Learning Articulated Shape with Keypoint Pseudo-labels from Web Images

Anastasis Stathopoulos
Rutgers University

Georgios Pavlakos
UC Berkeley

Ligong Han
Rutgers University

Dimitris Metaxas
Rutgers University

Abstract

*This paper shows that it is possible to learn models for monocular 3D reconstruction of articulated objects (*e.g. *horses, cows, sheep), using as few as 50-150 images labeled with 2D keypoints. Our proposed approach involves training category-specific keypoint estimators, generating 2D keypoint pseudo-labels on unlabeled web images, and using both the labeled and self-labeled sets to train 3D reconstruction models. It is based on two key insights: (1) 2D keypoint estimation networks trained on as few as 50-150 images of a given object category generalize well and generate reliable pseudo-labels; (2) a data selection mechanism can automatically create a "curated" subset of the unlabeled web images that can be used for training – we evaluate four data selection methods. Coupling these two insights enables us to train models that effectively utilize web images, resulting in improved 3D reconstruction performance for several articulated object categories beyond the fully-supervised baseline. Our approach can quickly bootstrap a model and requires only a few images labeled with 2D keypoints. This requirement can be easily satisfied for any new object category. To showcase the practicality of our approach for predicting the 3D shape of arbitrary object categories, we annotate 2D keypoints on 250 giraffe and bear images from COCO in just 2.5 hours per category.*

1. Introduction

Predicting the 3D shape of an articulated object from a single image is a challenging task due to its under-constrained nature. Various successful approaches [14, 19] have been developed for inferring the 3D shape of humans. These approaches rely on strong supervision from 3D joint locations acquired using motion capture systems. Similar breakthroughs for other categories of articulated objects, such as animals, remain elusive. This is primarily due to the scarcity of appropriate training data. Some works (such as CMR [15]) learn to predict 3D shapes using only 2D labels for supervision. However, for most object categories even 2D labels are limited or non-existent. We ask: *how can we learn models that predict the 3D shape of articulated objects in-the-wild when limited or no annotated images are available for a given object category?*

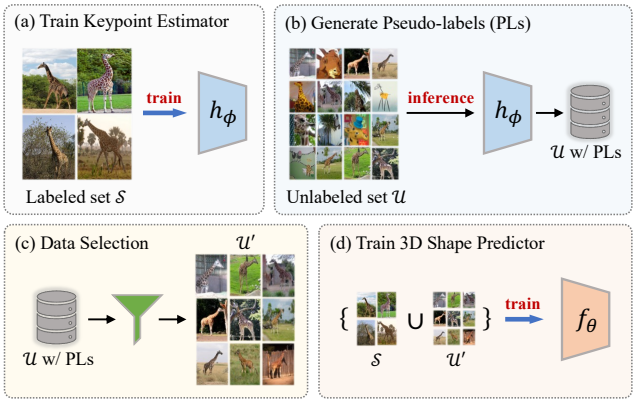

Figure 1. **Overview of the proposed framework**. It includes: (a) training a category-specific keypoint estimator with a limited labeled set $\mathcal{S}$, (b) generating keypoints pseudo-labels on web images, (c) automatic curation of web images to create a subset $\mathcal{U}'$, and (d) training a model for 3D shape prediction with images from $\mathcal{S}$ and $\mathcal{U}'$.

In this paper we propose an approach that requires as few as 50-150 images labeled with 2D keypoints. This labeled set can be easily and quickly created for any object category. Our proposed approach is illustrated in Figure 1 and summarized as follows: (a) train a category-specific keypoint estimation network using a small set $\mathcal{S}$ of images labeled with 2D keypoints; (b) generate 2D keypoint pseudo-labels on a large unlabeled set $\mathcal{U}$ consisting of automatically acquired web images; (c) automatically curate $\mathcal{U}$ by creating a subset of images and pseudo-labels $\mathcal{U}'$ according to a selection criterion; (d) train a model for 3D shape prediction with data from both $\mathcal{S}$ and $\mathcal{U}'$.

A key insight is that current 2D keypoint estimators [30, 34, 38] are accurate enough to create robust 2D keypoint detections on unlabeled data, even when trained with a limited number of images. Another insight is that images from $\mathcal{U}$ increase the variability of several factors, such as

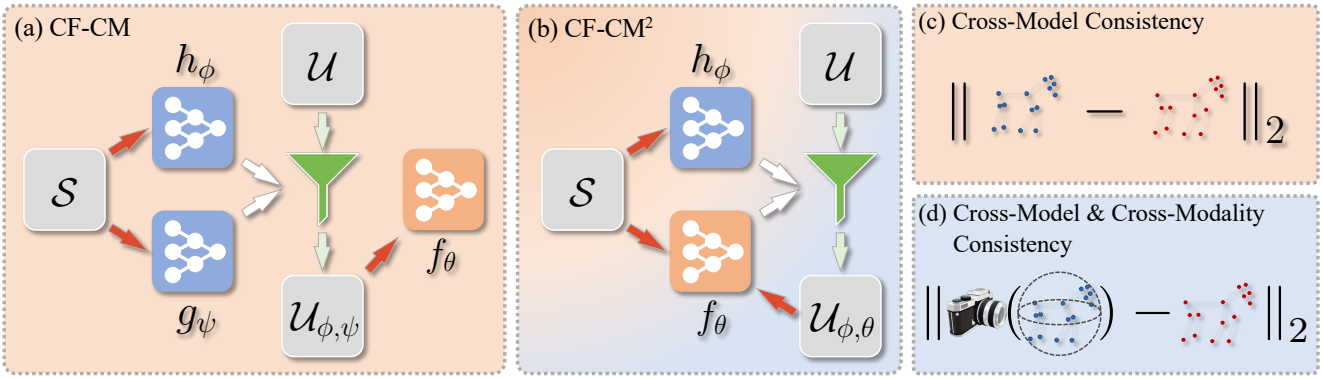

Figure 2. Given a small set $\mathcal{S}$ of images labeled with 2D keypoints, we train a 2D keypoint estimation netowrk h_ϕ and generate keypoint pseudo-labels on web images (set $\mathcal{U}$). We select a subset of $\mathcal{U}$ to train a 3D shape predictor f_θ. Two methods for data selection can be seen here: **(a) CF-CM**: an auxiliary 2D keypoint estimator g_ψ generates predictions on $\mathcal{U}$ and images with the smallest discrepancy between the keypoint estimates of h_ϕ and g_ψ are selected (criterion (c)); **(b) CF-CM^2**: f_θ is trained with samples from $\mathcal{S}$ and generates predictions on $\mathcal{U}$. Images with the smallest discrepancy between the keypoint estimates of h_ϕ and f_θ are selected (criterion (d)) to retrain f_θ.

