

End-to-End Learning of Multi-category 3D Pose and Shape Estimation

Yigit Baran Can¹
yigit.can@vision.ee.ethz.ch
Alexander Liniger¹
alex.liniger@vision.ee.ethz.ch
Danda Pani Paudel¹
paudel@vision.ee.ethz.ch
Luc Van Gool^{1,2}
vangool@vision.ee.ethz.ch

¹ Computer Vision Lab
ETH Zurich
² VISICS, ESAT/PSI
KU Leuven

Abstract

In this paper, we study the representation of the shape and pose of objects using their keypoints. We propose an end-to-end method that simultaneously detects 2D keypoints from an image and lifts them to 3D. The proposed method learns both 2D detection and 3D lifting only from 2D keypoint annotations. In addition to being end-to-end from images to 3D keypoints, our method also handles objects from multiple categories using a single neural network. We use a Transformer-based architecture to detect the keypoints, as well as to summarize the visual context of the image. This visual context information is used while lifting the keypoints to 3D, to allow context-based reasoning for better performance. Our method can handle occlusions as well as a wide variety of object classes. Our experiments on three benchmarks show that our method performs better than the state-of-the-art. Code <https://github.com/ybarancan/end2end3D>.

1 Introduction

A keypoint-based shape and pose representation is attractive because of its simplicity and ease of handling. Example applications include 3D reconstruction [10, 31, 40], registration [20, 26, 27, 52], and human body pose analysis [4, 6, 29, 39], recognition [17, 37], and generation [44, 53]. The keypoints are often detected as 2D image coordinates due to the ease of the corresponding annotation. But in many applications (e.g. augmented reality), both 3D shape and pose are required [47] for the subsequent geometric reasoning tasks.

Estimating keypoints in 3D [42, 43, 47, 55] has two pitfalls: (i) the need of 3D keypoints, pose, or multi-view for supervision; (ii) the lack of direct pose reasoning with respect to a canonical frame. Learning-based methods can provide both 3D keypoints and pose from a single image, making them suitable for applications from scene understanding [13] to augmented reality [28]. Template-based single view methods [58, 51] may also be used to obtain 3D keypoints and pose from 2D keypoints. However, besides requiring templates,

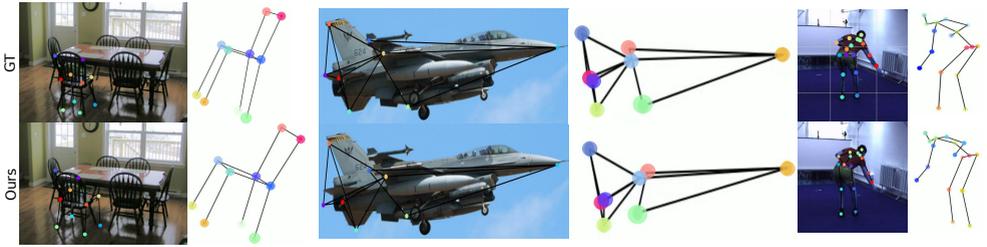


Figure 1: Our method can provide accurate 3D estimations for diverse categories directly from a single image. The 2D keypoints are detected and used with an image-based feature vector to produce 3D estimates.

they are known to be sensitive to self-occlusions [10]. Therefore, we adopt a learning-based method for single view inference of both the 3D keypoints and the pose of objects.

In this paper, we consider that only one image per object is available both during training and inference. This allows us to learn from diverse datasets, such as internet image collections, potentially offering us a high generalization ability. For better scalability, we also assume that only minimalistic supervision in the form of 2D keypoints and objects’ categories are available. Existing methods that learn 3D shape and pose from an image collection by object categories are also known as deep non-rigid structure-from-motion (NrSfM) due to their underlying assumption. The method proposed in this paper also belongs to the same class, which can be divided into single [23, 65, 48, 64] and multi-category [60] methods. Multi-category methods estimate 3D shape and pose of various classes of objects, and are interesting due to two main reasons, (i) computational: one single neural network can infer shapes and poses for objects from different categories; (ii) relational: possibility of establishing/exploiting relationships across categories. This is in contrast to single-subject methods such as [23, 49] where a different model is trained for each test sample. Instead we follow the standard setting and procedure used in [60, 63, 48].

Most existing methods [9, 23, 60, 65, 48, 64] that output pose estimates from images operate in two stages; 2D keypoint extraction followed by 3D shape and pose estimation. These two stages are often performed independently. We argue that these two stages are dependent and can mutually benefit from each other. Thus, 2D keypoints can be extracted while being suitable for the down-streaming task of 3D reasoning. In particular, we extract the visual context information along with the 2D keypoints from the keypoint extraction network. Later, both visual context and 2D keypoints are provided in a differentiable way to the network that lifts 2D keypoints to 3D. Our experiments clearly demonstrate the benefit of visual context information during 3D pose and shape recovery.

