MBW: Multi-view Bootstrapping in the Wild – Supplementary Material

A Ablation study - Iterations in MBW

In this section, we conduct an ablation study analyzing the effects of iterations in our proposed approach. In other words, we discuss the improvements in 2D and 3D landmark predictions as well as the implications of iterations in our proposed approach. We conduct our ablation study on the publicly available human benchmark dataset [4]. We initialise our pipeline (MBW) with 5% 2D input labels and 4 views from “Directions-1” sequence of “Subject-1” [4].

Improvement in 2D and 3D landmark predictions: As shown in Fig. 2a, we see that the major improvement in 2D landmark predictions could be observed between Iteration 0 and Iteration 1. Moreover, we see that as the iterations progress, the 2D landmark prediction error continues to reduce as seen in Fig. 2a. Furthermore, we notice that as the iterations progress, our pipeline continues to further denoise and improve the 2D landmark predictions as well as continues to generate accurate pseudo-labels as visible in Fig. 2b. Similarly, we see that as the iterations progress, the 3D reconstruction error (in PA-MPJPE) continues to decrease as visible in Tab. 1.

We also graphically visualize the effects of MBW at each stage in Fig. 3. With a learned MV-NRSfM over given data, we visualize the first two dimensions of the bottleneck. The initial two dimensions of the bottleneck show the overall spread of the given data. The red dots in this plot represents the initial set of 2D input labels. We color code this scatter plot based on 2D reprojection error. Specifically, the colors in Fig. 3 represent the error calculated from Eq. (2). As the iterations progress, we observe the reprojection error to continue to decrease as better 3D structures as well as 2D landmark predictions are learned iteratively.

Handling occlusions with geometry: Analyzing further, we investigate the type of improvement over different iterations. We notice that the main benefit of using learnable geometric self-supervision (see Sec. 3.2) is its capability to handle occlusions. Figure 1 shows that MBW, in conjunction with MV-NRSfM is able to denoise the 2D landmark predictions as seen in the Iter. 2 columns. Compared to Iter. 1, we observe that MV-NRSfM was able to denoise and then feed the pseudo-label to our iterative pipeline which resulted in correct annotations for cases with severe occlusions. Owing to the above observations, we show improvement of 2D landmark predictions over iterations, specifically in cases where the landmarks are occluded. Quantitative improvement is shown in Fig. 2a while the qualitative improvement is shown in Fig. 1 that shows improvements during Iter. 2, and Fig. 2b that shows improvement during Iter. 3 – where we observe the benefit of using the multi-view constraint of MV-NRSfM.

Denoising and its limitations: Since MV-NRSfM leverages the redundancy in shape variations among different frames, it is less sensitive to the variations of input views, and more capable of detecting outliers and denoising inlier 2D landmark estimates. More specifically, it has the capability of denoising the 2D inputs and providing a 3D structure based on its learned distribution. For the cases shown in Fig. 1 and Fig. 2b, we showcase the denoising capabilities of MV-NRSfM. However, we should note that MV-NRSfM is only able to denoise and refine the inlier estimates if the amount of noise in 2D input labels is small enough. There are two reasons: (i) Inaccurate camera matrix: If the 2D input is extremely noisy in one of the views, even if MV-NRSfM would degenerate to an accurate 3D structure, it would not be able to reliable project the generated 3D structure over the extremely noisy view because of inaccurate camera matrix calculated from OnP or PnP. (ii) Inaccurate 3D structure: If most of the views are noisy or if the baseline between cameras is not wide enough, MV-NRSfM cannot learn to enforce multi-view shape consistency thereby generating an inaccurate 3D structure.