camera viewpoints, articulations and image backgrounds, that are important for training generalizable models for 3D shape prediction. However, the automatically acquired web images contain a high proportion of low-quality images with wrong or heavy truncated objects. Naively using all web images and pseudo-labels during training leads to degraded performance as can be seen in our experiments in Section 4.2. While successful pseudo-label (PL) selection techniques [4, 5, 23, 29, 40] exist for various tasks, they do not address the challenges in our setting. These works investigate PL selection when the unlabeled images come from curated datasets (*e.g.* CIFAR-10 [20], Cityscapes [6]), while in our setting they come from the web. In addition, they eventually use all unlabeled images during training while in our case most of the web images should be discarded. To effectively utilize images from the web we investigate four criteria to automatically create a "curated" subset that includes images with high-quality pseudo-labels. These contain a confidence-based criterion as well as three consistency-based ones (see Figure 2 for two examples).

Through extensive experiments on five different articulated object categories (horse, cow, sheep, giraffe, bear) and three public datasets, we demonstrate that training with the proposed data selection approaches leads in considerably better 3D reconstructions compared to the fully-supervised baseline. Using all pseudo-labels leads to degraded performance. We analyze the performance of the data selection methods used and conclude that consistency-based selection criteria are more effective in our setting. Finally, we conduct experiments with varying number of images in the labeled set $\mathcal{S}$. We show that even with only 50 annotated instances and images from web, we can train models that lead to better 3D reconstructions than the fully-supervised models trained with more labels.

2. Related Work

Monocular 3D shape recovery. The task of 3D shape recovery from a single image is solved considerably well for the human category, with many approaches achieving impressive results [14, 17, 19, 31]. Their success can be attributed to the existence of large datasets with 3D [13, 27] and 2D annotations [1, 26]. However for other articulated object categories, such as animals, datasets with 3D annotations do not exist and even 2D labels are scarce and in some cases not available.

Recent works [10, 15, 18, 21, 22, 24] have addressed several aspects of this problem. In CMR [15], the authors train a system for monocular 3D reconstruction using 2D keypoint and mask annotations. Although their approach works well, it relies on a large number of 2D annotations, such as 6K training samples for 3D reconstruction of birds in CUB [36], which limits its direct applicability to other categories. Follow-up works [10, 24] attempt to eliminate this reliance by using part segmentations [24] or mask labels [10], but their performance is still inferior to CMR [15], and they require mask or part segmentation annotations during training, which are not always available. Instead, we propose a novel approach that addresses the scarcity of 2D keypoint annotations by effectively utilizing images from the web.

In [18], ACFM leverages temporal information for supervision and applies the learned model on a per-frame basis. In [21, 22] the authors generate 3D reconstructions in the form of a rigid [22] or articulated [21] template, training their models with masks and 2D keypoint labels from Pascal VOC [9] and predicted masks (using Mask-RCNN [11]) from ImageNet [7]. Despite using only a small labeled set and mask pseudo-labels, their approach is not fully automatic as they manually select the predicted masks for their

training set. Another line of work [2, 33], regresses the parameters of a statistical animal model [41], but their work only reconstructs the 3D shape of dogs, for which large-scale 2D annotations exist [2]. BARC [33] also uses a datasest with 3D annotations during a pre-training stage.

Learning with keypoint pseudo-labels. The use of keypoint pseudo-labels has been widely investigated in various works related to 2D pose estimation [3, 8, 23, 28, 29, 32, 39]. In [32, 39], the authors investigate learning with keypoint pseudo-labels for human pose estimation, while in [3,23,29] the goal is to estimate the 2D pose of animals. In the case of animal pose estimation, PLs are used for domain adaptation from synthetic data [23, 29] or from human poses [3]. What is common across these works is that they integrate PLs into the training set (in some cases progressively) and that at the end of the process all samples from the unlabeled set are included. This is not an issue as images in the unlabeled set come from curated datasets. While these approaches can potentially enhance the quality of PLs, a data selection mechanism is still necessary in our case because the unlabeled images come from the web, and most of them should be discarded. To this end, we adapt PL selection criteria form previous works, such as keypoint confidence-based filtering from [3] and multi-transform consistency from [23, 29, 32], to automatically curate the acquired web images.

3. Approach

Given a labeled set $\mathcal{S} = \{(I_i^{(s)}, x_i)\}_1^{N_s}$ comprising of N_s images with paired 2D keypoint annotations and an unlabeled set $\mathcal{U} = \{I_i^{(u)}\}_1^{N_u}$ containing N_u unlabeled images, our goal is to learn a 3D shape recovery model by effectively utilizing both labeled and unlabeled data. We use a limited annotated set N_s and unlabeled images from the web, thus N_u is much larger than N_s. We investigate how keypoint pseudo-labeling on unlabeled images can improve existing 3D reconstruction models. CMR [15] and ACSM [21] are chosen as two canoncial examples. Next, we present some relevant background on monocular 3D shape recovery, CMR and ACSM, while in Section 3.2 we present our proposed approach.

3.1. Preliminaries

Given an image I of an object, its 3D structure is recovered by predicting a 3D mesh M and camera pose π with a model f_θ whose parameters θ are learned during training.

CMR. The shape in CMR [15] is represented as a 3D mesh $M \equiv (V, F)$ with vertices $V \in \mathbb{R}^{|V|\times 3}$ and faces F, and is homeomorphic to a sphere. Given an input image I, CMR uses ResNet-18 [12] to acquire an image embedding, which is then processed by a 2-layer MLP and fed to two independent linear layers predicting vertex displacements $\Delta_V \in \mathbb{R}^{|V|\times 3}$ from a template shape $T \in \mathbb{R}^{|V|\times 3}$ and the camera pose. The deformed 3D vertex locations are then computed as $V = T + \Delta_V$. A weak-perspective camera model $\pi \equiv (s, \mathbf{t}, \mathbf{q})$ is used, where $s \in \mathbb{R}_+$ is the scale, $\mathbf{t} \in \mathbb{R}^2$ the translation and $\mathbf{q} \in \mathbb{R}^4$ the rotation of the camera in unit quaternion parameterization. Thus, CMR learns to predict $f_\theta(I) \equiv (\Delta_V, \pi)$. We refer the reader to [15] for more details.

