For (a), we might use RAT4D as a post-processor that refines a sequence of 3D predictions ensuring temporal consistency and proper alignment. For instance, given an initial sequence of 3D poses from 3D-LFM frame by frame, RAT4D could smooth and correct them in a rotation-equivariant manner (removing jitter due to viewpoint changes). For (b), we will train a new version of our single-frame model with augmented rotations and enforce that the 3D output rotates accordingly (this can be done by rotating input 2D keypoints by some angle in 3D space, projecting them, and making sure the model's 3D output is the rotated version of the original output). Techniques from equivariant neural networks literature will be consulted to ensure we do this in a sound way (Rotation invariance and equivariance in 3D deep learning: a survey) (Vector Neurons: A General Framework for SO(3)-Equivariant Networks). The goal is to dramatically improve robustness: e.g., if the camera is upside-down or at a weird angle, the model should still confidently output the correct 3D structure (just rotated). We'll quantify robustness by evaluating on augmented test sets with various rotations (and also perhaps using the Rotated 3DLoMatch benchmark for cross-frame matching as a sanity check of low-level equivariance (Rotation-Invariant Transformer for Point Cloud Matching) (Rotation-Invariant Transformer for Point Cloud Matching)). A stretch goal is to incorporate full SE(3) equivariance, i.e., handle translations as well (though translation is trivial for keypoints if we work in relative coordinates). Primarily, rotation equivariance is the big win.

5.6 Low-Cost 3D Data Collection Synthetic-to-Real (Optional) – Time permitting, we will explore avenues to further ease 3D data acquisition using consumer devices and synthetic data. One idea is to use smartphone sensors (commodity RGB-D or LiDAR on phones) to capture partial 3D scans of objects and use those to supervise our models. This could involve creating a small dataset where, say, we scan a toy from multiple angles with an iPhone to get a point cloud, then we generate images of that toy and ensure our model reconstructs the same shape – effectively using the scan as ground truth. Another idea is Synthetic-to-Real transfer: use simulated

environments (like Habitat or Unity) to generate training data of objects in motion (with perfect 3D labels), train models there, and then adapt them to real imagery. We might, for example, create a synthetic dataset of animals using 3D models and see if training on that helps the real-world performance via fine-tuning. Since synthetic data can provide types of supervision hard to get in real life (exact depth maps, part segmentation, etc.), it could augment our approach, especially for rare scenarios (e.g., extreme camera viewpoints or lighting – these can be simulated in Unity). The caution is domain gap, so techniques from domain randomization will be applied to make synthetic images more varied. The success criteria here would be if models pre-trained on synthetic 3D data require significantly less real data to achieve good performance, demonstrating a viable cheaper data route. This task is listed as optional because it is exploratory and depends on time; the priority remains the core self-supervised pipeline.

5.7 Integration and System Demonstration – As a culminating step, integrate the various components (if developed separately) into a unified system. For example, the final system might take multi-view video input for a new category, run an automated process (MBW 2.0) to get 3D labels, fine-tune/adapt the foundation model (3D-LFM/Mixer) with those, and then be ready to do single-view 3D inference on new videos of that category – outputting results with part labels and being robust to orientation. We aim to demonstrate this end-to-end on a few case studies, such as: Case Study 1: A new animal species at the zoo – we collect 5 minutes of video from 3 phone cameras around its enclosure, and within a day our system produces a working 3D model that can estimate the animal's pose from any single image or camera. Case Study 2: A human doing a novel activity with an object – we use a headset (egocentric) to record a first-person view; the system learns the 3D motion of both the human and object parts and can then predict/capture that motion from a single front camera as well. These will show the practicality and generality of the approach.

To keep the progress on track, we outline a tentative timeline (Gantt chart) for the next 12 months: Q2 2025 (Apr–Jun): Complete ICCV submission revisions; run additional ablations for MLP-Mixer model. Begin experiments on part-based reasoning (collect any necessary annotations, set up training with part losses). Initial RAT4D prototype integrated with single-view model (small-scale test).

Q3 2025 (Jul–Sep): Focus on cross-domain generalization – acquire egocentric data, perform domain adaptation experiments. Continue part-based model development and integrate results into main model. If ICCV paper accepted, camera-ready and publication duties. Simultaneously, start scaling to OOD categories experiments (pick 2–3 new categories, do few-shot adaptation tests). By end of Q3, aim to have a technical report or draft on part-based 3D lifting.

Q4 2025 (Oct–Dec): Emphasize robustness and equivariance. Implement full RAT4D training regime with data augmentations; evaluate extensively. By mid Q4, incorporate the ROT-aware model into the pipeline. Also, in Q4 gather any smartphone/synthetic data for optional task; run preliminary synthetic-to-real training if time. Begin writing thesis chapters for completed components.

Q1 2026 (Jan–Mar): Integration and final experiments. Run end-to-end system on case studies, refine any issues (e.g., if multi-view pipeline and single-view model need interfacing adjustments). Collect results, figures, and prepare for thesis proposal defense (if scheduled around this time) or for the final thesis document. Draft expected contributions and conclusions. Buffer time for any spill-over experiments or unforeseen challenges.