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

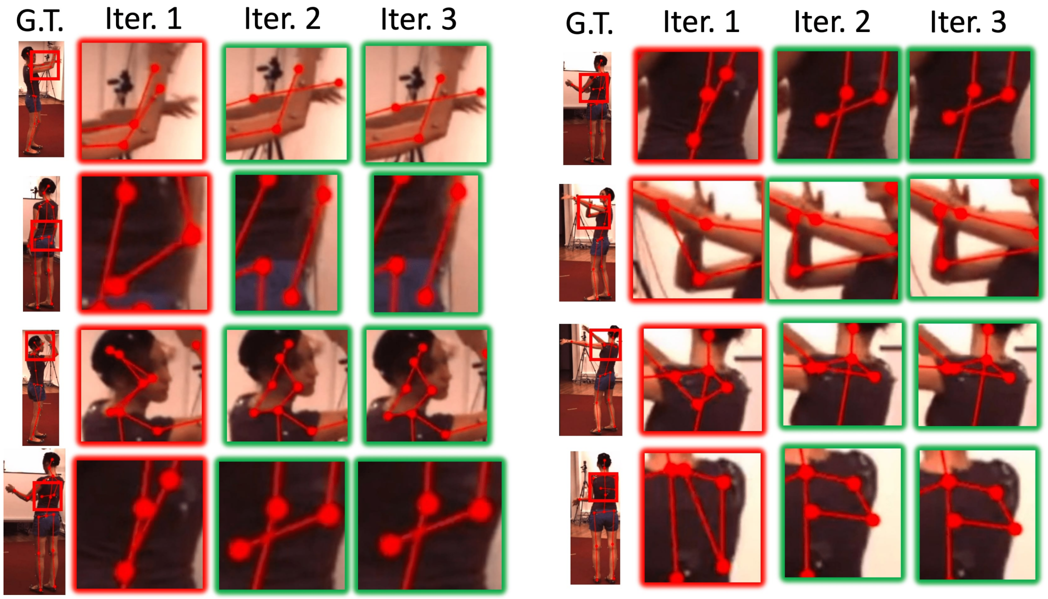
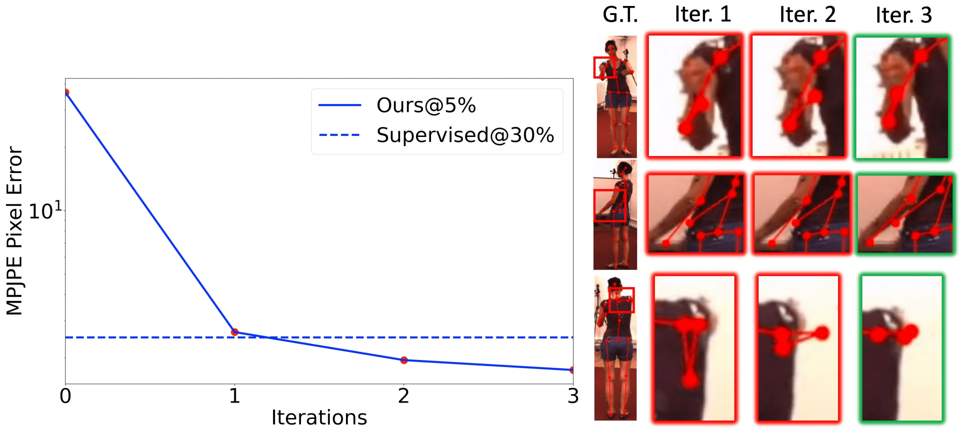


Figure 1: Improvements in 2D landmark predictions as the iterations progress. Specifically, MBW is able to show improvements in cases where the landmarks are occluded. MBW leverages multi-view shape consistency from MV-NRSfM to denoise the inliers from Iter. 1 and use them as pseudo-labels for the next iteration. The red box in G.T. shows where the location of occlusion as well as groundtruth landmark locations. The red glow boxes show the noisy inliers. The green glow boxes show accurate 2D landmark predictions. This figure shows improvements during Iter. 1



(a) Mean-Per-Joint-Position-Error in pixels (2D error).

(b) Improvements in 2D landmark prediction during Iter. 3

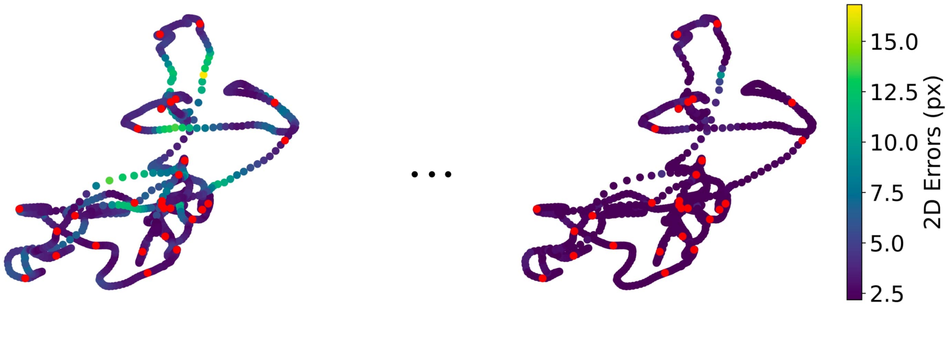
Figure 2: (a) 2D landmark predictions show improvement as the iterations progress. We plot absolute errors in 2D,*i.e.* we calculate Mean-Per-Joint-Position-Error in pixels for each iteration. (b) Similar to Fig. 1, we show improvements in cases with occlusion using the proposed pipeline. This figure shows improvements during Iter. 2

B Initial input labels and Active learning

For the inital set of 2D input labels, we sample uniformly across time and views (unlike Pereira et al. [8] that uses PCA to decide which frames to label). Although we pick labels from each view initially, we carry this action in the initial iteration to learn a good 3D shape prior that enforces multi-view consistency. For the subsequent iterations, we do not necessarily require to pseudo labels for all the views of an instance.

Table 1: Quantitative comparison showing improvement in 3D structure over each iteration on benchmark dataset [4].