We model the 3D shape using a dictionary learning approach, similar to [60], where the shape basis for the union of categories are learned. The instance-wise shape is then recovered with the help of the shape basis coefficients. However, it is known that the size of the shape basis requires careful tuning [60, 65]. In the multi-category setting, the latent space is shared by all categories and each category can have a different optimal shape basis size. Moreover, directly using the shape coefficients results in being over-sensitivity to small perturbations in the input. We show that both problems can be solved with a simple formulation, that sparsifies the shape basis by applying cut-off on the shape coefficients based on a learned threshold vector. This new formulation with a negligible number of additional parameters

allows for a much simpler network compared to sparse dictionary-based networks [23]. The major contributions of our work can be summarized as:

- End-to-end reconstructing of 3D shape and pose in a multiple category setup, using a single neural network.
- We propose to use auxiliary image context information to improve the performance.
- Our method achieves state-of-the-art results in the multi-category setting, with significant improvement.

2 Related Work

The task of lifting 2D keypoints of deformable objects to 3D from a single image has been mostly studied in the context of NrSfM, where the task is to recover the poses and viewpoints from multiple observations in time of an object [0]. Significant research in NrSfM exists such as sparse dictionary learning [22, 56], low-rank constraints [12], union of local subspaces [57], diffeomorphism [32], and coarse-to-fine low-rank reconstruction [0]. It is possible to use NrSfM frameworks to build category-specific models that can learn to estimate pose and viewpoint from a single image by treating the images of the same category as observations of a single object deformed at different time steps [8, 23, 24].

Obtaining the 3D structure of an object from a single image has been studied sparsely. In [19] instance segmentation datasets were used to train a model that outputs 3D meshes given an image. Correspondences between 2D-3D keypoints were also used to improve results [24]. While some recent methods can estimate the viewpoint and non-rigid meshes, these methods work on objects with limited diversity, such as faces [18, 56, 50].

The closest line of work to ours involves building a single model for a diverse set of input classes. C3DPO [30] proposed to learn the factorization of the object deformation and viewpoint change. They propose to enforce the transversal property through a separate canonicalization network that undoes rotation applied on a canonical shape. Park et al. proposed using Procrustean regression [54] to determine unique motions and shapes [55]. They also propose an end-to-end method using a CNN that can output 3D location of human keypoints from the image. However, their method cannot handle multiple object categories or occluded keypoints. Moreover, it requires temporal information in the form of sequences. Human pose estimation is also tackled in [9], where the authors propose a cyclic-loss and discriminator. They further boost their results by using temporal information and additional datasets for the training of their GAN. However, their method is limited to human pose estimation. Recently [43] extended Procrustean formulation with autoencoders and proposed a method that can infer 3D shapes without the need for sequence. However, their method requires a more complex network, two encoders, as well as Procrustean alignment optimization at test-time, which renders the method slow [48]. All these methods accept 2D keypoints as input rather than images and tackle the problem of obtaining 3D keypoint locations from a single image using a separate keypoint detector, such as a stacked hourglass network [46].

3 Multi-category from a Single View

We extract 3D structures in the form of 3D keypoints, given only an image of an object category. During training, we only have access to the 2D location of keypoints and the category label. For simplicity, we separate our solution into two parts: category and 2D

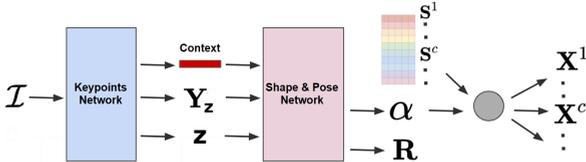


Figure 2: System pipeline. Multi-category from a single view.

keypoints extraction from the image and lifting them to 3D. We will first focus on lifting the given 2D keypoints to 3D, followed by the end-to-end network introducing our 2D keypoint extractor and the tight coupling with the lifter network.

3.1 Multi-category NrSfM

Let $\mathbf{Y}_i = [y_{i1}, \dots, y_{ik}] \in \mathbb{R}^{2 \times k}$ be a stacked matrix representation of k 2D keypoints from the i^{th} view. We represent the structure of the i^{th} view as $\mathbf{X}_i = \alpha_i^T \mathbf{S}$, using the shape basis $\mathbf{S} \in \mathbb{R}^{D \times 3k}$ and coefficients $\alpha_i \in \mathbb{R}^D$. For simplicity, we assume that the keypoints are centered and normalized and that the camera follows an orthographic projection model, represented by $\Pi = [\mathbf{I}_{2 \times 2} \ \mathbf{0}]$. Given the camera rotation matrix $\mathbf{R}_i \in \mathbf{SO}(3)$, as well as the centered and normalized keypoints, we can write $\mathbf{Y}_i = \Pi \mathbf{R}_i (\mathbb{I}_3 \odot \alpha_i^T \mathbf{S})$, where the operation $\mathbb{I}_3 \odot \mathbf{s}$ reshapes the row vector $\mathbf{s} \in \mathbb{R}^{1 \times 3k}$ to a matrix of the form $\mathbb{R}^{3 \times k}$. The recovery of shape and pose by NrSfM given n views can be written as¹,

$$\min_{\alpha_i, \mathbf{S}, \mathbf{R}_i \in \mathbf{SO}(3)} \sum_{i=1}^n \mathcal{L}(\mathbf{Y}_i, \Pi \mathbf{R}_i (\mathbb{I}_3 \odot \alpha_i^T \mathbf{S})). \quad (1)$$

where $\mathcal{L}(a, b)$ is a norm-based loss of the form $\|a - b\|$.

