Unsupervised 3D Structure Inference from Category-Specific Image Collections

Weikang Wang Dongliang Cao Florian Bernard
University of Bonn, Germany

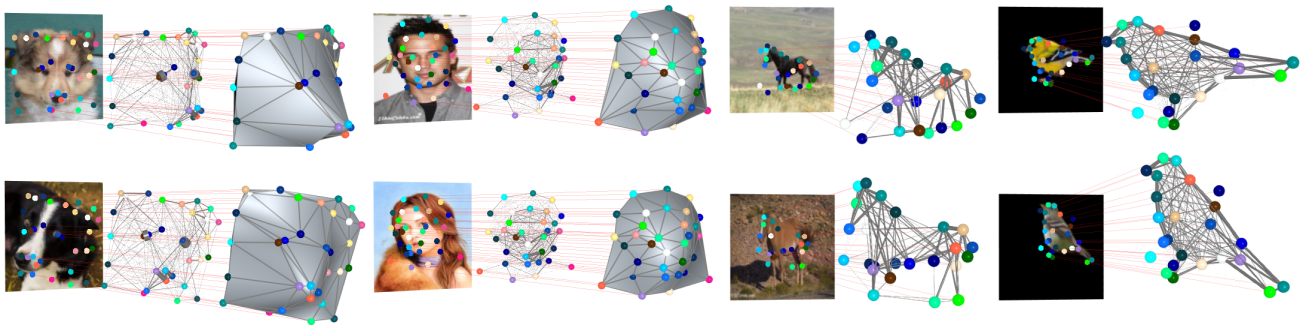

Figure 1. Our method detects consistent and reasonable 3D structure for general object categories from a single 2D image. For training, our method only requires a collection of images from the same category without any additional information.

Abstract

Understanding 3D object structure from image collections of general object categories remains a long-standing challenge in computer vision. Due to the high relevance of image keypoints (e.g. for graph matching, controlling generative models, scene understanding, etc.), in this work we specifically focus on inferring 3D structure in terms of sparse keypoints. Existing 3D keypoint inference approaches rely on strong priors, such as spatio-temporal consistency, multi-view images of the same object, 3D shape priors (e.g. templates, skeleton), or supervisory signals e.g. in the form of 2D keypoint annotations. In contrast, we propose the first unsupervised 3D keypoint inference approach that can be trained for general object categories solely from an inhomogeneous image collection (containing different instances of objects from the same category). Our experiments show that our method not only improves upon unsupervised 2D keypoint inference, but more importantly, it also produces reasonable 3D structure for various object categories, both qualitatively and quantitatively.

1. Introduction

Understanding 3D object structure from 2D images is a fundamental problem in computer vision that has been studied for decades. A range of different 3D structure representations have been considered, including keypoints [14, 25, 59, 69], meshes [55, 56, 62], and voxels [35, 45, 47]. Among these, keypoint-based 3D structure representations have the strong advantage that they are simple and flexible, while at the same time being highly relevant for diverse downstream tasks. However, obtaining 3D keypoints from 2D images is hard due to unknown depth, pose variations, different appearance, and the lack of explicit geometric information. Due to these difficulties, prior works mostly focus on 2D keypoints detection instead of 3D [9, 12, 31, 36, 66].

In this work we investigate whether it is possible to directly obtain 3D keypoints without using any explicit supervision. Our work is motivated by the fact that 2D image keypoints are observations of points in 3D space that are projected onto the 2D image plane. By explicitly understanding and inferring 3D geometry we are able to improve upon the quality of detected keypoints. Specifically, we experimentally demonstrate that projecting our predicted 3D keypoints onto the image plane leads to more reliable 2D keypoints compared to existing 2D keypoint predictors. Moreover, our method can serve the increasing amount of tasks that rely on 3D keypoints, e.g. 3D structure-aware generative models [58], 3D avatar generation [34], or 3D reconstruction [55, 56]. Despite keypoint prediction being a well-studied problem in the 2D domain, 3D keypoint inference is under-explored and we thus specifically contribute

towards this area.

Existing works on inferring 3D keypoints either use 3D point clouds or 2D images/videos from the same object instance as input data. While the former is conceptually a subset selection problem, in this work we solely focus on the more challenging latter case. Manually obtaining 3D keypoints in image data is much harder than annotating 2D keypoints. Thus, there are only very few datasets with 3D keypoints annotations, which in turn motivates unsupervised learning strategies that do not require expensive or even infeasible 3D annotations for training. Existing works that address 3D keypoint inference from image data use for example 2D keypoint annotations [38, 44], multiple views with known poses [11, 35, 48], spatial-temporal consistency in videos [9, 15], 2D/3D skeleton [55, 56] or templates [14]. In stark contrast, in this work we relax the assumptions about the availability of such supervisory signals and train a 3D keypoint predictor merely from category-specific image collections as input. Our method builds upon the assumption that object instances from the same category should have a similar 3D structure with moderate intra-class variations (e.g. faces, horses, birds). Table 1 summaries the differences among existing methods and Fig. 2 gives a visualization of some additional information within these methods and the comparison to our setting.

Methods	Unsup.	w/o Prior	Cat.-agnostic
Bulat et al. [3]	✗	✓	✓
Gou et al. [10]	✗	✓	✓
Zhao et al. [68]	✗	✓	✓
Tulyakov et al. [50]	✗	✓	✓
Pavlakos et al. [41]	✗	✓	✗
Zhang et al. [66]	✗	✓	✗
He et al. [14]	✗	✗	✓
Mildenhall et al. [35]	✓	✗	✓
Wu et al. [54]	✓	✗	✓
Moniz et al. [37]	✓	✗	✗
Supasorn et al. [48]	✓	✗	✓
Novotny et al. [38]	✓	✗	✓
Park et al. [40]	✓	✗	✓
Chen et al. [5]	✓	✗	✓
Reddy et al. [44]	✓	✗	✓
Zhou et al. [69]	✓	✓	✗
Dundar et al. [9]	✓	✓	✗
Xu et al. [59]	✓	✓	✗
Honari and Fua [15]	✓	✗	✓
Ours	✓	✓	✓

Table 1. In comparison to existing 3D keypoints inference methods, our approach combines a unique set of desirable properties: it can be trained in an **unsupervised** manner without requiring any additional information; it does **not require any shape prior** and can thus be applied to **different object categories**.