Iteration	3D (mm)↓
1	33.3
2	28.3
3	$\mathbf{23.1}$



(a) Iteration 1

(b) Iteration 3

Figure 3: We plot the first two dimensions of the MV-NRSfM bottleneck since it shows the overall coverage and distribution of the sequential data. The red points represent the frames that were given initial 2D input labels. The colorbar of these scatter plots represents the reprojection error (Eq. (2)) between the projected 3D structure from MV-NRSfM and 2D candidate predictions at the corresponding iteration. We observe that during the final iteration, the error reduces substantially as visible by colorbars indicating that the MV-NRSfM network generalizes reasonably well with limited initial labels, and during iterations, it expands its coverage over most of the samples in the dataset.

In the experiments shown in Sec. 4, we did not have a second set of manual annotations (active learning). We observe that since we propose a principled way to detect outliers, our pipeline could be readily used in the active learning domain where our proposed approach of iterations can be useful to dictate the next set of labeling in active learning. Thus, if we have a contiguous chunk of outliers in space and time for our captured video sequence, MBW could find this chunk of outliers in a principled way and is able to exactly specify where we should sample and collect more data for the next iteration if it is used for active learning.

From the above discussion, it is clear that our approach has a rich connection with active learning and could be readily used with the work from Feng et al. [2] that uses active learning to iteratively improve the performance of the network in each iteration. Although it is outside of the scope of the paper, we would like to note that the strengths of our results show its applicability in the active learning scenario.

C MV-NRSfM Architecture Details

As shown in Fig. 4, a 3D structure is drawn from a statistical shape distribution, and consequently projected to 2 or more views using a Perspective-n-Point or Orthographic-n-Point solver. Given a set of frames, the parameters of the shape distribution are adapted by minimizing the error between the predicted and the ground truth 2D keypoints.

The architecture of the multi-view neural shape prior is shown in Fig. 5. Motivated by hierarchical sparse coding [5], we implement the neural shape prior with an autoencoder with a bottleneck dimension of 8 (we keep the same bottleneck dimension for all of our experiments across all different object categories). First, the network h_e extracts the block sparse codes, Ψ. Thereafter, the bottleneck layer factorizes the block sparse code into a projection (camera) matrix and the unrotated vector sparse code, φ. We use the same encoder over all additional views. The vector sparse codes at the bottleneck stage are then pooled together and fed into a shape decoder, h_d to generate the 3D shape at the canonical pose. Finally, the canonical 3D structure is projected over all the views, using the closed-form solution of the PnP or OnP solver.

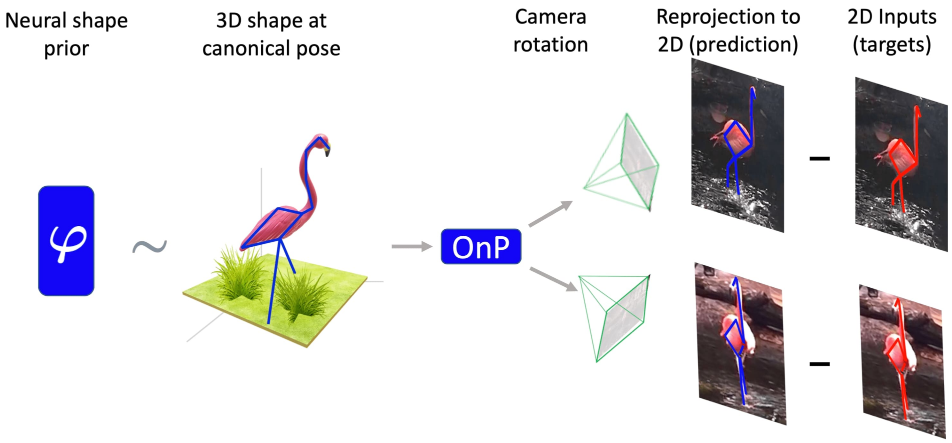


Figure 4: The 3D structure is drawn from a statistical shape distribution using neural shape priors and consequently projected to multiple views using the cameras calculated through OnP formulation. MV-NRSfM minimizes the 2D reprojection error between the predicted 2D projections and target (input) 2D projections.

D Bounding Box calculation

We assume that only a single object of interest (Chimpanzee in Fig. 6 is visible in each frame. In this section, we briefly explain our technique to estimate bounding boxes to reduce the problem into a single object. Using the initial set of 2D input labels, we propagate optical flow [9] and generate 2D candidate predictions over the entire sequence. During this initial (flow iteration) iteration, we calculate the bounding box for the entire sequence using the 2D candidate predictions – by taking the smallest and largest x and y coordinates of 2D candidate predictions. However, since the 2D candidate predictions coming from the optical flow are unreliable and noisy, we pad the bounding box with extra space by a fixed size to make the object visible, as shown in Fig. 6a. For the subsequent iterations, we calculate the bounding boxes from the 2D detector network predictions that are denoised through MV-NRSfM. Since we already denoise the 2D candidate predictions, we remove the extra padding from the bounding box calculation. An experimental video showing the bounding box visualization improvement from initial to final iteration is attached in the supplementary.

E MBW-Zoo Dataset

The URL to platform where the dataset can be viewed and downloaded: https://github.com/mosamdabhi/MBW-Data. Further information concerning the released data asset such as datasheets for dataset, and overall dataset documentation is discussed in the following subsections.