ACSM. The shape in ACSM [21] is also represented as a 3D mesh $M \equiv (V, F)$. The mesh M is a pre-defined template shape for each object category with fixed topology. Given a template shape T, its vertices are grouped into P parts. As a result, for each vertex v of T an associated membership $a_p^v \in [0, 1]$ corresponding to each part $p \in \{1, \dots, P\}$ is produced. The articulation δ of this template T is specified as a rigid transformation w.r.t. each parent part, *i.e.* $\delta \equiv \{(t_p, R_p)\}$, where the torso corresponds to the root part in the hierarchy. Given the articulation parameters δ, a global transformation $\mathcal{T}_p(\cdot, \delta)$ for each part can be computed using forward kinematics. Thus, the position of each vertex v of T after articulation can be computed as $\sum_p a_p^v \mathcal{T}_p(v, \delta)$. Therefore, to recover the 3D structure of an object with ACSM we need to predict the camera pose π and articulation parameters δ. ACSM [21] uses the same image encoder with CMR [15]. The camera pose is parameterized and predicted as in CMR, while independent fully-connected heads regress the articulation parameters δ. Thus, ACSM learns to predict $f_\theta(I) \equiv (\delta, \pi)$. For more details, we refer the reader to [21].

Keypoint and mask supervision. CMR and ACSM are supervised with 2D annotations, by ensuring that the predicted 3D shape matches with the 2D evidence when projected onto the image space using the predicted camera. Let's assume k 2D keypoint locations $x_i \in \mathbb{R}^{k\times 2}$ as label for image I_i. Each one of these k 2D keypoints has a semantic correspondence with a 3D keypoint. The 3D keypoints can be computed from the mesh vertices. In CMR the location of 3D keypoints is regressed from that of the mesh vertices using a learnable matrix A, while mesh vertices corresponding to 3D keypoints are pre-determined for ACSM. This leads to k 3D keypoints $\hat{X}_i \in \mathbb{R}^{k\times 3}$ that can be projected onto the image using the predicted camera $\hat{\pi}$. In this setting, keypoint supervision is provided using a re-projection loss on the keypoints:

$$L_{kp} = ||x_i - \hat{\pi}(\hat{X}_i)||_2. \quad (1)$$

Similarly, assuming an instance segmentation mask m_i is provided as annotation for I_i, we can render the 3D mesh $M = (V_i, F)$ through the predicted camera $\hat{\pi}_i$ using a differentiable renderer [16] $\mathcal{R}(\cdot)$ and provide mask supervision with the following loss:

$$L_{mask} = ||m_i - \mathcal{R}(V_i, F, \hat{\pi}_i)||_2. \quad (2)$$

In ACSM, a second network $f'_{\theta'}$ is trained to map each pixel in the object's mask to a point on the surface of the template shape. This facilitates computation of several mask losses. In our experiments with ACSM we only train network f_θ with the keypoint reprojection loss in Eq. (1).

3.2. 3D Shape Recovery Approach

Our proposed approach involves the following steps: (1) annotation of 2D keypoints on N_s images (when no labels are available for the given category) to create a labeled set $\mathcal{S}$; (2) training a 2D keypoint estimation network with samples from $\mathcal{S}$ and generating 2D keypoint pseudo-labels on an image collection from the web (unlabeled set $\mathcal{U}$); (3) selecting data from $\mathcal{U}$ to be used for training according to a criterion; (4) training a 3D shape predictor with samples from $\mathcal{S}$ and $\mathcal{U}$. We explain all the steps in detail below.

Data annotation. We annotate N_S images with 2D keypoints, following standard keypoint definitions used in existing datasets. For instance, we use 16 keypoints for quadruped animals as defined in Pascal [9]. The only requirement is to annotate images containing objects with the variety of poses and articulations we wish to model.

Generating keypoint pseudo-labels. In pseudo-labeling an initial model is trained with labels from a labeled set $\mathcal{S}$ and is used to produce artificial labels on an unlabeled set $\mathcal{U}$. In this work, instead of tasking the 3D shape predictor f_θ to produce pseudo-labels for itself, we propose to generate 2D keypoint pseudo-labels using the predictions of a 2D keypoint estimation network [38] h_ϕ. This is a natural option since 2D keypoint estimators are easier to train with limited data and yield more precise detections compared to reprojected keypoints obtained from a 3D mesh.

Most current methods for 2D pose estimation [30,34,38] solve the task by estimating K gaussian heatmaps. Each heatmap encodes the probabiblity of a keypoint at a location in the image. The location of each keypoint can be estimated as the location with the maximum value in the corresponding heatmap, while that value can be used as a confidence estimate for that keypoint. We train a 2D keypoint estimation network h_ϕ with labels from $\mathcal{S}$ and use it to generate keypoint pseudo-labels on $\mathcal{U}$. As such, the unlabeled set is now $\mathcal{U}_\phi = \{(I_i^{(u)}, \tilde{x}_i^{(u)}, \tilde{c}_i^{(u)})\}_1^{N_u}$, where $\tilde{x}_i^{(u)} \in \mathbb{R}^{k\times 2}$ are the estimated locations and $\tilde{c}_i^{(u)} \in [0,1]^k$ the corresponding confidence estimates for k keypoints of interest in image $I_i^{(u)}$.