In the context of multi-class NrSfM, our method extracts 3D structures of objects from a wide variety of classes. Thus, $(\mathbb{I}_3 \odot \alpha_i^T \mathbf{S}) \in \mathbb{R}^{3 \times k}$, should be able to express the 3D structure of objects with different number of keypoints. Let \mathbf{Z} represent the set of object categories and $z_i \in \mathbf{Z}$ be the category of sample i . Let each category $z \in \mathbf{Z}$ be represented by k_z keypoints, thus we have a total of $k = \sum_z k_z$ keypoints. To “access” the correct keypoints we have a subset selection vector $\zeta_z \in \{0, 1\}^k$ that indicates which dimensions relate to category z . Given these multi-category definitions, we can reformulate 1 as

$$\min_{\alpha_i, \mathbf{S}, \mathbf{R}_i \in \mathbf{SO}(3)} \sum_{i=1}^n \mathcal{L}(\mathbf{Y}_i \circ \zeta_{z_i}, \Pi \mathbf{R}_i (\alpha_i^T \mathbf{S}) \circ \zeta_{z_i}), \quad (2)$$

where \circ is the broadcasted elementwise multiplication.

In the above formulation, \mathbf{R}_i and α_i are inputs, hence category dependent, while \mathbf{S} is shared among all categories. To formulate the problem as a learning-based approach, let α_i be the output of a function of input \mathbf{Y}_i , i.e. $\alpha(\mathbf{Y}_i)$. Let us separate the function $\alpha(\cdot)$ into two composite functions $\alpha(\mathbf{Y}_i) = g(f(\mathbf{Y}_i))$, with $g(\cdot)$ being an affine function, $g(\mathbf{v}) = \mathbf{W}_g \mathbf{v} + \mathbf{b}_g$ with $\mathbf{v} \in \mathbb{R}^F$, $\mathbf{W}_g \in \mathbb{R}^{D \times F}$, $\mathbf{b}_g \in \mathbb{R}^D$. We do not place any restriction on the function $f(\cdot)$ other than taking some observation \mathbf{Y}_i and outputting a vector of dimension F . Moreover, let us rewrite \mathbf{R}_i as a function of the input y , i.e. $R(y)$. Representing all the parameters with θ , the problem definition becomes

$$\min_{\theta} \sum_i \mathcal{L}(\mathbf{Y}_i \circ \zeta_{z_i}, \Pi(R(\mathbf{Y}_i))([\mathbf{W}_g f(\mathbf{Y}_i) + \mathbf{b}_g] \mathbf{S}) \circ \zeta_{z_i}). \quad (3)$$

¹We will omit \mathbb{I}_3 and the transposition of α throughout the rest of the paper for the ease of notation.

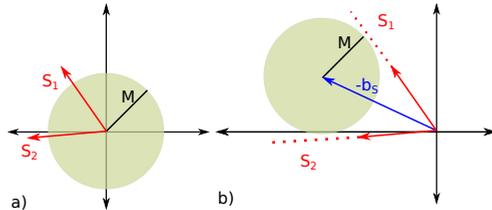


Figure 3: Cut-off coefficients translate the original latent space (a) by b_s to the representation in (b). This enables the latent space vectors (S_1 and S_2) to have non-negative coefficients. The latent space vectors (S_1 and S_2) as well as the translation term b_s are learned from the data while adding negligible number of parameters.

In the above formulation, shape basis coefficients α_i are latent codes with latent space basis vectors W_g and a translation term b_g , which are shared for all categories. Projecting features of objects from different categories into a shared latent space lets the method extract cross-categorical geometric relationships, please refer to Supp for visuals. Moreover, it substantially simplifies computations since we do not require a separate network for each class.

3.2 Cut-off Shape Coefficients

Equation 3 is under-determined unless there are additional constraints imposed on the system. The most common constraint is restricting the dimension D of the shape basis coefficients α_n [4, 5]. However, selecting the optimal cardinality requires careful tuning [6, 7]. Since our method extracts 3D structures of objects from a wide variety of classes, the latent space has to accommodate latent codes from a wide range of inputs. Since most objects share some common characteristics, using different manifolds for each class results in failure to utilize cross-class information and an increase in the complexity of the method. On the other hand, the dimensionality of the optimal manifold is different for each class. Thus, ideally, we would like to automatically select a manifold for each input in a way that maximizes the performance. Note that the optimal manifold selection does not only depend on the object class, and we encourage cross-class rules for manifold assignment.

The manifold selection problem can be posed as integer problem, where given a sample Y_n , the network selects a subset of the basis vectors \mathbf{S} . This can be formalized using a binary selection vector $\mathbb{I}_n \in \{0, 1\}^D$ where $\sum_d \mathbb{I}_n[d] \leq D; \forall n$. Given the the basis coefficients $\beta_n \in \mathbb{R}^D$, the representation of Y_n is $\psi_n = (\mathbb{I}_n \circ \beta_n)\mathbf{S}$. Since this formulation is non-differentiable, we propose a differentiable alternative which we call cut-off coefficients. The idea is to truncate negative shape coefficients to zero allowing the network a differential way to select basis vectors. To gain back the expressiveness of full range shape coefficients we introduce a bias term, which allows the network to learn basis which are suited for non-negative coefficients. This idea is visualized in Fig 3. The latent space in the figure is 2-dimensional and the effect of the proposed formulation is translating the latent space by a vector b_s such that the coefficients of the latent vectors are non-negative for any input. Thus, we arrive at $\hat{\psi}_n = \beta_n \mathbf{S} + b_s$ where $\beta_n \in \mathbb{R}_{\geq 0}^D$ and $b_s \in \mathbb{R}^B$. Note that the non-negativity constraint can be simply implemented using a ReLU based truncation ($\beta_n = \text{ReLU}(\beta'_n)$). Furthermore, we want to highlight that this formulation does not reduce the expressiveness of the network, even if we would not re-optimize the basis, please see Supp for the proof.