E.1 Overview

We release a challenging dataset and consisting image frames of tail-end distribution categories (such as Fish, Colobus Monkeys, Chimpanzees, etc.) with their corresponding 2D, 3D, and Bounding-Box labels generated from minimal human intervention. Some of the prominent use cases of this dataset include not only sparse 2D and 3D landmark prediction, but also dense reconstruction tasks such as dense deformable shape reconstruction, novel view rendering (NeRF), and finally this dataset could also be used for advancing Simultaneous Localization and Mapping (SLAM) frameworks. We also submit the codebase that we used to generate the above labels for in-the-wild object categories.

The data was collected by two smartphone cameras without any constraints: meaning no guidance or instructions were given as to how the data should be collected. The intention was to mimic the data

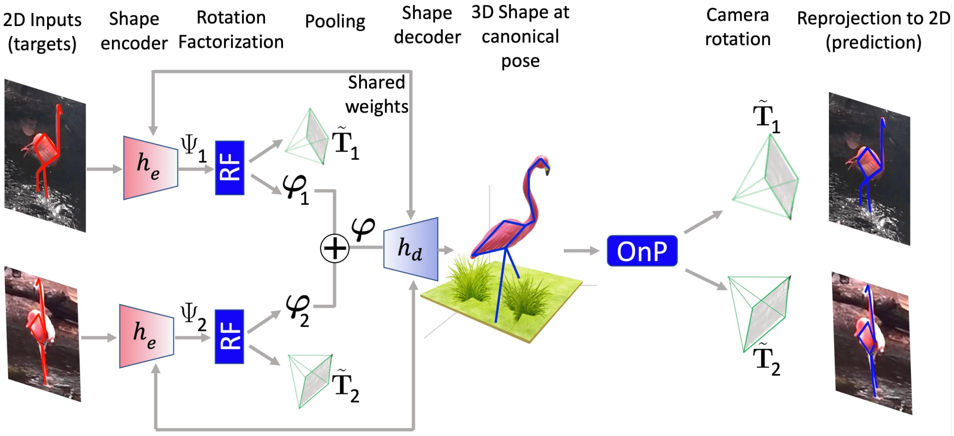
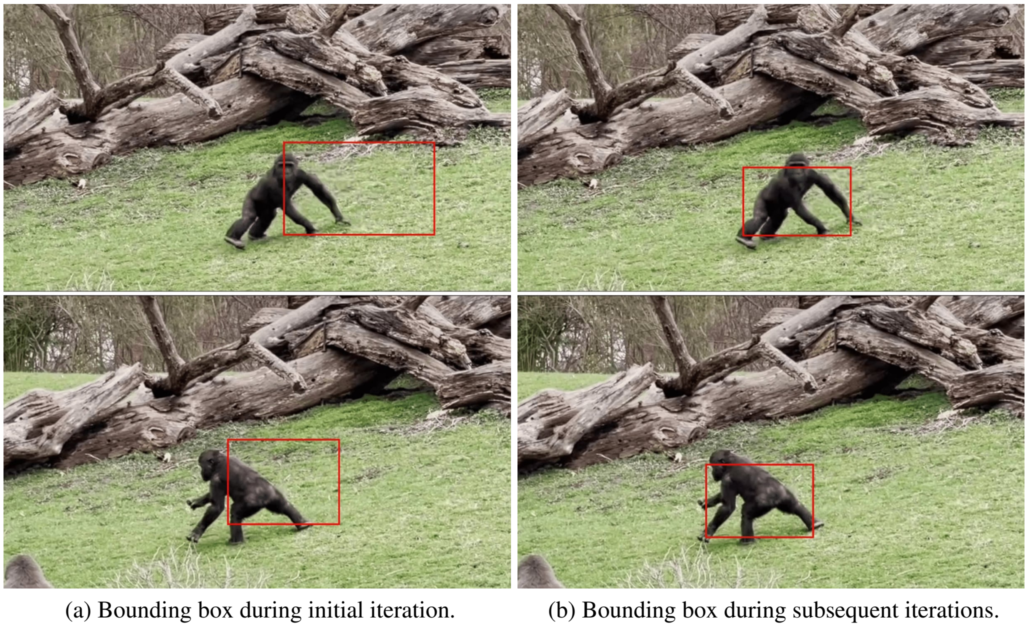


Figure 5: Architecture showing autoencoder-based MV-NRSfM approach. The 2D inputs (targets) from multiple views act as an input to encoder $\boldsymbol{h}_e$ that extracts the block sparse code Ψ from the corresponding views. A Rotation Factorization (RF) layer at the bottleneck stage, factorizes the block sparse code into the respective camera matrices and the unrotated vector sparse code φ. The codes are then fused via *pooling* function into a single code that acts as an input to the shape decoder $\boldsymbol{h}_d$. The shape decoder predicts the 3D structure in the canonical frame while enforcing equivariant view consistency.



(a) Bounding box during initial iteration.