Learning with keypoint pseudo-labels. Given a labeled set $\mathcal{S} = \{(I_i^{(s)}, x_i)\}_1^{Ns}$ containing images with keypoint labels x_i, we train f_θ using the supervised loss:

$$L_{kp}^{\mathcal{S}} = \sum_{i=1}^{Ns} ||x_i - \hat{\pi}(\hat{X}_i)||_2, \tag{3}$$

where $\hat{\pi}$ is the predicted camera pose and $\hat{X}_i$ the 3D keypoints computed from 3D mesh vertices V_i. Similarly, given a set $\mathcal{U}_\phi = \{(I_i^{(u)}, \hat{x}_i^{(u)}, \hat{c}_i^{(u)})\}_1^{N_u}$ comprising of images with keypoint pseudo-labels $(\tilde{x}_i, \tilde{c}_i)$, we train using the following modification of the keypoint reprojection loss:

$$L_{kp}^{\mathcal{U}} = \sum_{i=1}^{Nu} ||\tilde{c}_i(\tilde{x}_i - \hat{\pi}(\hat{X}_i))||_2. \tag{4}$$

We can train using data from both $\mathcal{S}$ and $\mathcal{U}_\phi$ with the following objective: $L = L_{kp}^S + L_{kp}^U$. However, as $Ns \ll N_u$, the training process is dominated by samples from $\mathcal{U}_\phi$. When the quality of samples in $\mathcal{U}_\phi$ is low, as it is the case with uncurated data from the web, training with all the samples from the unlabeled set leads to degraded performance. To mitigate this issue, we train using a subset of $\mathcal{U}_\phi$ containing useful images and high-quality pseudo-labels that are chosen according to a selection criterion.

3.2.1 Data Selection

Given a number of images N to be used by $\mathcal{U}_\phi$, we select them according to one of the following criteria.

Keypoint Confidence. This method, referred to as *KP-conf*, selects samples from $\mathcal{U}_\phi$ based on the confidence score of the PLs. In particular, it selects N images with with the highest per-sample sum of keypoint confidences $\sum_k \tilde{\mathbf{c}}_k^{(u)}$.

Multi-Transform Consistency. Intuitively, if one scales or rotates one image the prediction of a good 2D keypoint estimator should change accordingly. In other words, it should be equivariant to the those geometric transformations. We can thus use the consistency to multiple equivariant transformations to select the images with high-quality PLs. We denote this consistency-filtering approach as *CF-MT*.

Cross-Model Consistency. Using this consistency filtering approach, referred to as *CF-CM*, we select samples based on the consistency of two 2D keypoint estimators, h_ϕ and g_ψ, with different architectures. Our insight is that the two models have different structural biases and can learn different representations from the same image. Thus, we use the discrepancy in their predictions as a proxy for evaluating the quality of the generated pseudo-labels. For each image I from $\mathcal{U}$, we generate keypoint pseudo-labels $\tilde{x}_\phi$ and $\tilde{x}_\psi$ with h_ϕ and g_ψ respectively, and calculate the discrepancy between the PLs with the following term:

$$D_{\text{CF-CM}} = ||\tilde{x}_\phi - \tilde{x}_\psi||_2. \tag{5}$$

We select N samples with the minimum discrepancy creating the subset $\mathcal{U}_{\phi,\psi} \subset \mathcal{U}_\phi$.

Cross-Model Cross-Modality Consistency. With this consistency filtering criterion, which we dub *CF-CM*^{2}, we select samples based on their discrepancy between the predictions of the 2D keypoint estimator h_ϕ and the reprojected keypoints from the 3D shape predictor f_θ. Given an image

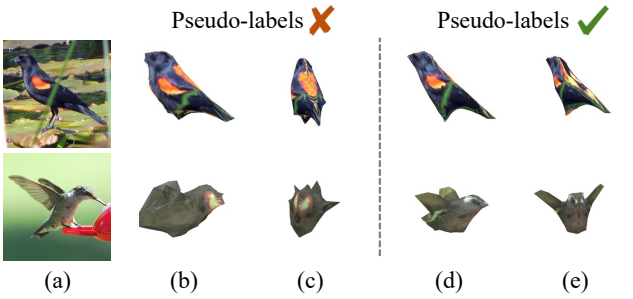

Figure 3. **Sample results on CUB**. For each sample, we show the predicted shape and texture from the inferred camera view (b, d) and a novel view (c, e) for models trained with and without keypoint pseudo-labels.

I from $\mathcal{U}$, we calculate the locations for 3D keypoints $\tilde{X}_\theta$ from the predicted 3D mesh vertices (as explained in Section 3.1). The corresponding 2D keypoints $\tilde{x}_\theta$ are calculated by projecting $\tilde{X}_\theta$ onto the image with the predicted camera parameters $\tilde{\pi}_\theta$, *i.e.* $\tilde{x}_\theta = \tilde{\pi}_\theta(\tilde{X}_\theta)$. As such, we calculate the discrepancy between the keypoints $\tilde{x}_\phi$ and $\tilde{x}_\theta$ as follows:

$$D_{\text{CF-CM}^2} = ||\tilde{x}_\phi - \tilde{\pi}_\theta(\tilde{X}_\theta)||_2. \quad (6)$$

Our insight for this consistency check comes from the fact that the two models employ different paradigms for estimating 2D keypoints and have distinct failure modes. 2D keypoint detectors use a bottom-up approach to determine the locations of 2D keypoints. While their predictions are more pixel-accurate than the reprojected keypoints from a 3D mesh, they are not constrained in any way and mainly rely on local cues. Conversely, 3D shape predictors, such as f_θ, estimate keypoints as part of a 3D mesh, which constrains their predictions with a specific mesh topology. This makes their predictions more robust.

3.3. Additional details

Here, we present some additional details of our approach. We include extensive implementation details in the supplementary material.

Data acquisition. For each target object category we download images from Flickr using the urls that belong to the public and freely usable YFCC100M [35]. However, some images contain wrong objects or objects with heavy truncation. There are also repeated images.

Initial filtering. We remove repeated images with dHash [37]. Then, we detect bounding boxes using Mask-RCNN [11] with ResNet50-FPN [12,25] backbone from detectron2. Only detections with confidence above a threshold τ are retained. We use $\tau = 0.95$ for all object categories.

Keypoints from 2D pose estimation network. We train the *SimpleBaselines* [38] pose estimator, with an ImageNet pretrained ResNet-18 [12] backbone, with labels from $\mathcal{S}$ and use it to generate keypoint PLs $(\tilde{x}_\phi^{(u)}, \tilde{c}_\phi^{(u)})$ on $\mathcal{U}$.

Data Selection. We curate the web images using a selection criterion. We evaluate the effectiveness of the four selection criteria presented previously. We use each criterion to select the best N data samples from the self-labeled web images. We present experiments with $N \in \{1K, 3K\}$.

Training. We train a 3D shape predictor f_θ with keypoint reprojection losses as defined in Eq. (3) & (4).