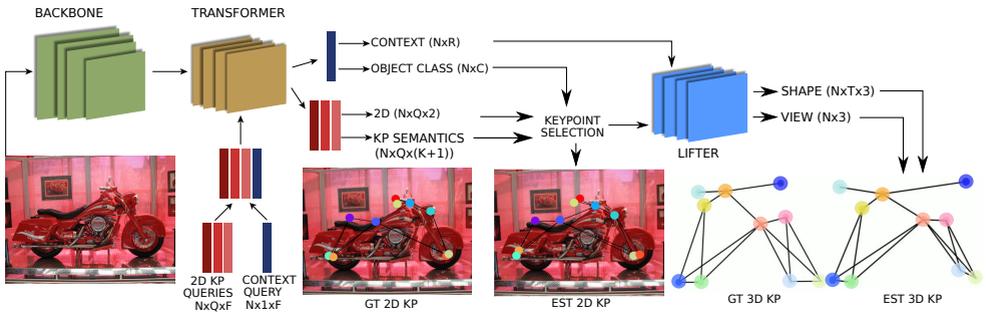


Figure 4: Our method uses a transformer architecture to extract 2D keypoints, object class, and a context vector. The detected keypoints and the context vector are processed by a fully connected network to output 3D keypoints and viewpoint from only a single image. The whole end-to-end training is only supervised by 2D keypoint annotations.

Applying the proposed cut-off coefficient approach to our problem Eq. 3 we get,

$$\mathcal{L}(\mathbf{Y} \circ \zeta, \Pi(R(\mathbf{Y})(\text{ReLU}(W_g f(\mathbf{Y}) + b_g) \mathbf{S} + b_S) \circ \zeta)). \quad (4)$$

Thus, we get the discussed advantages, mainly that our method learns to adaptively pick the active basis vectors, thus selecting the dimensionality of the manifold. Furthermore, all parts including the shape basis vectors \mathbf{S} , the coefficient generating function $W_g f(\cdot) + b_g$ and the bias b_S , are learned from the data. Note that the proposed formulation does not require ISTA iterations [12] which is employed in [23] through a specific encoder-decoder architecture. Moreover, the learnable bias term differentiates our method from sparsity iteration while not changing the expressiveness of the original model.

The proposed formulation has the property of allowing for sparseness, which encourages the representation of the objects in the shape space, i.e. coefficients β to be disentangled [9, 15]. Intuitively, as the number of active (non-zero) coefficients increases, more different combinations of the shape basis vectors \mathbf{S} can arrive at the same solution. By allowing to automatically cut off coefficients, the network can learn a small number of shape coefficients to represent changes from one object to another, thus each coefficient can learn to represent a different major variation. Moreover, the cut-off imposed by the ReLU implies that a small change in the coefficients will, likely, not result in any change in the output if the coefficient is inactive, which improves the robustness of the overall method.

4 End-to-End Learning from Images

The image of an object can be used for more than only 2D keypoint extraction. We propose to detect the 2D keypoints from the image and extract a context vector that can be used in conjunction with the 2D keypoints to obtain a better 3D estimation. The detected 2D keypoints and context vector are used by the lifter network, in our end-to-end trainable pipeline.

4.1 Keypoints from Images

We require a method that can output the locations of an object category dependent pre-defined set of keypoints. Therefore, the problem at hand naturally extends to object classi-

fication. Moreover, in order to fully utilize the image, the keypoint network should produce a context feature representation from the image that can guide the lifter network. Thus, the desired function is $T(\mathcal{I}) = (\mathbf{Y}_z, \mathbf{z}, \rho)$ where \mathcal{I} is the image, \mathbf{Y}_z are the category dependent keypoint locations, \mathbf{z} the category of the object and ρ is the context vector.

We propose to use a DETR-based [14] architecture at the core of function T . Thus, the input image is processed by the backbone (Resnet50 [16]), and the resulting feature map is fed to the transformer. The transformer uses two sets of learned query vectors. The first set is related to keypoints, where each query vector q represents a keypoint. To formalize this, let the maximum number of keypoints among all categories be $\max_z k_z$ be K . Thus, we can extract 2D normalized locations $\delta \in [0, 1]^2$ and the semantics $\omega \in \{0, 1\}^K$ of each keypoint by processing the corresponding query vector using two MLPs. The semantics of a keypoint is category dependent and encoded as a one-hot vector. For example, a given entry in ω can correspond to the front right of a car or the left rear leg of a chair. Entries of ω with indices larger than k_z are zero. The true semantics are denoted by Ω .

To help the lifter network estimate the 3D keypoints, the visual context in the image is important. Thus, we use a second set of learned query vectors, which gets processed by the transformer together with the keypoint queries. The output of the transformer for the context query is then processed by two MLPs. The first outputs a N_ρ dimensional context vector and the second $\hat{\mathbf{z}}$ the one-hot encoded category probability. The category probability is used in conjunction with the keypoint type estimates to obtain the correct 2D keypoint representation, while the context vector is used by the lifter network together with the 2D keypoint representation, see Fig. 4.