(b) Bounding box during subsequent iterations.

Figure 6: (a) Bounding box visualization during the initial iteration - calculated from the optical flow 2D landmark predictions. Unless readjusted, since optical flow predictions tend to accrue errors, we see that the calculated bounding box exhibit a similar error (elongation to the right). (b) During subsequent iterations, the bounding box predictions become accurate due to less noisy 2D candidate predictions.

captured casually by anyone holding a smartphone grade camera. Due to this reason, the cameras were continuously moving in space changing their extrinsics with respect to each other, capturing an in-the-wild dynamic scene. Thus, this dataset could be used to benchmark robust algorithms in various computer vision tasks.

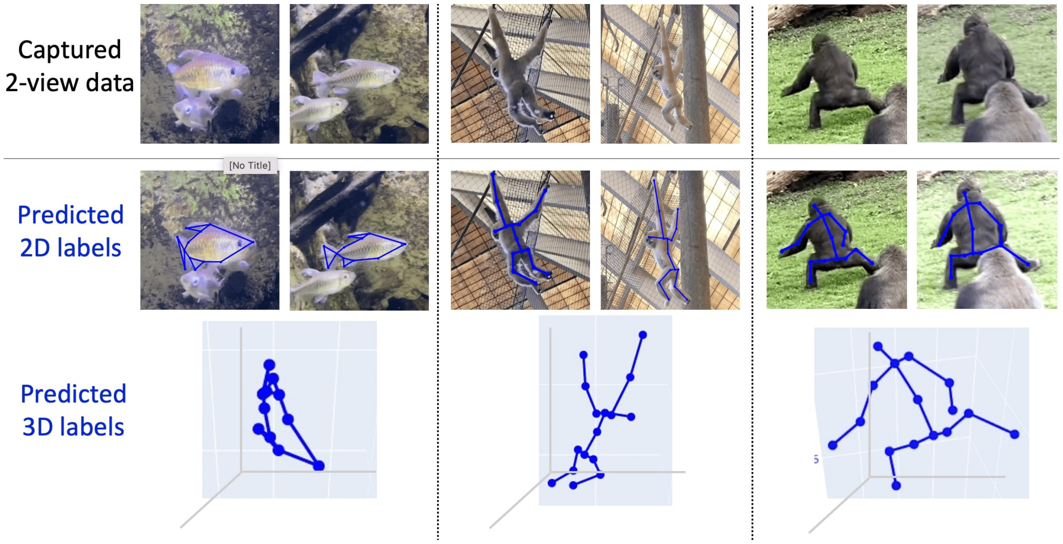


Figure 7: Overview of the dataset. The dataset consists of 2-view synchronized video sequences captured from a zoo visit. MBW provides landmark predictions (labels) that we provide for the categories shown above.

E.2 Collection process

We capture 2-View videos from handheld smartphone cameras. In our case, we used an iPhone 11 Pro Max and an iPhone 12 Pro Max to capture the video sequences. We use Final Cut Pro to manually synchronize the 2-View video sequences using the audio signal and time stamps. Please note that all we require are 2-view synchronized image frames and manual annotations for 1-2% of the data. No camera calibration (intrinsics or extrinsics) is required to run MBW.

E.3 Frames visualization

In total, there are 16154 instances in this dataset from 7 different object categories, coming from 2 camera views. Sample instances are visualized in Fig. 8.

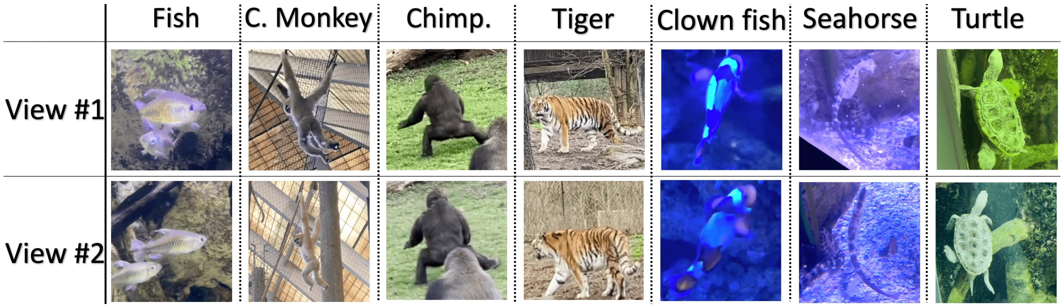


Figure 8: Visualization of instances from 7 different object categories.

E.4 Joint connections visualization

We manually annotate 1-2% of the image frames per view. Our annotation consists of the 2D landmark keypoints. The location of landmarks is chosen to extract articulated information from the objects. Joint connection visualization is shown below.

How to choose the number of joints to track? Different applications require the tracking of different keypoints (landmarks/joints) of tail-end distribution objects. Thus, our approach (MBW) gives user the freedom to decide which keypoints are of interest and hence should be chosen to track.

In the released dataset, we chose keypoints that explain the articulation or deformation of the object catgeory, that we visualize the in Fig. 9.