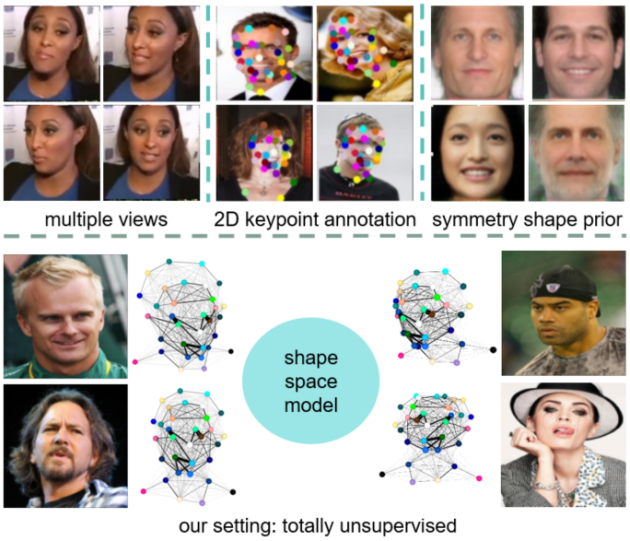

Figure 2. (Top row from left to right) Existing methods require different additional information for 3D structure inference, e.g. multiple views (from [12]), 2D keypoint annotations (from [12]), and a symmetry prior (from [54]). (Bottom row) Our setting does not require such information but instead also infers a 3D shape space model that is learned in an unsupervised manner.

Our method enables to infer 3D keypoints from category-specific image collections without relying on additional prior information. To this end, we build upon the recent unsupervised 2D keypoint detection method by He et al. [12] and extend it by additionally learning a sparse 3D shape space model to represent 3D structure. To regularise the 3D structure, a self-supervised re-projection loss is introduced to couple 3D keypoints with corresponding 2D keypoints. By combining our 3D shape space model, re-projection loss, and additional regularisation loss terms, we demonstrate that our proposed method not only improves 2D keypoint accuracy, but also produces reasonable 3D structure. We summarise our main contributions as:

- For the first time we obtain 3D keypoints from general category-specific image collections without relying on additional prior information.
- Our method obtains reasonable 3D keypoints based on our learnable 3D shape space model that is fused with a state-of-the-art unsupervised 2D keypoint detector.
- We demonstrate that our method both improves unsupervised 2D keypoint detection accuracy and obtains reasonable 3D shape on different benchmarks.

2. Related Work

In the following we summarize relevant 2D and 3D keypoint prediction methods.

2.1. 2D Keypoint Detection

Supervised methods. Various domain-specific supervised 2D keypoint detection methods have been proposed that are trained on a large amount of labeled datasets, e.g. for faces [2, 57, 71], human bodies [1, 4, 18, 21, 32], or vehicles [43], etc. However, labeling keypoints for images is time-consuming, laborious and error-prone, especially when there are occlusions, extreme pose variations, or appearance changes. As a consequence, labeling errors will subsequently influence network training and downstream task performance.

Unsupervised methods. Due to these reasons, an increasing amount of attention is paid to unsupervised methods. One popular concept is multi-view geometric consistency. Thewlis et al. [49] detect keypoints by recovering transformations between one image and its deformed counterpart from pre-defined deformations. Without relying on pre-defined deformations, various video-based methods [9, 19, 20, 22, 24, 36, 46] infer keypoints based on a temporal coherence prior. Instead of using multiple views as input, Xu et al. [60] use unpaired input to disentangle appearance and poses by conducting a swapping-reconstruction strategy and encouraging reconstructed images to resemble the original. He et al. [13] and He et al. [11] use a generative adversarial training framework to constrain keypoint locations by interpreting them as latent codes for an image generator. However, collaboratively learning intermediate keypoints and complex adversarial models is unstable and does not scale. Starting from some initial point sets, Mallis et al. [31] filter keypoints by optimizing alternatively between clustering of keypoints features across a dataset, and training a keypoint detector with pseudo-labels offered by cluster centers. However, this method depends heavily on the initial point set. Zhang et al. [67] address single-image keypoint inference using an auto-encoder architecture to interpret keypoints as a latent representation, while reconstructing the input image from the concatenated keypoints and respective pixel feature descriptors. He et al. [12] use a similar idea by representing keypoints as latent features in an auto-encoder manner. However, they reconstruct the input from an edge map created from keypoints combined with random masking of appearance information. In this way, paired input data is not needed to disentangle appearance and poses, which makes this method applicable in wider scenarios.

2.2. 3D Keypoint Inference

Compared to 2D keypoints detection, 3D keypoints detection is an under-explored topic, particularly in unsupervised settings without strong priors.

Supervised methods. Some earlier methods [3, 10, 68] use a two-stage strategy that infers 2D keypoints first, and then predicts depth for these keypoints. These methods not only need 3D keypoints or depth annotations for training, but are also sub-optimal due to their two-stage procedure. Tulyakov et al. [50] generalize a cascaded regression method to 3D for estimating 3D face landmarks in an end-to-end manner from images. Pavlakos et al. [41] generalize 2D heatmaps directly to 3D by inferring 3D keypoints, while Zhang et al. [66] improve this method via a joint voxel and coordinate regression to avoid the curse of dimensionality. However, these methods all needs 3D keypoints annotations, so that their scope of application is limited due to the absence of annotated 3D keypoints in many cases. Reddy et al. [44] propose a trifocal loss to learn (possibly occluded) 3D keypoints from multiple views by using 2D keypoints annotations. Closely related to 3D shape inference, sparse non-rigid structure-from-motion methods also take 2D keypoints as input to infer 3D structure of objects [8, 23, 38–40, 52, 64, 65]. Based on [12], He et al. [14] predict 3D keypoints using dozens of images with 2D annotations together with a skeleton template.