We train the network with two supervision signals. First, we perform direct supervision of the 2D keypoints and the category-specific outputs ω and \mathbf{z} , where we use Hungarian matching to select the supervision targets. Second, by training end-to-end, the keypoint extraction network also receives supervision via the lifter network, which helps to learn the lifting and keypoint regression jointly. It is also the indirect supervision signal that guides the learning of the context ρ . This end-to-end connection of the lifter and keypoint extraction network is in sharp contrast to existing papers, which focused on either of the two parts. Our experiments show that the combination of the two can greatly improve performance.

4.2 End-to-end Pipeline

Given the end-to-end joint 2D-3D model, the first step in the training loop is Hungarian matching over the keypoint queries and the GT keypoints. For this, the loss to minimize is given by $\mathcal{L}_H = \mathcal{L}_l + \mathcal{L}_k$ where $\mathcal{L}_l = \|\mathbf{y} - \delta\|_1$ and $\mathcal{L}_k = \mathcal{L}_{CE}(\Omega, \omega)$. The Hungarian matching output provides the set of query vectors that are one-to-one matched to true keypoints. We reformat selected location estimates $\hat{\delta}$ using the matched semantics $\hat{\omega}$ and the category estimate $\hat{\mathbf{z}}$ into the form given in Eq. 2. Let this extracted 2D keypoint representation be $\hat{\mathbf{Y}}$ and the true keypoints be $\bar{\mathbf{Y}}$. Adding the category loss of the keypoint network, we arrive at the following set of losses: Location loss $\mathcal{L}_l = \|\bar{\mathbf{Y}} - \hat{\delta}\|_1$; KP Type loss $\mathcal{L}_k = \mathcal{L}_{CE}(\Omega, \hat{\omega})$; Category loss $\mathcal{L}_b = \mathcal{L}_{CE}(\bar{\mathbf{z}}, \hat{\mathbf{z}})$; Reprojection loss $\mathcal{L}_r = \mathcal{L}(\bar{\mathbf{Y}} \circ \zeta, \Pi R(\hat{\mathbf{Y}}, \rho) f(\hat{\mathbf{Y}}, \rho) \circ \zeta)$. We use Huber loss for \mathcal{L}_r . For the total loss, the different terms are combined using hyperparameters.

During evaluation, where we cannot use Hungarian matching, we first get the object category estimate $\hat{\mathbf{z}}$. Then, for each keypoint type defined for that category, we take the location of the most likely proposal and convert it into the form given in Eq. 2 to obtain $\hat{\mathbf{y}}$. The combination of $\hat{\mathbf{y}}$ and the context vector is processed by the lifter network to output 3D pose and view. The lifter network is given in Fig. 5. The architecture is designed to

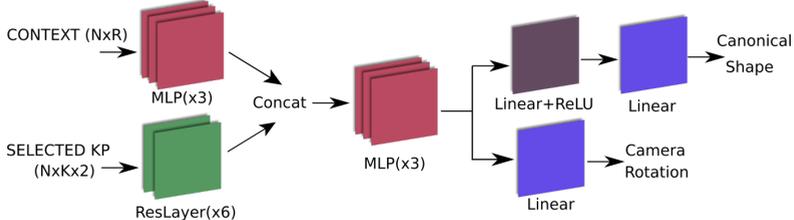


Figure 5: Our lifter architecture combines context vector with the estimated keypoints to produce improved pose estimates.

allow for easy pre-training. We first pre-train the transformer to estimate the locations of 2D keypoints and feed the lifter network true 2D locations alongside the context vector. After that we end-to-end train the whole method.

5 Experiments

5.1 Baselines

We experiment on the **Synthetic Up3D (S-Up3D)**, **PASCAL3D+** and **Human3.6M** datasets. For all datasets, we use the pre-processed versions of [31]. There are only a few NrSfM methods that can handle a setting as diverse as our method. We compare against C3DPO [30] and PAUL [43] in all datasets since they can produce accurate estimates in a wide range of datasets and settings. We also report results of EMSfM [45] and GbNrSfM [44] on the S-Up3D and PASCAL3D datasets. We compare against [9, 25, 35] only in Human3.6M dataset since they cannot handle occlusions or multiple object categories. We refer to the end-to-end method of [35] as Proc-CNN. Note that, obtaining an Orthographic-N-point (OnP) [41] solution requires an optimization at test time which renders methods that depend on OnP [35, 43] slower than feed-forward methods such as ours. Also, [9], which we refer to as Geo, uses extra datasets and temporal information. In Pascal3D, we also compare against CMR [49]

We report results with three settings: 1) Lifter and transformer are trained separately without context vector (Ours/TR); 2) Lifter and transformer are trained end-to-end without context vector (Ours w/o Context); 3) The proposed end-to-end training with context vector (Ours). We also experiment with the stacked hourglass network [46] to extract 2D keypoints.

5.2 Evaluation protocol

Following [30], we report absolute mean per joint position error $\mathbf{MPJPE}(X, Y) = \sum_{k=1}^K \|X_k - Y_k\|/K$ as well as $\mathbf{Stress}(X, Y) = \sum_{i < j} \left| \|X_i - X_j\| - \|Y_i - Y_j\| \right| / (K(K-1))$, where we center both the estimates and ground truth at zero mean. For all datasets, we follow the same canonical train/test split and evaluation protocol as [30].