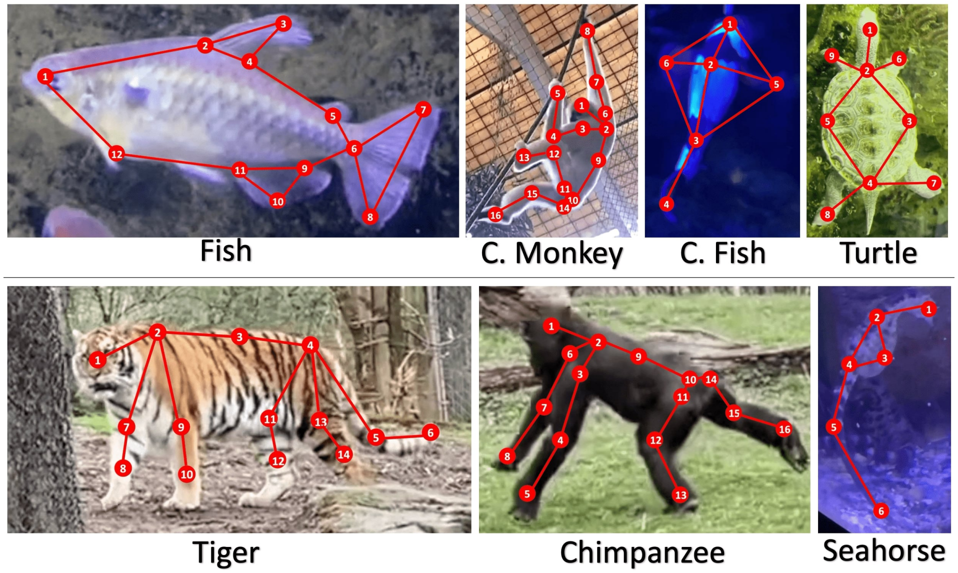


Figure 9: Visualization of keypoint connections on different object categories.

F Dataset format

The dataset (can be viewed and downloaded by the reviewers from here [1]. The downloaded dataset is divided into two directories: `annot/` and `images/`. As names suggest, the `annot/` directory contains annotations and `images/` directory consists of 2-view synchronized image frames.

Annotations format The annotation format is discussed in Tab. 2.

Table 2: The annotations are provided as a `.pkl` file. The pickle files consists of following keys.

Key	Description
`W_GT`	Manual annotation. Non-NaN values for ≈ 2% of data. NaN values for the rest.
`W_Predictions`	2D landmark predictions (labels) generated from MBW.
`S_Pred`	3D landmark predictions (labels) generated from MBW (up-to-scale).
`BBox`	Bounding box crops generated from MBW
`confidence`	Flag specifying confidence (Eq. (2)) for the MBW predictions.

G Datasheets for dataset

As part of making dataset collection more easy and amenable in a wildly unconstrained setup, we collect a dataset of zoo animals using smartphone grade cameras, and annotate (≈ 2%) of the collected frames with 2D keypoint landmarks. We call this dataset the Multiview Bootstrapping in the wild (MBW) Zoo dataset; what follows below is the datasheet [3] describing this data.

G.1 Motivation

1. **For what purpose was the dataset created?** *(Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.)*

   This dataset was created to generate 2D and 3D labels of articulated objects in unconstrained settings. Such unconstrained and casually collected datasets have a wide variety of applications including entertainment, neuroscience, psychology, ethology, and many fields of medicine. Large offline labeled datasets do not exist for all but the most common articulated object categories (e.g., humans, hands, cars). Hand labeling these landmarks within a video sequence is a laborious task. Learned landmark detectors can help, but can be error-prone when trained from only a few examples. As part of contribution of this paper, we provide this dataset where 2D and 3D labels are generated in very challenging scenarios, using our approach.

Note that the user is required to only provide handheld videos from 2 or more views and manual 2D keypoint labels for 15 frames per video. No camera intrinsics or extrinsics information is required to generate the labels shown in this dataset.

2. **Who created this dataset?** *(e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)*?

   This dataset was created by the corresponding author, Mosam Dabhi. At the time of creation, Mosam is a graduate student at Carnegie Mellon University (CMU) in Pittsburgh, Pennsylvania, USA.

3. **Who funded the creation of the dataset?** *(If there is an associated grant, please provide the name of the grant or and the grant name and number.)*

   [N/A]

4. **Any other comments?**

   [N/A]

G.2 Composition

1. **What do the instances that comprise the dataset represent** *(e.g., documents, photos, people, countries)*?

   Each instance is an image of an articulated object (zoo animals, birds, and fish). Approximately 2-5% of images have corresponding manual 2D landmark annotation. From here on, we use the term landmark prediction and label interchangeably for convenience. For the remaining images in the sequences, the 2D and 3D landmark labels, as well as bounding box crops labels are generated by MBW. In particular, each entity also has a confidence flag that is a boolean specifying how much reliable is the label generated by MBW.