This timeline is aggressive but feasible given the momentum from current results and the modular nature of tasks (some can be done in parallel by different team members/collaborators). Regular checkpoints (monthly internal reviews) will ensure we stay on track. In conclusion, the proposed work will significantly advance our system's capabilities in terms of generality (more categories), semantic richness (parts), and robustness (equivariance, domain transfer), moving us closer to a truly scalable 3D perception framework applicable "in the wild." 6. Expected Contributions By the end of this PhD research, we expect to deliver the following contributions to the robotics and vision community: A Unified 3D Perception Framework: A comprehensive system for 3D data generation and interpretation in the wild, encompassing data bootstrapping, foundation model inference, and adaptive learning for new categories. This framework will be one of the first to cover the entire pipeline from raw images to high-level 3D understanding (including semantic parts) in a general-purpose manner.

Publications: We anticipate several high-impact publications stemming from this work. So far:

NeurIPS 2022: "MBW: Multi-view Bootstrapping in the Wild" – published

(MBW: Multi-view Bootstrapping in the Wild — OpenReview) (MBW: Multi-view Bootstrapping in the Wild — OpenReview).

CVPR 2024: "3D-LFM: Lifting Foundation Model" – published (3D-LFM: Lifting Foundation Model) (3D-LFM: Lifting Foundation Model).

ICCV 2025: a paper on the MLP-Mixer for category-specific token mixing (currently under review).

Additionally, a journal paper or RA-Letters submission consolidating the MBW+3D-LFM pipeline or the part-based extension is planned for late 2025, which could serve as a chapter in the thesis as well. Each of these publications advances the state of the art in 3D vision: MBW in data annotation, 3D-LFM in multi-category reconstruction, etc.

Open-Source Code and Datasets: All code developed (for MBW, 3D-LFM, RAT4D, etc.) will be released as open source, extending our existing repositories (MBW: Multi-view Bootstrapping in the Wild — OpenReview) (3D-LFM: Lifting Foundation Model). We have already released the MBW Zoo dataset (MBW: Multi-view Bootstrapping in the Wild — OpenReview); we plan to also release:

The foundation model checkpoint for 3D-LFM (so others can use it as a pre-trained model for their tasks).

Any new dataset annotations from part-based learning (e.g., if we annotate or collect part labels for certain categories).

A possible small egocentric 3D keypoint dataset if we end up labeling some egocentric video for evaluation. By providing these resources, we ensure the research is reproducible and can catalyze further work on 3D learning with limited supervision.

Demonstration of Few-Shot 3D Learning: Through our experiments and case studies, we will demonstrate the practical feasibility of few-shot 3D learning. For instance, one expected result is a qualitative showcase where our model learns to reconstruct a new animal species with just, say, 10 minutes of video and 10 annotated frames (something unimaginable with previous techniques). This will be an important proof-of-concept for adaptive robotic vision systems.

Robust 3D Recognition for Robotics: The integration of equivariance and cross-domain generalization means our contributions will be directly relevant to robotics applications. We expect to contribute a module (or at least empirical evidence) that shows how enforcing physics-based symmetry (rotations) in network design leads to

more reliable perception (Rotation-Invariant Transformer for Point Cloud Matching) (Rotation-Invariant Transformer for Point Cloud Matching). Robotics researchers can take this insight to design better models for, say, object pose estimation or SLAM systems that operate in diverse orientations.

Thesis Manuscript (70-90 pages): A polished thesis document (in LaTeX format per CMU RI standards) that compiles all this work. It will serve as a single reference for anyone interested in 3D data generation in the wild, covering literature, methodology, experiments, and including diagrams for clarity. The thesis will formalize concepts like "multi-view bootstrapping" and "lifting foundation models," hopefully becoming a cited piece of literature in its own right for future students tackling related problems.

Community Impact: By addressing the long tail of 3D data and showing results on categories like animals, our work might impact fields beyond core robotics – e.g., biology (studying animal movement via our automatically reconstructed 3D poses), animation (quickly rigging 3D models from a few videos), and AR/VR (where dynamic 3D content of various objects is needed). We will make an effort to engage with those communities (perhaps via workshops or dataset challenges using MBW or 3D-LFM) to drive adoption of our methods.

Embodied AI Integration: Finally, a less tangible but important contribution is the integration of our 3D perception modules into an embodied AI setting (this could be part of a lab demo or a collaboration). For instance, plugging our model into a robot's perception pipeline to allow it to perceive a new object's 3D pose and then pick it up. Such a demo would illustrate end-to-end value: the robot can be shown a new tool, our system generates its 3D model, and then the robot uses that to plan grasping – all with minimal human programming.

By delivering on these contributions, the thesis will solidify a complete story: starting from a fundamental problem (lack of 3D supervision) and ending with a deployed solution (a generalist 3D perception system). This will significantly push forward the boundaries of what is considered possible in learning 3D from limited and in-the-wild data.

Chapter 9

Conclusions

Appendix A

Appendix

Bibliography