4. Experiments

Evaluation metrics. While our models are able to reconstruct the 3D shape as a mesh, evaluating their accuracy remains a challenge due to the absence of 3D ground truth data for the current datasets and object categories. Following prior work [15,18,21,22] we evaluate our models quantitatively using the standard "Percentage of Correct Keypoints" (PCK) metric. For $PCK(t)$, the reprojection is "correct" when the predicted 2D keypoint location is within $t \times max(w, h)$ distance of the ground-truth location, where w and h are the width and the height of the object's bounding box. We summarize the PCK performance at different thresholds by computing the area under the curve (AUC), $\text{AUC}_{a_1}^{a_2} = \int_{a_1}^{a_2} PCK(t)\, dt$. We use $a_1 = 0.05$ and $a_2 = 0.1$ and refer to $\text{AUC}_{a_1}^{a_2}$ as AUC in our evaluation. Following [10], we also evaluate the predicted camera viewpoint obtained by our models. We calculate the rotation error $\text{err}_R = \arccos\left(\frac{\text{Tr}(\hat{R}^T\tilde{R})-1}{2}\right)$ between predicted camera rotation $\hat{R}$ and pseudo ground-truth camera $\tilde{R}$ computed using SfM on keypoint labels. Finally, we offer qualitative results to measure the quality of the predicted 3D shapes.

4.1. Simulated Semi-Supervised Setting

First, we investigate the effectiveness of keypoint PLs in a controlled setting. For our experiments we use the CUB dataset [36] that consists of 6K images with bounding box, mask and 2D keypoint labels for training and testing. We study the impact of training with various levels of supervision on 3D reconstruction performance. We split CUB's training set into a labeled set $\mathcal{S}$ and an unlabeled set $\mathcal{U}$ by randomly choosing samples from the initial training set to be included in $\mathcal{S}$, while placing the rest in $\mathcal{U}$. This is a controlled setting where we expect the performance of the models trained with labels from $\mathcal{S}$ and keypoint pseudo-labels from $\mathcal{U}$ to be lower-bounded by that of models trained using only $\mathcal{S}$ and upper-bounded by the fully-supervised baseline that uses the whole training set. We use CMR [15] for the 3D shape prediction network f_θ.

Results on CUB. In Figure 4, we compare the performance of models trained with different number of labeled and pseudo-labeled instances. Following CMR, we also report the mIoU metric. We observe that training with 2D keypoint pseudo-labels consistently improves 3D reconstruction performance by a significant margin across all metrics.

		Horse		Cow		Sheep	
		AUC (↑)	err_R (↓)	AUC (↑)	err_R (↓)	AUC (↑)	err_R (↓)
	ACSM (Mask) [21]	34.9 (↓15.9)	44.3 (↑14.1)	30.7 (↓16.9)	71.0 (↑36.3)	28.2 (↓18.6)	73.1 (↑39.3)
	ACSM (KP+Mask) [21]	37.4 (↓13.4)	76.4 (↑46.2)	-	-	-	-
	ACSM-ours	50.8	30.2	47.6	34.7	46.8	33.8
	ACSM-ours + KP-all	50.2 (↓0.6)	31.9 (↑1.7)	43.9 (↓3.9)	40.1 (↑5.4)	43.5 (↓3.5)	40.7 (↑6.9)
$N = 1K$	ACSM-ours + KP-conf	53.9 (↑3.1)	30.8 (↑0.6)	47.6	38.3 (↑3.6)	48.1 (↑1.3)	38.6 (↑4.8)
$N = 1K$	ACSM-ours + CF-MT	55.1 (↑4.3)	30.0 (↓0.2)	50.7 (↑3.1)	38.9 (↑4.2)	51.1 (↑4.3)	33.3 (↓0.5)
$N = 1K$	ACSM-ours + CF-CM	54.7 (↑3.9)	30.0 (↓0.2)	50.9 (↑3.3)	38.6 (↑3.9)	50.6 (↑3.8)	32.8 (↓1.0)
$N = 1K$	ACSM-ours + CF-CM^2	54.2 (↑3.4)	30.9 (↑0.7)	49.4 (↑1.8)	**32.7 (↓2.0)**	48.6 (↑1.8)	32.1 (↓1.7)
$N = 3K$	ACSM-ours + KP-conf	54.3 (↑3.5)	30.0 (↓0.2)	50.2 (↑2.6)	40.2 (↑5.5)	49.1 (↑2.3)	35.7 (↑1.0)
$N = 3K$	ACSM-ours + CF-MT	55.4 (↑4.6)	30.0 (↓0.2)	49.2 (↑1.6)	39.1 (↑4.4)	51.0 (↑4.2)	33.9 (↑0.1)
$N = 3K$	ACSM-ours + CF-CM	**55.9 (↑5.1)**	29.8 (↓0.4)	50.0 (↑2.4)	38.8 (↑4.1)	48.4 (↑1.6)	34.1 (↑0.3)
$N = 3K$	ACSM-ours + CF-CM^2	55.5 (↑4.7)	**28.1 (↓2.1)**	**52.6 (↑5.0)**	36.2 (↑1.5)	**53.0 (↑6.2)**	**31.3 (↓2.5)**

Table 1. **Evaluation on Pascal.** N is the number of selected images from the web. ACSM-ours + KP-all uses all available web images. Performance change from the fully-supervised baseline ACSM-ours is shown in green/red.

The improvement is larger when the initial labeled set $\mathcal{S}$ is small (*e.g.* 100 images). This is a very interesting finding. The quality of the generated keypoint pseudo-labels decreases when the 2D keypoint estimator is trained with fewer samples. However, the keypoint pseudo-labels are accurate enough to provide a useful signal for training CMR, largely improving its performance.

In Figure 3, we show sample 3D reconstruction results comparing CMR trained with: i) 300 labeled instances and ii) the same labeled instances and additional keypoint pseudo-labels (for 5,700 images). We show the predicted shape and texture from the predicted camera view and a side view. From Figure 3 we can see that the model trained with keypoint pseudo-labels can accurately capture challenging deformations (*e.g.* open wings), while the other model struggles.

4.2. Learning with Web Images

We go beyond standard semi-supervised approaches that use unlabeled images from curated datasets by effectively utilizing images from the web for training. We use $N_S = 150$ images annotated with 2D keypoints and evaluate our proposed approach.