2. **How many instances are there in total (of each type, if appropriate)?**

   In total, there are 16154 instances in this dataset from 7 different object categories, coming from 2 camera views. The overall dataset statistics in Tab. 3 reflect the above description.

Table 3: Dataset composition specifications.

Object	Frames (#)	Joints (#)	Manual labels (%)	Labels from MBW
Fish	1456	12	2.7	**Available**
Colobus Monkey	392	16	5.1	**Available**
Chimpanzee	204	16	6.3	**Available**
Tiger	1829	14	0.4	**Available**
Clownfish	910	6	2.0	**Available**
Seahorse	480	6	2.2	**Available**
Turtle	2806	9	[N/A]	[N/A]

Note: We are unable to provide the stereo baseline distance (m) and stereo angle (°) since the data was captured where the cameras were continuously moving thereby changing these metrics.

3. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** *(If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable)?*

   It is a sample of all videos captured casually in an unconstrained environment such as zoo. It is not intended to be representative: the data was collected randomly in the order of visit. This data was collected with the intent to show the applicability of MBW in challenging data scenarios – specifically to label articulated objects in the wild at scale.

4. **What data does each instance consist of?** *("Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description).*

   Please refer Sec. F and the corresponding Table 2.

5. **Is there a label or target associated with each instance? If so, please provide a description.**

   As noted in the Table 2, labels for each instance are associated for the categories with flag **Available**. For the rest, only initial ≈ 2% `W_GT` labels are provided, since we did not run MBW that could provide us with prediction labels. We release this dataset to set a benchmark for solving challenging 2D and 3D landmark prediction tasks for in-the-wild unconstrained video captures.

6. **Is any information missing from individual instances?** *(If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.)**

   The prediction labels (`W_Predictions` , `S_Pred` , `BBox`, and `confidence`) are missing for the categories where MBW is not run as shown in Table 2.

7. **Are relationships between individual instances made explicit? (e.g., users' movie ratings, social network links)?** *If so, please describe how these relationships are made explicit.*

   Instances are unrelated.

8. **Are there recommended data splits (e.g., training, development/validation, testing)?** *If so, please provide a description of these splits, explaining the rationale behind them.*

   Since the sole purpose of this data collection was to generate labels from scratch, we expect this data to be used solely for generating labels for unlabeled data. Thus, we do not explicitly provide a training/validation/testing split; however, we recognize that people may wish to do this, or to do some form of cross-validation. We would suggest cross-validation and test split by dividing the manual labels into 80/10/10 split and pick the samples via a uniform sampling strategy.

9. **Are there any errors, sources of noise, or redundancies in the dataset?** *If so, please provide a description.*

   Since the inital 2D keypoints were manually labeled, there could be some errors in the manual annotations since they were visually localized and clicked as labels.

10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please*

provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.)

The dataset needs to be downloaded from here [1].

11. **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)** *If so, please provide a description.*

    No.

12. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** *If so, please describe why.*

    No.

13. **Does the dataset relate to people?** *If not, you may skip the remaining questions in this section.*

    No

14. **Does the dataset identify any subpopulations (e.g., by age, gender)?** *(If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.)*

    [N/A]

15. **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** *(If so, please describe how)*

    [N/A]

16. **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** *(If so, please provide a description.)*

    [N/A]

17. **Any other comments?**

    [N/A]

G.3 Collection Process

1. **How was the data associated with each instance acquired?** *(Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.)*

    The data was collected by two smartphone cameras without any constraints: meaning no guidance or instructions were given as to how the data was collected. The intention was to mimic the data captured casually by anyone holding a smartphone grade camera. Due to this reason, the cameras were continuously moving in space changing their extrinsics with respect to each other thereby making this a dynamic setup.

2. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)** *How were these mechanisms or procedures validated?*

We captured 2-View videos from handheld smartphone cameras. In our case, we used an iPhone 11 Pro Max and an iPhone 12 Pro Max to capture the video sequences at 30 frames-per-second (fps). We use Final Cut Pro to manually synchronize the 2-View video sequences using the audio signal and time stamps. Please note that all we require are 2-view synchronized image frames and manual annotations for 1-2% of the data. As mentioned above, we are unable to provide the stereo baseline distance and stereo angles since the data was captured where the cameras were continuously moving thereby changing these metrics.

3. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

   Please refer question # 2 and question# 3 of Sec. G.2.

4. **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

   The lead author was helped by Shraddha Thakkar who graciously volunteered to capture the data during their visit to a Zoo.

5. **Over what timeframe was the data collected?** (*Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles? If not, please describe the timeframe in which the data associated with the instances was created.)*

   The dataset was collected on March 19, 2022.

6. **Were any ethical review processes conducted (e.g., by an institutional review board)?** *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   No review processes were conducted with respect to the collection of this data. Manual annotation was conducted by visually localizing the joints on the objects whose accuracy was confirmed by visual inspection.

7. **Does the dataset relate to people?** (*If not, you may skip the remaining questions in this section.*)

   No.

8. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

   [N/A] .