6 Results

In order to show the performance of individual components, we separate the results into two parts: (i) using GT 2D keypoints and (ii) estimating the keypoints directly from the image

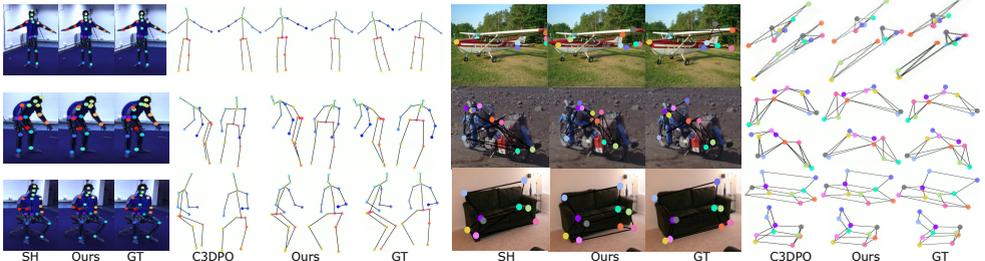


Figure 6: Visual results on the H36m and Pascal3D datasets. 2D keypoint estimates of ours, HRNet (SH) from [11] an GT, plus 3D structures from 2 angles for C3DPO, ours and GT. The images in the figure are cropped around the humans for visualization purposes. Our network produces better 2D keypoint estimations even when the points are occluded.

and producing the 3D pose with these estimates. Note that the latter setting corresponds to image-to-3D task.

Method	Pascal3D		Human3.6M		S-Up3D	
	MPJPE	Stress	MPJPE	Stress	MPJPE	Stress
Geo-SH ^{‡*}	-	-	51	-	-	-
EM-SfM	131.0	116.8	-	-	0.107	0.061
GbNrSfM	184.6	111.3	-	-	0.093	0.062
PoseGAN	-	-	130.9	51.8	-	-
Proc ^{†*}	-	-	86.4	-	-	-
PAUL [†]	30.9	-	88.3	-	0.058	-
C3DPO-base	53.5	46.8	135.2	56.9	0.160	0.105
C3DPO	38.0	32.6	101.8	43.5	0.068	0.040
Ours	29.5	26.6	92.8	42.6	0.057	0.035

Method	Pascal3D		Human3.6M	
	MPJPE	Stress	MPJPE	Stress
Geo-SH ^{‡*}	-	-	68	-
CMR/SH	74.4	53.7	-	-
C3DPO/SH	57.4	41.4	145.0	84.7
Proc-SH ^{†*}	-	-	124.5	-
Proc-CNN ^{†*}	-	-	108.9	-
PAUL-SH [†]	-	-	132.5	-
Ours/SH	56.1	39.0	140.7	80.9
Ours/TR	61.3	47.9	114.0	58.8
Ours w/o Cont	57.6	42.9	113.8	56.7
Ours	51.6	35.3	107.7	55.4

Table 1: Results on Pascal3D, Human3.6M and S-Up3D datasets with **Left** GT keypoints and **Right** estimated keypoints. [†]: Uses test time optimization. ^{*}: Requires temporal sequences for training. [‡]Uses additional datasets for training.

6.1 Results with GT Keypoints

We present our results for the lifter network when the GT keypoints are used in Table 1. Our method outperforms all methods that do not perform test-time optimization apart from [9], which uses additional datasets and temporal information. Our method outperforms all methods in the Pascal3D dataset where our method’s multi-class focus is shown best. We also outperform all other methods in the S-Up3D dataset. Comparing C3DPO-base and Ours, the boost the cut-off coefficients provide can be seen. Our method is only slightly worse than the Procrustean network [53] in the Human3.6M dataset although they use sequences for training and test-time Procrustean optimization. It can be seen that our method produces the best overall results while being applicable in all datasets.

6.2 Results with Estimated Keypoints

The results with estimated keypoints, i.e. direct pose estimation from the image, are given in Table 1. In both datasets our method outperforms the competitors. We see that the boost

mainly comes from the proposed joint training and the context vector. Especially test-time Procrustean optimization methods, even when competitive with GT keypoints, suffer considerably using estimated keypoints, visible in the Human3.6M results. For the Pascal dataset, neither PAUL [48] nor Procrustean network [43] even report numbers. The context vector improves performance and allows for the reprojection loss to provide gradients more easily to the earlier layers of the network. Our method provides the best overall results in different datasets. This is the result of the proposed flexible end-to-end framework that can be readily applied to any dataset.

6.3 Disentanglement

To evaluate the effect of the proposed cut-off weights on the latent space, we measure the mutual coherence of the latent space basis vectors W_g . The linear combinations of these vectors create the latent code that is then decoded into the 3D keypoints via S . Thus, the mutual coherence of the latent basis vectors provides a measure of disentanglement. Table 2 shows that sparse cut-off coefficients encourage latent basis vectors to be less correlated. Please refer to Supplementary Material for visual samples for disentanglement and cross categorical geometric relationships explored by the proposed method.