Datasets. We train our models with 2D keypoints from Pascal [9] using the same train/test splits as prior work [21, 22]. For categories with less than 150 labeled samples, we augment the labeled training set $\mathcal{S}$ by manually labeling some images from COCO [26]. We use Pascal's test split and the Animal Pose dataset [3] for evaluation.

Implementation details. We train the *SimpleBaselines* [38] 2D pose estimator using 150 images with 2D keypoint labels and generate pseudo-labels on web images. For criterion CF-MT, we use scaling and rotation transformations. For criterion CF-CM, we train Stacked HourGlass [30] as the auxiliary 2D pose estimator. We use ACSM [21] for 3D shape prediction. We found that some template meshes provided in the publicly available ACSM codebase contained incorrect keypoint definitions for certain animals' limbs, with inconsistencies between the right and left sides. After rectifying this issue, we trained ACSM exclusively with keypoint labels and achieved better performance than the one reported in [21], establishing a stronger baseline. We denote our implementation of ACSM trained with 150 images labeled with 2D keypoints as *ACSM-ours*.

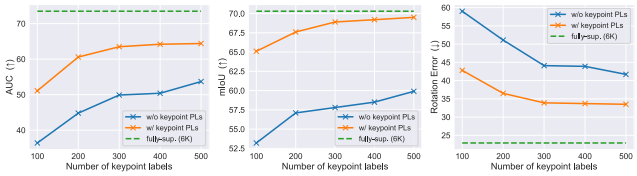

Figure 4. **Evaluation on CUB.** We show the AUC, mIoU and camera rotation error (in degrees) for models trained with different number of labeled and pseudo-labeled instances. Performance across all metrics is significantly improved with keypoint PLs.

Baselines. We compare the performance of the models trained with selected web images and pseudo-labels with the fully-supervised baseline (*ACSM-ours*) as well as with the baseline that uses all pseudo-labels during training, denoted as *ACSM-ours+KP-all*. To put our results into perspective, we also compare with the models from ACSM [21]. The available models are trained using mask labels from Pascal [9] and manually picked mask pseudo-labels from ImageNet [7]. For horses, the authors provide a model trained with keypoint labels from Pascal as well.

Results on Pascal. First, we evaluate the performance of different methods on Pascal's test set. The results are pre-

		Horse		Cow		Sheep	
		AUC (↑)	err_R (↓)	AUC (↑)	err_R (↓)	AUC (↑)	err_R (↓)
	ACSM (Mask) [21]	47.4 (↓19.3)	38.0 (↑17.3)	46.2 (↓13.4)	51.8 (↑22.3)	31.4 (↓25.8)	65.6 (↑50.0)
	ACSM (KP+Mask) [21]	51.0 (↓15.7)	59.9 (↑39.2)	-	-	-	-
	ACSM-ours	66.7	20.7	59.6	29.5	57.2	15.6
	ACSM-ours + KP-all	69.1 (↑2.4)	20.9 (↑0.2)	61.2 (↑1.6)	27.8 (↓1.7)	29.1 (↓28.1)	19.9 (↑4.3)
$N = 1K$	ACSM-ours + KP-conf	72.9 (↑6.2)	19.6 (↓1.1)	65.0 (↑5.4)	31.9 (↑2.4)	58.5 (↑1.3)	20.3 (↑4.7)
$N = 1K$	ACSM-ours + CF-MT	74.6 (↑7.9)	19.9 (↓0.8)	67.0 (↑7.4)	26.2 (↓3.3)	59.3 (↑2.1)	**14.6 (↓1.0)**
$N = 1K$	ACSM-ours + CF-CM	**75.1 (↑8.4)**	19.6 (↓1.1)	67.2 (↑7.6)	25.9 (↓3.6)	59.3 (↑2.1)	14.7 (↓0.9)
$N = 1K$	ACSM-ours + CF-CM^2	73.0 (↑6.3)	19.7 (↓1.0)	67.2 (↑7.6)	**23.5 (↓6.0)**	58.7 (↑1.5)	15.3 (↓0.3)
$N = 3K$	ACSM-ours + KP-conf	74.3 (↑7.6)	19.6 (↓1.1)	67.8 (↑7.8)	28.0 (↓1.5)	57.3 (↑0.1)	15.3 (↓0.3)
$N = 3K$	ACSM-ours + CF-MT	74.4 (↑7.7)	19.6 (↓1.1)	66.0 (↑6.4)	29.7 (↑0.2)	59.6 (↑2.4)	15.1 (↓0.5)
$N = 3K$	ACSM-ours + CF-CM	74.1 (↑7.4)	19.6 (↓1.1)	66.4 (↑6.8)	29.3 (↓0.2)	59.4 (↑2.2)	15.8 (↑0.2)
$N = 3K$	ACSM-ours + CF-CM^2	73.8 (↑7.1)	**19.5 (↓1.2)**	**69.6 (↑10.0)**	25.2 (↓4.3)	**60.2 (↑3.0)**	14.9 (↓0.7)

Table 2. **Evaluation on Animal Pose.** N is the number of selected images from the web. ACSM-ours + KP-all uses all available web images. Training with selected web images significantly improves the fully-supervised baseline. We show the performance change from ACSM-ours in green/red.

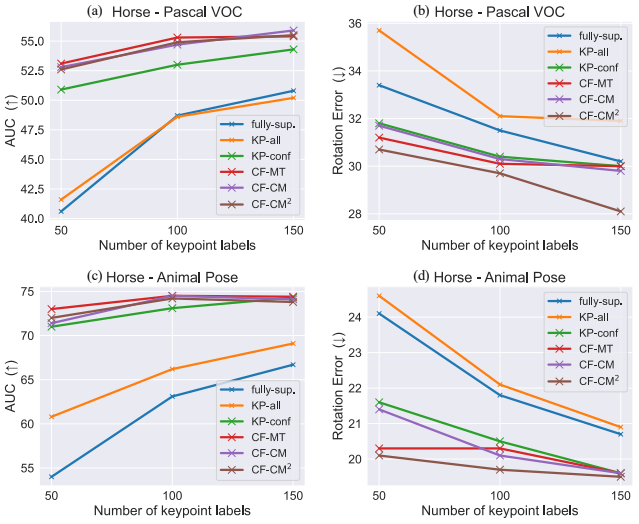

Figure 5. **Evaluation under varying initial supervision.** We show the AUC and camera rotation error (in degrees) on Pascal (top) and Animal Pose (bottom) for the use of our proposed framework under various levels of initial supervision from $\mathcal{S}$. KP-conf, CF-MT, CF-CM, CF-CM^2 use PLs from $N = 3K$ images.

sented in Table 1. From Table 1 we observe that naively including all pseudo-labels during training (ACM-ours+KP-all) leads in degraded performance. Data selection with KP-conf leads to some improvements, but models trained with pseudo-labels selected from consistency-filtering, *i.e.* CF-MT, CF-CM and CF-CM^2, perform consistently better. The improvement over the fully-supervised baseline is quite substantial for those models. For instance, ACSM-ours+CF-CM^2 has a relative improvement (in AUC) of **10.9%** over ACSM-ours and **17.0%** over ACSM-ours+KP-all, averaged across all categories shown in Table 1.