Unsupervised methods. In order to get rid of sub-optimal two-stage approaches or 2D/3D keypoints annotations, some works aim at unsupervised 3D structure inference. Supasorn et al. [48] use paired views and known transformations between them as input to infer 3D keypoints. Chen et al. [5] use multiple views as input to infer 3D keypoints by averaging 2D keypoints and depth values detected in each view. Using a pair of views as input, Honari and Fua [15] propose a two-stream auto-encoder to encode 2D keypoints of each view, and then use triangulation to lift two sets of 2D keypoints to 3D. Various unsupervised 3D human pose estimation methods [25, 59, 69], either use multiple views or video as input, or they rely on some domain-specific losses as constraints.

Different from all mentioned 3D keypoints inference settings, our method does not rely on 2D/3D keypoints annotations, 3D template/priors/skeletons, multiple views/video, or domain-specific 3D structure. Instead, for the first time we infer 3D structure only from category-specific collections of images for a broad range of object categories.

3. Background

Our method builds upon the recent AutoLink method [12] that addresses the task of unsupervised keypoint detection in 2D. In this section we give a brief overview of AutoLink.

AutoLink comprises an auto-encoder with 2D keypoints as bottleneck latent embeddings. For training, it employs a reconstruction loss that compares the input image with a reconstructed image that is obtained based on edge maps and a randomly masked input image.

Formally, given an image $I \in \mathbb{R}^{W \times H \times 3}$, the encoder takes I as input and outputs K heatmaps, denoted as $H_1, H_2, \ldots, H_K$, where $H_i \in [0,1]^{W \times H}$. The i-th heatmap represents the likelihood for each pixel to be the

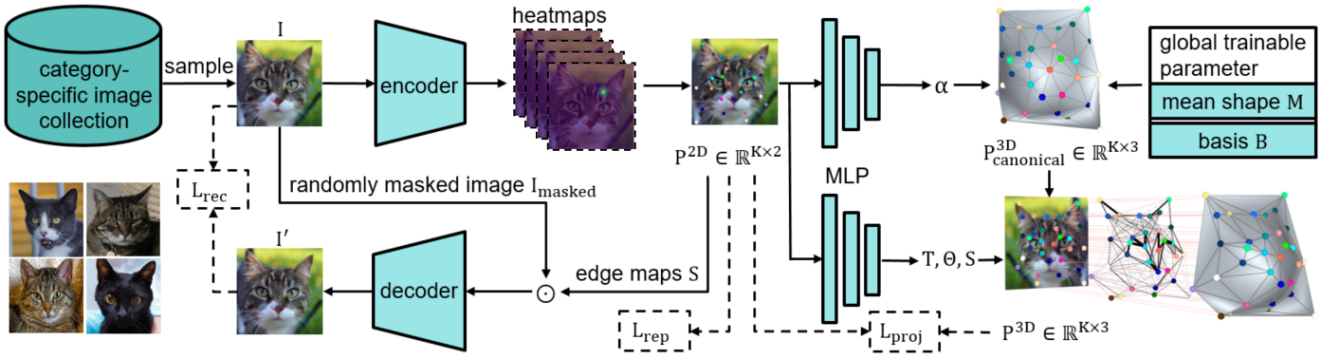

Figure 3. Overview of our pipeline. We sample images I from a category-specific image collection, which are then piped into an encoder that predicts K heatmaps. These heatmaps are used to extract a $K \times 2$ keypoint matrix P^{2D}. From P^{2D}, an edge map S is constructed, which is then, together with a randomly masked version of the input image I_{masked}, fed into the decoder to reconstruct image I'. Moreover, an MLP takes P^{2D} as input and predicts with different heads, the coefficients α, the 3D translation T, and the 3D rotation R. At the same time we optimize for a globally trainable mean shape M and basis B, which define a 3D shape space model used to represent our 3D keypoints, see Eq.(5). The main difference to AutoLink [12], which predicts 2D keypoints only, is that we infer 3D keypoints and a 3D shape space model.

specified keypoint i, where K represents the total number of different keypoints. With that, the location of each 2D keypoint $P_i^{\text{2D}} = (x_i, y_i), i = 1, 2, \ldots, K$ can be represented as

$$P_i^{\text{2D}} = \sum_p \frac{H_i(p)}{\sum_{p'} H_i(p')} p, \tag{1}$$

where $p, p' \in [-1, 1] \times [-1, 1]$ are normalized pixel coordinates.

For each pair of 2D keypoints, P_i^{2D} and P_j^{2D}, $1 \leq i, j \leq K$, the method constructs an edge map $S_{ij} \in \mathbb{R}^{H \times W}$ defined on normalized pixel coordinates p via

$$S_{ij}(p) = \exp(-d_{ij}(p)/\sigma^2), \tag{2}$$

where $d_{ij}(p)$ is the L_2 distance of pixel p to the line segment connecting P_i^{2D} and P_j^{2D}, and $\sigma \in \mathbb{R}$ controls the thickness of the edge. The final edge map $S \in \mathbb{R}^{W \times H}$ summarizing all edges between all keypoints is given by

$$S(p) = \max_{1 \leq i,j \leq K} \omega_{ij} S_{ij}(p), \tag{3}$$

where ω_{ij} are learnable parameters that are shared by all images and optimized during training.

Finally, the edge map S and the (randomly masked) input image I_{masked} are concatenated along the channel axis and fed into a decoder to reconstruct the input. The whole AutoLink pipeline is trained end-to-end only based on the reconstruction loss

$$\mathcal{L}_{\text{rec}} = \|F(D(S \odot I_{\text{masked}})) - F(I)\|_2, \tag{4}$$

where D is the decoder, F is a pre-trained and fixed feature extractor (i.e. VGG network), $\odot$ represents channel-wise concatenation, and I_{masked} is the randomly masked original input image with 80% mask area.

4. Unsupervised 3D Structure Inference

In the following we explain our method. To this end, we introduce our 3D shape space model for inferring 3D keypoint structure in Section 4.1. In section 4.2 we explain estimating the object pose. Section 4.3 introduces our unsupervised losses. Our overall pipeline is summarized in Fig. 3.