Method	S-Up-3D	Pascal3D	Human3.6M
Standard (no cut-off)	0.89	0.84	0.44
Ours (cut-off)	0.36	0.38	0.24

Table 2: Mutual coherence of the latent space basis vectors W_g with respect to the proposed cut-off formulation on all datasets.

7 Conclusion

We study the problem of estimating 3D pose and shape from a single image for objects of multiple categories, in an end-to-end manner. Our learning framework relies only on 2D annotations of keypoints for supervision, and exploits the relationships between keypoints within and across categories. The proposed end-to-end learning process offers a structured and unified approach for the image-to-3D problem. Our experiments show that end-to-end training and the use of contextual information improve the performance substantially. Our method is the first of its kind, providing a framework that can be applied to any dataset. We also outperform all the compared methods in estimating 3D shape and pose directly from images, on three benchmark datasets.

References

- [1] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(7):1442–1456, 2011. doi: 10.1109/TPAMI.2010.201. URL <https://doi.org/10.1109/TPAMI.2010.201>.

- [2] Adrien Bartoli, Vincent Gay-Bellile, Umberto Castellani, Julien Peyras, Søren I. Olsen, and Patrick Sayd. Coarse-to-fine low-rank structure-from-motion. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society, 2008. doi: 10.1109/CVPR.2008.4587694. URL <https://doi.org/10.1109/CVPR.2008.4587694>.
- [3] Yoshua Bengio. Learning deep architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, 2009. doi: 10.1561/2200000006. URL <https://doi.org/10.1561/2200000006>.
- [4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016.
- [5] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *2000 Conference on Computer Vision and Pattern Recognition (CVPR 2000), 13-15 June 2000, Hilton Head, SC, USA*, pages 2690–2696. IEEE Computer Society, 2000. doi: 10.1109/CVPR.2000.854941. URL <https://doi.org/10.1109/CVPR.2000.854941>.
- [6] 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.
- [7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13. URL https://doi.org/10.1007/978-3-030-58452-8_13.
- [8] Geonho Cha, Minsik Lee, and Songhwa Oh. Unsupervised 3d reconstruction networks. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3848–3857. IEEE, 2019. doi: 10.1109/ICCV.2019.00395. URL <https://doi.org/10.1109/ICCV.2019.00395>.
- [9] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, and James M. Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5714–5724. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00586. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Chen_Unsupervised_3D_Pose_Estimation_With_Geometric_Self-Supervision_CVPR_2019_paper.html.
- [10] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. *Int. J. Comput. Vis.*, 107(2):101–122,

2014. doi: 10.1007/s11263-013-0684-2. URL <https://doi.org/10.1007/s11263-013-0684-2>.
- [11] Zheng Dang, Fei Wang, and Mathieu Salzmann. 3d registration for self-occluded objects in context. *arXiv preprint arXiv:2011.11260*, 2020.
- [12] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.
- [13] Clara Fernandez-Labrador. *Indoor Scene Understanding using Non-Conventional Cameras*. PhD thesis, Université de Bourgogne Franche-Comté (COMUE)(UBFC), FRA.; Universidad . . . , 2020.
- [14] Katerina Fragkiadaki, Marta Salas, Pablo Andrés Arbeláez, and Jitendra Malik. Grouping-based low-rank trajectory completion and 3d reconstruction. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 55–63, 2014. URL <https://proceedings.neurips.cc/paper/2014/hash/b53b3a3d6ab90ce0268229151c9bde11-Abstract.html>.
- [15] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pages 315–323. JMLR.org, 2011. URL <http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf>.
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL <https://doi.org/10.1109/CVPR.2016.90>.
- [17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [18] Simon Jenni and Paolo Favaro. Self-supervised multi-view synchronization learning for 3d pose estimation. In Hiroshi Ishikawa, Cheng-Lin Liu, Tomás Pajdla, and Jianbo Shi, editors, *Computer Vision - ACCV 2020 - 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - December 4, 2020, Revised Selected Papers, Part V*, volume 12626 of *Lecture Notes in Computer Science*, pages 170–187. Springer, 2020. doi: 10.1007/978-3-030-69541-5_11. URL https://doi.org/10.1007/978-3-030-69541-5_11.

- [19] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, volume 11219 of *Lecture Notes in Computer Science*, pages 386–402. Springer, 2018. doi: 10.1007/978-3-030-01267-0_23. URL https://doi.org/10.1007/978-3-030-01267-0_23.
- [20] Laurent Kneip, Hongdong Li, and Yongduek Seo. Upnp: An optimal $o(n)$ solution to the absolute pose problem with universal applicability. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, 2014.
- [21] Filippos Kokkinos and Iasonas Kokkinos. To the point: Correspondence-driven monocular 3d category reconstruction. *CoRR*, abs/2106.05662, 2021. URL <https://arxiv.org/abs/2106.05662>.
- [22] Chen Kong and Simon Lucey. Prior-less compressible structure from motion. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4123–4131. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.447. URL <https://doi.org/10.1109/CVPR.2016.447>.
- [23] Chen Kong and Simon Lucey. Deep non-rigid structure from motion. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1558–1567. IEEE, 2019. doi: 10.1109/ICCV.2019.00164. URL <https://doi.org/10.1109/ICCV.2019.00164>.
- [24] Chen Kong, Rui Zhu, Hamed Kiani, and Simon Lucey. Structure from category: A generic and prior-less approach. In *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, pages 296–304. IEEE Computer Society, 2016. doi: 10.1109/3DV.2016.38. URL <https://doi.org/10.1109/3DV.2016.38>.
- [25] Yasunori Kudo, Keisuke Ogaki, Yusuke Matsui, and Yuri Odagiri. Unsupervised adversarial learning of 3d human pose from 2d joint locations. *CoRR*, abs/1803.08244, 2018. URL <http://arxiv.org/abs/1803.08244>.
- [26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015.
- [27] Q.-T. Luong and O.D. Faugeras. The fundamental matrix: theory, algorithms, and stability analysis. *International Journal of Computer Vision*, 17:43–75, 1995.
- [28] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651, 2015.
- [29] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2823–2832, 2017.