		Giraffe		Bear	
		AUC (↑)	err_R (↓)	AUC (↑)	err_R (↓)
	ACSM-ours	79.1	34.7	59.2	**28.9**
	ACSM-ours + KP-all	75.6 (↓3.5)	39.5 (↑4.8)	58.2 (↓1.0)	35.4 (↑6.5)
$N = 1K$	ACSM-ours + KP-conf	83.3 (↑4.2)	31.4 (↓3.3)	62.7 (↑3.5)	31.0 (↑2.1)
$N = 1K$	ACSM-ours + CF-MT	84.4 (↑5.3)	32.1 (↓2.6)	62.6 (↑3.6)	31.9 (↑3.0)
$N = 1K$	ACSM-ours + CF-CM	84.1 (↑5.0)	32.5 (↓2.2)	62.8 (↑3.6)	31.6 (↑2.7)
$N = 1K$	ACSM-ours + CF-CM^2	**86.6 (↑7.5)**	**28.8 (↓5.9)**	62.2 (↑3.0)	30.7 (↑1.8)
$N = 3K$	ACSM-ours + KP-conf	82.8 (↑3.7)	35.6 (↑0.9)	59.9 (↑0.7)	35.0 (↑6.1)
$N = 3K$	ACSM-ours + CF-MT	83.5 (↑4.4)	33.1 (↓1.6)	63.1 (↑3.9)	31.7 (↑2.8)
$N = 3K$	ACSM-ours + CF-CM	83.6 (↑4.5)	33.4 (↓1.3)	**63.4 (↑4.2)**	31.5 (↑2.6)
$N = 3K$	ACSM-ours + CF-CM^2	84.5 (↑5.4)	32.8 (↓1.9)	62.8 (↑3.6)	32.3 (↑3.4)

Table 3. **Evaluation on COCO.** N is the number of selected images from the web. ACSM-ours + KP-all uses all available web images. Performance change from the fully-supervised baseline is shown in green/red.

Results on Animal Pose. Next, we evaluate our previously trained models on Animal Pose dataset and present results in Table 2. Results are consistent with those in Pascal. Training models with all pseudo-labels results in large performance degradation in some cases (see Sheep in Table 2). KP-conf improves over the supervised baseline, but the improvement is higher for CF-MT, CF-CM and CF-CM^2. Results from Tables 1 & 2 suggest that consistency-filtering is more effective than confidence-based filtering in our setting.

Results on COCO. In this part, we demonstrate the practicality of our approach in predicting the 3D shape of object categories even in the absence of initial annotations for those categories. We evaluate our approach on giraffes and bears. We annotate 250 images from COCO with 2D keypoints in a process that takes only ≈2.5 hours per-category. Given a mesh template, we only need to associate the 2D

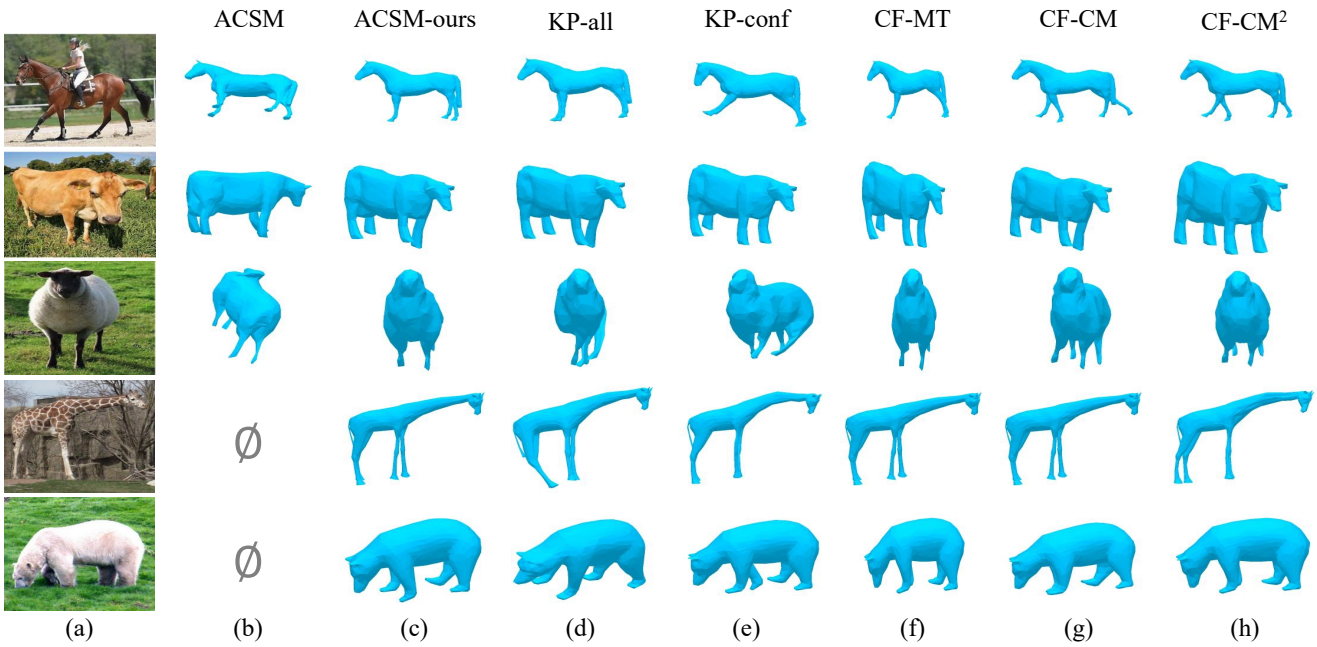

Figure 6. **Sample results on 3D shape recovery of quadrupeds.** For each input image, we show the predicted 3D shape from the inferred viewpoint. (b) original ACSM (models for giraffes and bears are not available), (c) our ACSM implementation, (d) ACSM-ours trained with all PLs, (e) ACSM-ours + KP-conf, (f) ACSM-ours + CF-MT, (g) ACSM-ours + CF-CM, (h) ACSM-ours + CF-CM^2.