9. **Were the individuals in question notified about the data collection?** (*If so, please describe or show with screenshots or other information how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*)

   [N/A] .

10. **Did the individuals in question consent to the collection and use of their data?** *(If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    [N/A] .

11. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** *(If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

[N/A] .

12. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** (*If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*)

    [N/A] .

13. **Any other comments.**

    No.

G.4 Preprocessing/cleaning/labeling

1. **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** *(If so, please provide a description. If not, you may skip the remainder of the questions in this section.)*

   We did not do any specific preprocessing or cleaning of the data except what is mentioned in question 1 of Sec. G.3. Manual annotations were labeled by visually localizing the joints on the objects in the limited (1-2%) image frames.

2. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** *(If so, please provide a link or other access point to the "raw" data.)*

   Yes, the image frames released under the `images/` directory are sampled from the original "raw" video.

3. **Is the software used to preprocess/clean/label the instances available?** (*If so, please provide a link or other access point.*)

   Yes. We used open-source Matplotlib library (`https://matplotlib.org`) to visualize and label the joints. For further details, please refer question # 1 of this subsection.

4. **Any other comments?**

   None.

G.5 Uses

1. **Has the dataset been used for any tasks already?** *(If so, please provide a description.)*

   Yes, the dataset has already been used for the task of 2D and 3D landmark predictions (sparse keypoints) by MBW. The task specifications are discussed in the MBW paper.

2. **Is there a repository that links to any or all papers or systems that use the dataset?** *(If so, please provide a link or other access point.)*

   No.

3. **What (other) tasks could the dataset be used for?**

   This dataset could be used for the following computer vision tasks:

   - Dense 3D reconstruction of the given articulated object categories [6].
   - Scene flow and optical flow generation tasks.
   - Novel view rendering - owing to synchronized multi-view video sequences (NeRF) [7].
   - Estimation of cameras in space to aid the applications of robotics, like approaches in Simulataneous Localization and Mapping (SLAM).

4. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** *(For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?)*

   No, to the best of our knowledge.

5. **Are there tasks for which the dataset should not be used?** *(If so, please provide a description.)*

   Please refer question 3 of this subsection. Apart from that, our answer to this question is: No, to the best of our knowledge.

6. **Any other comments?**

   None.

G.6 Distribution

1. **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** *(If so, please provide a description.)*

   Yes, the dataset is freely available under **CC-BY-NC** license.

2. **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** *(Does the dataset have a digital object identifier (DOI)?*

   The dataset can be accessed from [1]. The DOI for the dataset can be found on this GitHub page.

3. **When will the dataset be distributed?**

   Please refer to the question above.

4. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *(If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.)*

   The dataset is distributed under a **CC-BY-NC** license.

5. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** *(If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.)*

   No.

6. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** *(If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.)*

   Not to our knowledge.

7. **Any other comments?**

   No.

G.7 Maintenance

1. **Who is supporting/hosting/maintaining the dataset?**

   The authors are maintaining and hosting the dataset information page on GitHub, while the dataset itself is hosted on Zenodo platform to have a persistent DOI.

2. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

   E-mail address of the corresponding author is provided at the dataset access page [1].

3. **Is there an erratum?** *(If so, please provide a link or other access point.)*

   Currently, no. As errors are encountered, future versions of the dataset may be released (but will be versioned). The information to access the latest version (with updated DOI) will all be provided in the same GitHub location.

4. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances')?** *(If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?)*

   Same as previous.

5. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** *(If so, please describe these limits and explain how they will be enforced.)*

   No.

6. **Will older versions of the dataset continue to be supported/hosted/maintained?** *(If so, please describe how. If not, please describe how its obsolescence will be communicated to users.)*

   Yes; all data will be versioned.

7. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** *(If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.)*

   Errors may be submitted by opening issues on GitHub. More extensive augmentations may be accepted at the authors' discretion.

8. **Any other comments?**

   None.

References

[1] Mosam Dabhi. MBW Zoo data, 2022. URL `https://github.com/mosamdabhi/MBW-Data`.

[2] Qi Feng, Kun He, He Wen, Cem Keskin, and Yuting Ye. Active learning with pseudo-labels for multi-view 3d pose estimation. *arXiv preprint arXiv:2112.13709*, 2021.

[3] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

[4] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[5] Chen Kong and Simon Lucey. Deep non-rigid structure from motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1558–1567, 2019.

[6] Suryansh Kumar, Luc Van Gool, Carlos EP de Oliveira, Anoop Cherian, Yuchao Dai, and Hongdong Li. Dense non-rigid structure from motion: A manifold viewpoint. *arXiv preprint arXiv:2006.09197*, 2020.

[7] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

[8] Talmo D Pereira, Nathaniel Tabris, Arie Matsliah, David M Turner, Junyu Li, Shruthi Ravindranath, Eleni S Papadoyannis, Edna Normand, David S Deutsch, Z Yan Wang, et al. Sleap: A deep learning system for multi-animal pose tracking. *Nature methods*, pages 1–10, 2022.

[9] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.