4.1. Trainable 3D Shape Space Model

Rather than working with high-dimensional and expensive 3D heatmaps, we infer 3D geometry by training a 2D keypoint predictor together with a 3D shape space model in an end-to-end manner, see Fig. 3. Our encoder first predicts 2D keypoint heatmaps from a given 2D image, which are then transformed into a 2D keypoint matrix using Eq. (1). Then a multi-layer perceptron (MLP) takes the 2D keypoint matrix as input to predict input-dependent coefficients α of a 3D shape model. At the same time, the mean shape M and the set of basis vectors B of this 3D shape space model are optimized as global trainable parameters. Overall, this allows to obtain the 3D keypoints (in canonical pose) as

$$P_{\text{canonical}}^{\text{3D}} = \text{mat}(M + \alpha B), \tag{5}$$

where M is the mean shape of size $1 \times 3K$, B is the basis matrix of size $n \times 3K$ and α are the coefficients of size $1 \times n$, with n denoting the number of basis functions. The operation $\text{mat}()$ reshapes a $1 \times 3K$ row vector to a $K \times 3$ matrix. By training these components simultaneously with our 2D heatmap predictor, we can ensure that during training our 2D heatmap predictor receives gradients that contain explicit information about 3D geometry.

We emphasize that opposed to most existing 3D shape space learning methods that require known keypoint locations, we simultaneously optimize for keypoint locations

while using our trainable 3D shape space model as 3D geometry-aware regularizer. Unlike non-rigid structure-from-motion methods [8, 38–40] that utilize similar ideas, our approach does not require multiple images of the same object from different views. Instead, we use an inhomogeneous collection of category-specific images, where the 3D shape space model serves as flexible and adaptive shape prior that can adjust to the different geometries of objects (within the considered object category). As such, our approach is much more flexible, since it can handle single-view image collections as for example available in web collections.

4.2. Pose Estimation

After getting the 3D keypoints $P^{\text{3D}}_{\text{canonical}}$ (in canonical pose) using Eq. (5), an input-dependent transformation matrix and scaling factor are estimated that transform $P^{\text{3D}}_{\text{canonical}}$ to the posed 3D keypoints P^{3D}, such that the projection of P^{3D} onto the 2D image plane coincides with the 2D structure of the object within the 2D image.

Non-rigid object deformations are taken care of by the 3D shape model, so that we represent the object pose solely via a rigid body transformation matrix $(R, T) \in \text{SE}(3)$. With the 2D keypoint matrix P^{2D} as input, an MLP computes the translation $T \in \mathbb{R}^{1\times 3}$ and the rotation angles $\Theta \in \mathbb{R}^{1\times 3}$, which are converted to a rotation matrix via the Rodrigues formula [30]:

$$R = I_3 + \frac{\sin(\theta)}{\theta}\,[\Theta]_\times + \frac{1-\cos(\theta)}{\theta^2}\,[\Theta]^2_\times\,, \tag{6}$$

where $\theta = \|\Theta\|$ and the skew-symmetric operator $[\,\cdot\,]_\times$ turns a vector in $\mathbb{R}^{1\times 3}$ into a skew-symmetric matrix. Finally the 3D keypoint matrix P^{3D} is given as

$$P^{\text{3D}} = P^{\text{3D}}_{\text{canonical}} R + \mathbf{1}_K T, \tag{7}$$

where $\mathbf{1}_K$ is a K-dimensional vector of all ones.

4.3. Unsupervised Losses

Our re-projection loss penalizes the discrepancy between the projected 3D keypoints and the 2D heatmap keypoint predictions, i.e.

$$\mathcal{L}_{\text{proj}} = \|P^{\text{3D}}\Pi - P^{\text{2D}}\|_F, \quad \text{with} \quad \Pi = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \\ 0 & 0 \end{bmatrix} \tag{8}$$

being a simple orthographic projection (of P^{3D} from Eq. (8)) combined with scaling $s = (s_1, s_2)$, which is predicted by an MLP. This loss ensures consistent 2D keypoint predictions between different images, since the 3D shape structure is shared across the whole image collection via the 3D shape space model.

However, combining $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{proj}}$ alone will lead to degenerate solutions (i.e. all keypoints converge to a single location). In order to avoid this situation, we consider the repulsion loss

$$\mathcal{L}_{\text{rep}} = -\sum_{i=1}^{K} \|p_i - \mathcal{N}(p_i)\|_2 \exp(-\|p_i - \mathcal{N}(p_i)\|_2/h), \tag{9}$$

where p_i is a 2D keypoint (a row of P^{2D}), $\mathcal{N}(p_i)$ is the nearest keypoint (other than p_i) of p_i in P^{2D}, and h is a temperature parameter that controls the repulsion decrease rate. The effect of Eq. (9) is that neighboring keypoints do not move too close to each other.

Our total loss combines all terms and is defined as

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_1 \mathcal{L}_{\text{proj}} + \lambda_2 \mathcal{L}_{\text{rep}}, \tag{10}$$

where λ_1, λ_2 are weights of the individual terms.

5. Experiments

We evaluate our method on various datasets and compare to existing keypoint prediction methods. Implementation details are provided in the supplementary document.

5.1. Experimental Setup

Datasets. We evaluate our method on six datasets: CELEBA WILD [28], CUB-200-2011 [51], 300W-LP [71], AFLW2000-3D [71], HORSE [70] and AFHQ [6]. The CELEBA WILD dataset [28] contains celebrity faces in in-the-wild environments, along with 2D keypoint annotations for each image. Following the setting of He et al. [12], we remove the images whose face covers less than 30% of the area. The 300W-LP [71] dataset is a large synthetic face dataset containing 61,225 samples across large poses (1,786 from IBUG, 5,207 from AFW, 16,556 from LFPW and 37,676 from HELEN). The AFLW2000-3D dataset was introduced along with the 300W-LP dataset by the same data acquisition method, containing 2,000 facial images, each with 68 3D keypoints annotations. We train on the 300W-LP dataset and test on the AFLW2000-3D dataset to evaluate the quantitative 3D keypoint detection performance. We use subsets from the AFHQ dataset [6], namely CAT, DOG, TIGER, FOX, CHEETAH, WOLF, JAGUAR and LION, to get qualitative 3D structure inference results. CUB-200-2011 consists of 11,788 images of birds. Following the settings in [12, 29], we align all birds facing left, and remove seabirds. The HORSE dataset is extracted from the CycleGAN dataset [70] by removing images with multiple horses and aligning them to face left. CUB-200-2011, HORSE, AFHQ and CELEBA WILD are used to generate qualitative results. We quantitatively evaluate unsupervised 2D keypoint inference using the CELEBA WILD dataset and compare it with the recent state-of-the-art method by He et al. [12].