keypoint labels with vertices from the 3D mesh. This process is done in meshlab in ≈5 minutes. We train our models with semi-supervised learning as described in Section 3.2. We use 150 images for training, while the remaining images are used for testing. We compare models that utilize pseudo-labels with the fully-supevised baseline trained with 150 images. The results are presented in Table 3. The results are consistent with those on Pascal and Animal Pose. Again, using all pseudo-labels leads in degraded performance. Data selection is essential, and in most cases leads to improved performance compared to the fully-supervised baseline. The performance gains are higher for consistency-based filtering approaches.

Results with different number of labels. In Figure 5, we present results on 3D reconstruction of horses with different number N_s of images labeled with 2D keypoints. For all experiments, we use $N = 3K$ images with pseudo-labels from $\mathcal{U}$ and vary only the number of initial labels N_s from $\mathcal{S}$. From Figure 5, we observe that models with keypoint pseudo-labels outperform the fully-supervised baseline even when N_s is just 50. More importantly, we notice that models trained with $N_s = 50$ and keypoint pseudo-labels outperform fully-supervised models with $N_s = 150$. We also observe, that consistency-based filtering methods are more effective than filtering with keypoint confidence scores. Finally, while all consistency-based filtering approaches lead to similar AUC performance, we notice that CF-CM^2 leads to lower camera rotation errors in all cases.

Qualitative Examples. In Figure 6, we show sample 3D reconstructions for all models. We observe that ACSM-ours recovers more accurate shapes than the original ACSM version. Training with all PLs results in unnatural articulations in the predicted shapes. Additionally, we observe that some articulations are captured only when training with selected samples through consistency-filtering (*e.g.* horse legs in the first row for CF-CM^2).

5. Summary

In summary, we investigated the amount of supervision necessary to train models for 3D shape recovery. We showed that is possible to learn 3D shape predictors with as few as 50-150 images labeled with 2D keypoints. Such annotations are quick and easy to acquire, taking less than 2 hours for each new object category, making our approach highly practical. Additionally, we utilized automatically acquired web images to improve 3D reconstruction performance for several articulated objects, and found that a data selection method was necessary. To this end, we evaluated the performance of four data selection methods and found that training with data selected through consistency-filtering criteria leads to better 3D reconstructions.

Acknowledgements: This research has been partially funded by research grants to D. Metaxas through NSF IUCRC CARTA-1747778, 2235405, 2212301, 1951890, 2003874.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014. 2

[2] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop. In *ECCV*, pages 195–211. Springer, 2020. 3

[3] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *CVPR*, pages 9498–9507, 2019. 3, 6

[4] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *AAAI*, volume 35, pages 6912–6920, 2021. 2

[5] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, pages 2613–2622, 2021. 2

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 2, 6

[8] Xuanyi Dong and Yi Yang. Teacher supervises students how to learn from partially labeled images for facial landmark detection. In *ICCV*, pages 783–792, 2019. 3

[9] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 2, 4, 6

[10] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *ECCV*, pages 88–104, 2020. 2, 5

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 2, 5

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, June 2016. 3, 5

[13] 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 TPAMI*, 36(7):1325–1339, 2013. 2

[14] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 1, 2

[15] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, pages 371–386, 2018. 1, 2, 3, 5

[16] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, pages 3907–3916, 2018. 3

[17] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, pages 11127–11137, 2021. 2

[18] Filippos Kokkinos and Iasonas Kokkinos. Learning monocular 3d reconstruction of articulated categories from motion. In *CVPR*, pages 1737–1746, 2021. 2, 5

[19] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, pages 2252–2261, 2019. 1, 2

[20] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 2

[21] Nilesh Kulkarni, Abhinav Gupta, David F Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *CVPR*, pages 452–461, 2020. 2, 3, 5, 6, 7

[22] Nilesh Kulkarni, Abhinav Gupta, and Shubham Tulsiani. Canonical surface mapping via geometric cycle consistency. In *ICCV*, pages 2202–2211, 2019. 2, 5, 6

[23] Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *CVPR*, pages 1482–1491, 2021. 2, 3

[24] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *ECCV*, pages 677–693, 2020. 2

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 5

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 2, 6

[27] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, pages 506–516. IEEE, 2017. 2

[28] Olga Moskvyak, Frederic Maire, Feras Dayoub, and Mahsa Baktashmotlagh. Semi-supervised keypoint localization. *ICLR*, 2021. 3

[29] Jiteng Mu, Weichao Qiu, Gregory D Hager, and Alan L Yuille. Learning from synthetic animals. In *CVPR*, pages 12386–12395, 2020. 2, 3

[30] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499, 2016. 1, 4, 6

[31] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 2

[32] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *CVPR*, pages 4119–4128, 2018. 3

[33] Nadine Rüegg, Silvia Zuffi, Konrad Schindler, and Michael J Black. Barc: Learning to regress 3d dog shape from images by exploiting breed information. In *CVPR*, pages 3876–3884, 2022. 3

[34] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 1, 4

[35] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 5

[36] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 5

[37] Li Weng and Bart Preneel. A secure perceptual hash algorithm for image content authentication. In *Communications and Multimedia Security*, pages 108–121, 2011. 5

[38] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, pages 466–481, 2018. 1, 4, 5, 6

[39] Rongchang Xie, Chunyu Wang, Wenjun Zeng, and Yizhou Wang. An empirical study of the collapsing problem in semi-supervised 2d human pose estimation. In *ICCV*, pages 11240–11249, 2021. 3

[40] Yinghao Xu, Fangyun Wei, Xiao Sun, Ceyuan Yang, Yujun Shen, Bo Dai, Bolei Zhou, and Stephen Lin. Cross-model pseudo-labeling for semi-supervised action recognition. In *CVPR*, pages 2959–2968, 2022. 2

[41] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *CVPR*, pages 6365–6373, 2017. 3