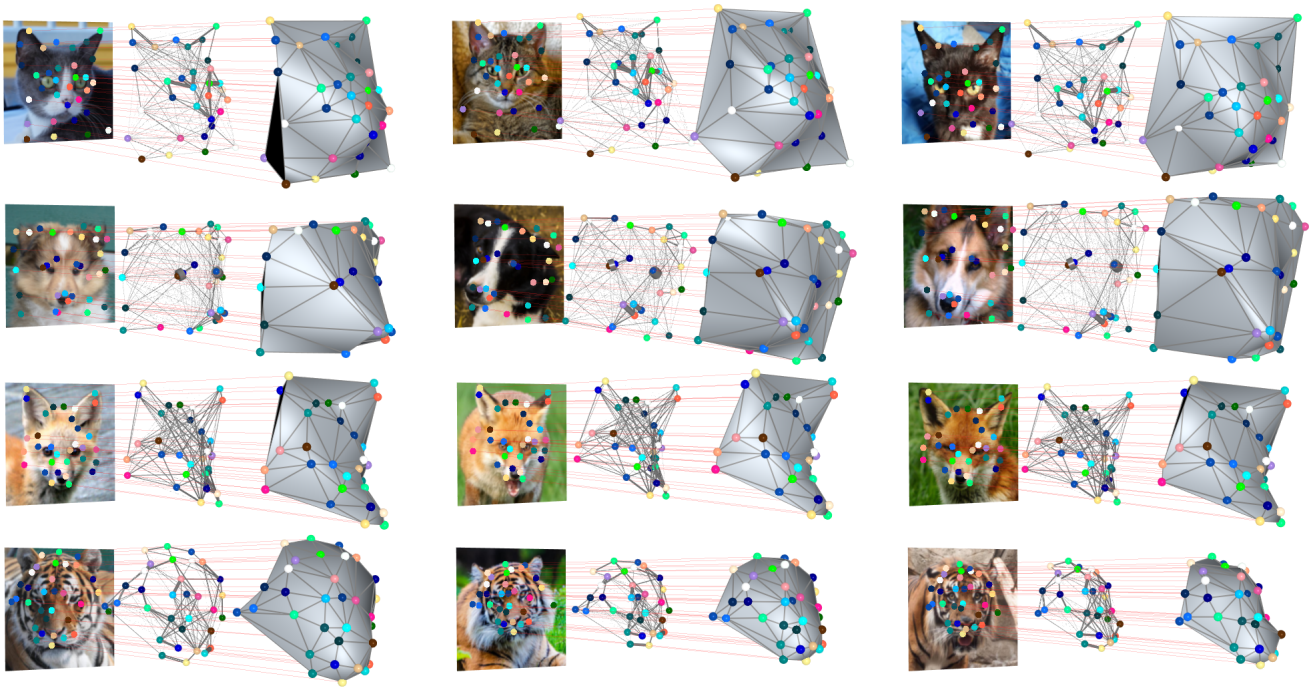

Figure 4. Visualization of the 3D structure for different instances of the AFHQ dataset (CAT, DOG, FOX and TIGER subset in each row) . Each 3D structure is shown in two forms: (i) 3D keypoints with linking edges between them (the edge thickness is proportional to the edge map coefficients ω, which indicate their relative importance); (ii) Delaunay triangulation of the 3D keypoints rendered as shaded mesh. Different keypoints are shown in different colors to demonstrate consistency across different images (columns).

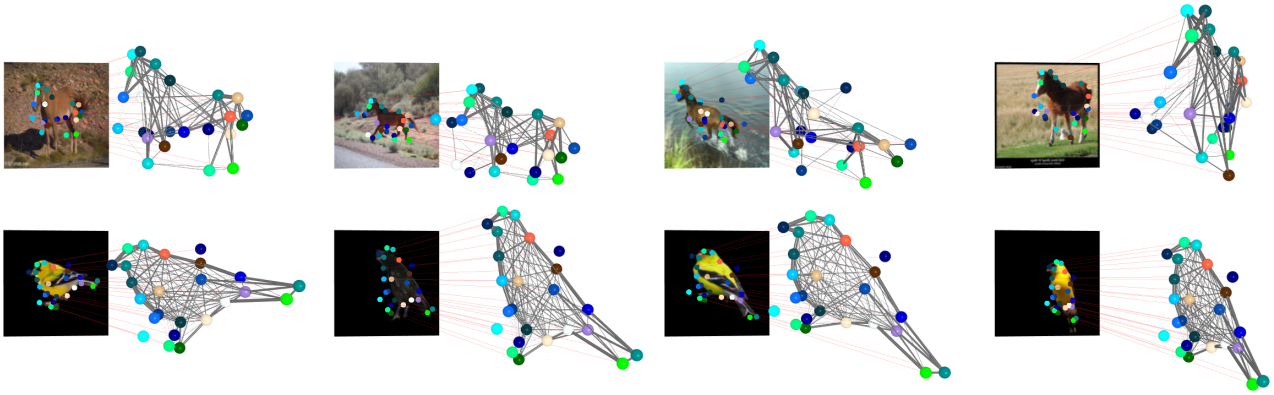

Figure 5. Visualization of the 3D structure for different images from the HORSE and CUB-200-2011 datasets (rows).

Metrics. For unsupervised 2D keypoint inference, we follow the metrics used by He et al. [12], which is the normalized mean L_2 error (NME) of detected 2D keypoints aligned to the ground truth. For the 3D case, we generalize NME to 3D by calculating the normalized mean L_2 of detected 3D keypoints aligned to ground truth 3D keypoints. In both cases, the normalization is the ocular distances of each image.

5.2. Qualitative Results

We visualize the learned 3D structure on four subsets of AFHQ (Cat, Dog, Fox and Tiger; visualization of results of other subsets are in supplementary document), CUB-200-2011, and HORSE datasets in Fig. 4 and 5, respectively. From the visualizations we can directly recognize the 3D structure of each shown object category. For the four subsets of AFHQ dataset, despite these animal categories sharing many similarities, we can recognize the individual animal face characteristics from predicted 3D keypoints with

linking edges. Moreover, we find that the set of thick edges, i.e. the ones with large learned edge coefficient ω, coincide with the most notable structures of each category, while negligible edges remain very thin. Finally, we demonstrate that keypoints are consistently detected among different instances of objects per category, which can be seen based on the consistency of the color of keypoints (both in 2D and 3D). More qualitative results and videos are included in the supplementary document and attached files.

Since there does not exist any other 3D keypoint inference method that is applicable in our unsupervised setting (cf. Table 1), as comparison we consider a simple baseline that first predicts 2D keypoints using AutoLink [12], and then lifts 2D keypoints to 3D by pinpointing learned 2D keypoints on the depth map (given by the pre-trained unsupervised monocular depth estimator MiDaS [42]). Fig. 6 shows a comparison on CELEBA WILD dataset.

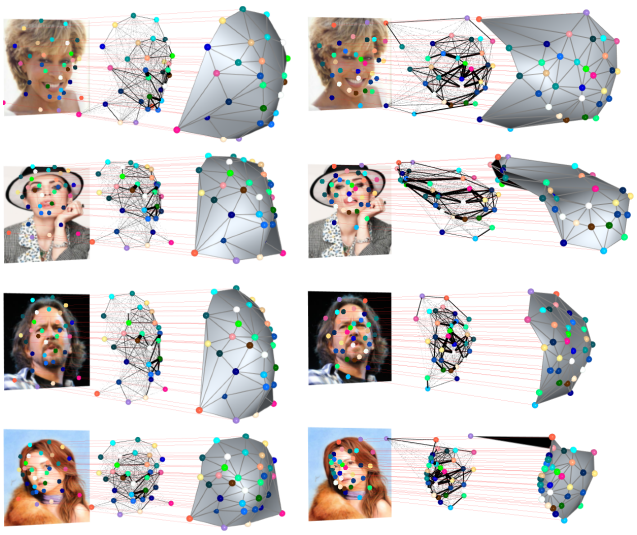

Figure 6. Comparison of predicted 3D structures between our method (left) and AutoLink [12] combined with MiDaS [42] (right). Our method obtains more consistent 2D keypoint predictions and more realistic 3D structure.

5.3. Quantitative Results

2D keypoint inference. First, we show that our method improves upon unsupervised 2D keypoint inference. Table 2 shows the NME values of our method and of various other unsupervised methods on the CELEBA WILD dataset. We also compare to the method by He et al. [12] for various choices of K on the CELEBA WILD dataset to show the robustness of our method, as shown in Table 3. We can see that our method outperforms all competitors across all settings.

Method	**K=8**
DFF [7]	31.30%*
SCOPS (w/o saliency) [17]	22.11%†
SCOPS (w/saliency) [17]	15.01%†
Liu et al. [27]	12.26%†
Huang et al. [16]	8.4%†
GANSeg [13]	6.18%†
Thewlis et al. [49]	31.30%*
Zhang et al. [67]	40.82%*
LatentKeypointGAN [11]	21.90%†
LatentKeypointGAN-tuned [11]	5.63%†
Lorenz et al. [29]	11.41%‡
IMM [19]	8.74%‡
AutoLink [12]	5.39%
Ours	**5.21%**

Table 2. Normalized L_2 error (NME) for 2D keypoints inference of various unsupervised methods on CELEBA WILD datasets for $K = 8$. (* reported from Collins et al. [7]; † reported from He et al. [12]; ‡ reported from Liu et al. [27])

Method	**K=8**	**K=16**	**K=24**	**K=32**
AutoLink [12]	5.39%	4.69%	3.99%	3.77%
Ours	**5.21%**	**3.97%**	**3.54%**	**3.48%**

Table 3. Normalized L_2 error (NME) of AutoLink and our method for different numbers of keypoints using the CELEBA WILD dataset.

Supervised	**Unsupervised**		
3DDFA	AutoLink+Unsup3d	AutoLink+MiDaS	Ours
4.94%	11.47%	9.23%	**8.48%**

Table 4. Normalized L_2 error (NME) of our method, two unsupervised methods (AutoLink + MiDaS and AutoLink + Unsup3d) and one supervised method (3DDFA) for 3D keypoints inference (training on the 300W-LP dataset and testing on the AFLW2000-3D dataset).

3D keypoint inference. As stated in Sec. 5.2, there does not exist any other 3D keypoint inference method that is applicable in our unsupervised setting. So we define two unsupervised baselines by combining AutoLink [12] with MiDaS [42], and Autolink [12] with Unsup3d [54]. Moreover, we compare against the supervised method 3DDFA [71] to understand the gap between unsupervised and supervised methods. Table 4 summarises the NME results, in which our method achieves better performance compared to the two-stage baselines.

6. Ablation Study

In this section we discuss our ablation studies. Results are shown in Fig. 7. The first and second row show the results

with full training losses (i.e. Eq. (10)) with the number of keypoints being 32 and 16, respectively. We can see that with an increasing number of keypoints more detailed structure can be captured by our 3D keypoints. Furthermore, we remove $\mathcal{L}_{\text{rep}}$ from our overall loss to train our framework (third row). The structure collapses into a degenerate keypoint configuration, which confirms the importance of $\mathcal{L}_{\text{rep}}$.

A more detailed analysis on the behavior of $\mathcal{L}_{\text{rep}}$, and on the influence of the number of basis functions n of our shape model, are provided in the supplementary document.

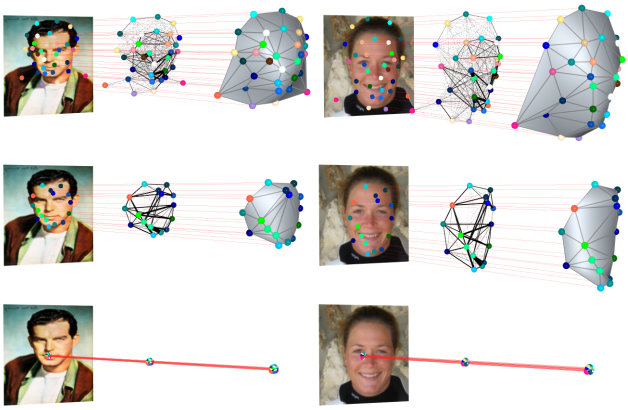

Figure 7. Ablation study on CELEBA WILD dataset. The first and second rows show results for a different number of keypoints (32 and 16). The third row show results without repulsion loss $\mathcal{L}_{\text{rep}}$.

7. Discussion

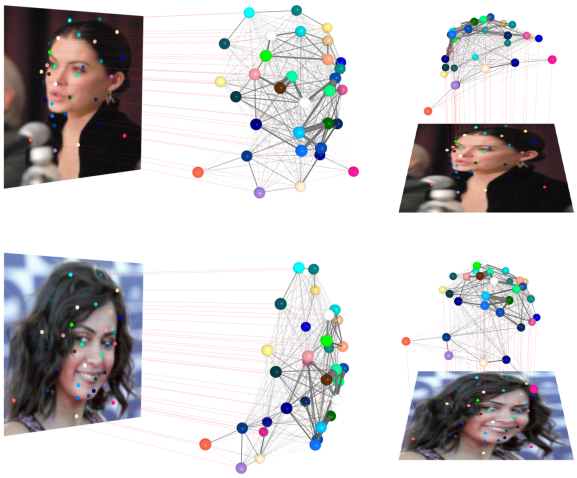

Figure 8. Pose estimation visualization of two instances from CELEBA WILD dataset from two views.

Large pose estimation has always been an important yet challenging issue in many 3D shape inference settings [26, 33, 55, 56, 61, 63]. Recent works have found that when the pose variations falls into a small range, it can be learned directly by a network together with shape [53]. We observe similar results in our setting. For the CELEBA WILD, AFHQ (instances with moderate pose variations), CUB-200-2011 and HORSE (aligned facing direction) datasets, Fig. 4 and 5 have shown multiple instances in different poses with correct pose estimation. Fig. 8 gives more pose estimation results. However, for datasets consisting of instances with large pose variations, our method may fail to optimize for the correct poses. For example, our test on the HUMAN3.6M dataset [18] shows that, even though our model learns plausible human body structure, it fails to optimize for the correct pose of each instance, which also leads to incorrect instance-specific deformations, as shown in Fig. 9 – for such cases with strong articulation, domain-specific approaches that exploit domain knowledge (e.g. about the human kinematic skeleton structure) are at the time being more suitable. Yet, inferring such articulated objects is a highly relevant and interesting direction for future work.

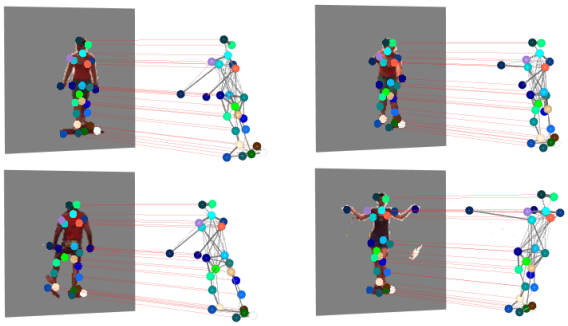

Figure 9. Incorrect pose estimation visualization of HUMAN3.6M dataset. For severely articulated objects additional domain knowledge is currently needed (e.g. a human body kinematic skeleton).

8. Conclusion

Overall, our method is the first unsupervised approach that can infer 3D keypoints from category-specific image collections without the need of any additional prior information. Conceptually, our main novelty lies in simultaneously learning 3D keypoints along with a 3D shape space model from an inhomogeneous collection of category-specific images (e.g. opposed to homogeneous collections of multi-view images or videos, which show the *same* object from different views or in different poses). We experimentally demonstrate that our method not only improves upon 2D keypoint accuracy, but also infers plausible 3D structure for various object categories. We expect our work to be useful for diverse downstream tasks, for example related to 3D reconstruction or 3D shape-based generative models.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 3

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030, 2017. 3

[3] Adrian Bulat, Tzimiropoulos, and Georgios. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*, pages 616–624. Springer, 2016. 2, 3

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 3

[5] Boyuan Chen, Pieter Abbeel, and Deepak Pathak. Unsupervised learning of visual 3d keypoints for control. In *International Conference on Machine Learning*, pages 1539–1549. PMLR, 2021. 2, 3

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 5

[7] Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. Deep feature factorization for concept discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 336–352, 2018. 7

[8] Hui Deng, Tong Zhang, Yuchao Dai, Jiawei Shi, Yiran Zhong, and Hongdong Li. Deep non-rigid structure-from-motion: A sequence-to-sequence translation perspective. *arXiv preprint arXiv:2204.04730*, 2022. 3, 5

[9] Aysegul Dundar, Kevin J Shih, Animesh Garg, Robert Pottorf, Anrew Tao, and Bryan Catanzaro. Unsupervised disentanglement of pose, appearance and background from images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3883–3894, 2021. 1, 2, 3

[10] Chao Gou, Yue Wu, Fei-Yue Wang, and Qiang Ji. Shape augmented regression for 3d face alignment. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*, pages 604–615. Springer, 2016. 2, 3

[11] Xingzhe He, Bastian Wandt, and Helge Rhodin. Latentkeypointgan: Controlling gans via latent keypoints. *arXiv preprint arXiv:2103.15812*, 2021. 2, 3, 7

[12] Xingzhe He, Bastian Wandt, and Helge Rhodin. Autolink: Self-supervised learning of human skeletons and object outlines by linking keypoints. *Advances in Neural Information Processing Systems*, 35:36123–36141, 2022. 1, 2, 3, 4, 5, 6, 7

[13] Xingzhe He, Bastian Wandt, and Helge Rhodin. Ganseg: Learning to segment by unsupervised hierarchical image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1225–1235, 2022. 3, 7

[14] Xingzhe He, Gaurav Bharaj, David Ferman, Helge Rhodin, and Pablo Garrido. Few-shot geometry-aware keypoint localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21337–21348, 2023. 1, 2, 3

[15] Sina Honari and Pascal Fua. Unsupervised 3d keypoint estimation with multi-view geometry. *arXiv preprint arXiv:2211.12829*, 2022. 2, 3

[16] Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8662–8672, 2020. 7

[17] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 869–878, 2019. 7

[18] 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. 3, 8

[19] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. *Advances in neural information processing systems*, 31, 2018. 3, 7

[20] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8787–8797, 2020. 3

[21] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 196–214. Springer, 2020. 3

[22] Yunji Kim, Seonghyeon Nam, In Cho, and Seon Joo Kim. Unsupervised keypoint learning for guiding class-conditional video prediction. *Advances in neural information processing systems*, 32, 2019. 3

[23] Chen Kong and Simon Lucey. Deep non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1558–1567, 2019. 3

[24] Tejas D Kulkarni, Ankush Gupta, Catalin Ionescu, Sebastian Borgeaud, Malcolm Reynolds, Andrew Zisserman, and Volodymyr Mnih. Unsupervised learning of object keypoints for perception and control. *Advances in neural information processing systems*, 32, 2019. 3

[25] Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R Venkatesh Babu, and Anirban Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6152–6162, 2020. 1, 3

[26] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021. 8

[27] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Unsupervised part segmentation through disentangling appearance and shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8355–8364, 2021. 7

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 5

[29] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10955–10964, 2019. 5, 7

[30] Y Ma, S Soatto, J Kosecka, and S Sastry. An invitation to 3d vision: From images to models. *Springer verlag*, 2003. 5

[31] Dimitrios Mallis, Enrique Sanchez, Matthew Bell, and Georgios Tzimiropoulos. Unsupervised learning of object landmarks via self-training correspondence. *Advances in Neural Information Processing Systems*, 33:4709–4720, 2020. 1, 3

[32] 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 *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 3

[33] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6351–6361, 2021. 8

[34] Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. Keypointnerf: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *European conference on computer vision*, pages 179–197. Springer, 2022. 1

[35] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[36] Matthias Minderer, Chen Sun, Ruben Villegas, Forrester Cole, Kevin P Murphy, and Honglak Lee. Unsupervised learning of object structure and dynamics from videos. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3

[37] Joel Ruben Antony Moniz, Christopher Beckham, Simon Rajotte, Sina Honari, and Chris Pal. Unsupervised depth estimation, 3d face rotation and replacement. *Advances in neural information processing systems*, 31, 2018. 2

[38] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7688–7697, 2019. 2, 3, 5

[39] Sungheon Park, Minsik Lee, and Nojun Kwak. Procrustean regression: A flexible alignment-based framework for non-rigid structure estimation. *IEEE Transactions on Image Processing*, 27(1):249–264, 2017.

[40] Sungheon Park, Minsik Lee, and Nojun Kwak. Procrustean regression networks: Learning 3d structure of non-rigid objects from 2d annotations. In *European Conference on Computer Vision*, pages 1–18. Springer, 2020. 2, 3, 5

[41] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017. 2, 3

[42] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 7

[43] N Dinesh Reddy, Minh Vo, and Srinivasa G Narasimhan. Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1906–1915, 2018. 3

[44] N Dinesh Reddy, Minh Vo, and Srinivasa G Narasimhan. Occlusion-net: 2d/3d occluded keypoint localization using graph networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7326–7335, 2019. 2, 3

[45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1

[46] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019. 3

[47] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 1

[48] Suwajanakorn Supasorn, Noah Snavely, Jonathan J Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. *Advances in neural information processing systems*, 31, 2018. 2, 3

[49] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proceedings of the IEEE international conference on computer vision*, pages 5916–5925, 2017. 3, 7

[50] Sergey Tulyakov, László A Jeni, Jeffrey F Cohn, and Nicu Sebe. Viewpoint-consistent 3d face alignment. *IEEE transactions on pattern analysis and machine intelligence*, 40(9): 2250–2264, 2017. 2, 3

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

[52] Chaoyang Wang, Chen-Hsuan Lin, and Simon Lucey. Deep nrsfm++: Towards unsupervised 2d-3d lifting in the wild. In *2020 International Conference on 3D Vision (3DV)*, pages 12–22. IEEE, 2020. 3

[53] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 8

[54] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1–10, 2020. 2, 7

[55] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Dove: Learning deformable 3d objects by watching videos. *International Journal of Computer Vision*, pages 1–12, 2023. 1, 2, 8

[56] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8792–8802, 2023. 1, 2, 8

[57] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2138, 2018. 3

[58] Zhijie Wu, Xiang Wang, Di Lin, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Sagnet: Structure-aware generative network for 3d-shape modeling. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 1

[59] Chenxin Xu, Siheng Chen, Maosen Li, and Ya Zhang. Invariant teacher and equivariant student for unsupervised 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3013–3021, 2021. 1, 2, 3

[60] Yinghao Xu, Ceyuan Yang, Ziwei Liu, Bo Dai, and Bolei Zhou. Unsupervised landmark learning from unpaired data. *arXiv preprint arXiv:2007.01053*, 2020. 3

[61] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Lassie: Learning articulated shapes from sparse image ensemble via 3d part discovery. *Advances in Neural Information Processing Systems*, 35:15296–15308, 2022. 8

[62] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4853–4862, 2023. 1

[63] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021. 8

[64] Haitian Zeng, Yuchao Dai, Xin Yu, Xiaohan Wang, and Yi Yang. Pr-rrn: pairwise-regularized residual-recursive networks for non-rigid structure-from-motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5600–5609, 2021. 3

[65] Haitian Zeng, Xin Yu, Jiaxu Miao, and Yi Yang. Mhr-net: Multiple-hypothesis reconstruction of non-rigid shapes from 2d views. In *European Conference on Computer Vision*, pages 1–17. Springer, 2022. 3

[66] Hongwen Zhang, Qi Li, and Zhenan Sun. Joint voxel and coordinate regression for accurate 3d facial landmark localization. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2202–2208. IEEE, 2018. 1, 2, 3

[67] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2694–2703, 2018. 3, 7

[68] Ruiqi Zhao, Yan Wang, C Fabian Benitez-Quiroz, Yaojie Liu, and Aleix M Martinez. Fast and precise face alignment and 3d shape reconstruction from a single 2d image. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*, pages 590–603. Springer, 2016. 2, 3

[69] Shihao Zhou, Mengxi Jiang, Qicong Wang, and Yunqi Lei. Towards locality similarity preserving to 3d human pose estimation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1, 2, 3

[70] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 5

[71] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 3, 5, 7