- [30] David Novotný, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3DPO: canonical 3d pose networks for non-rigid structure from motion. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7687–7696. IEEE, 2019. doi: 10.1109/ICCV.2019.00778. URL <https://doi.org/10.1109/ICCV.2019.00778>.
- [31] 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 International Conference on Computer Vision*, pages 7688–7697, 2019.
- [32] Shaifali Parashar, Mathieu Salzmann, and Pascal Fua. Local non-rigid structure-from-motion from diffeomorphic mappings. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 2056–2064. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00213. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Parashar_Local_Non-Rigid_Structure-From-Motion_From-Diffeomorphic_Mappings_CVPR_2020_paper.html.
- [33] Sungheon Park, Minsik Lee, and Nojun Kwak. Procrustean regression: A flexible alignment-based framework for nonrigid structure estimation. *IEEE Transactions on Image Processing*, 27(1):249–264, 2017.
- [34] Sungheon Park, Minsik Lee, and Nojun Kwak. Procrustean regression: A flexible alignment-based framework for nonrigid structure estimation. *IEEE Trans. Image Process.*, 27(1):249–264, 2018. doi: 10.1109/TIP.2017.2757280. URL <https://doi.org/10.1109/TIP.2017.2757280>.
- [35] Sungheon Park, Minsik Lee, and Nojun Kwak. Procrustean regression networks: Learning 3d structure of non-rigid objects from 2d annotations. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, pages 1–18. Springer, 2020. doi: 10.1007/978-3-030-58526-6_1. URL https://doi.org/10.1007/978-3-030-58526-6_1.
- [36] Mihir Sahasrabudhe, Zhixin Shu, Edward Bartrum, Riza Alp Güler, Dimitris Samaras, and Iasonas Kokkinos. Lifting autoencoders: Unsupervised learning of a fully-disentangled 3d morphable model using deep non-rigid structure from motion. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 4054–4064. IEEE, 2019. doi: 10.1109/ICCVW.2019.00500. URL <https://doi.org/10.1109/ICCVW.2019.00500>.
- [37] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *2011 International Conference on Computer Vision*, pages 667–674. IEEE, 2011.

- [38] Jingnan Shi, Heng Yang, and Luca Carlone. Optimal pose and shape estimation for category-level 3d object perception. *arXiv preprint arXiv:2104.08383*, 2021.
- [39] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pages 1297–1304. Ieee, 2011.
- [40] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *Int. J. Comput. Vision*, 80(2):189–210, 2007.
- [41] Carsten Steger. Algorithms for the orthographic-n-point problem. *Journal of Mathematical Imaging and Vision*, 60(2):246–266, 2018.
- [42] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, and Rudolph Triebel. Augmented autoencoders: Implicit 3d orientation learning for 6d object detection. *International Journal of Computer Vision*, 128(3):714–729, 2020.
- [43] Supasorn Suwajanakorn, Noah Snavely, Jonathan Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. *arXiv preprint arXiv:1807.03146*, 2018.
- [44] Hao Tang, Dan Xu, Gaowen Liu, Wei Wang, Nicu Sebe, and Yan Yan. Cycle in cycle generative adversarial networks for keypoint-guided image generation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2052–2060, 2019.
- [45] Lorenzo Torresani, Aaron Hertzmann, and Christoph Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(5):878–892, 2008. doi: 10.1109/TPAMI.2007.70752. URL <https://doi.org/10.1109/TPAMI.2007.70752>.
- [46] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1653–1660. IEEE Computer Society, 2014. doi: 10.1109/CVPR.2014.214. URL <https://doi.org/10.1109/CVPR.2014.214>.
- [47] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1519, 2015.
- [48] Chaoyang Wang and Simon Lucey. Paul: Procrustean autoencoder for unsupervised lifting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 434–443, 2021.
- [49] Chaoyang Wang, Chen Kong, and Simon Lucey. Distill knowledge from nrsfm for weakly supervised 3d pose learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 743–752. IEEE, 2019. doi: 10.1109/ICCV.2019.00083. URL <https://doi.org/10.1109/ICCV.2019.00083>.

- [50] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild (extended abstract). In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4854–4858. ijcai.org, 2021. doi: 10.24963/ijcai.2021/665. URL <https://doi.org/10.24963/ijcai.2021/665>.
- [51] Heng Yang and Luca Carlone. In perfect shape: Certifiably optimal 3d shape reconstruction from 2d landmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 621–630, 2020.
- [52] Zi Jian Yew and Gim Hee Lee. 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In *European Conference on Computer Vision*, pages 630–646. Springer, 2018.
- [53] Stefanos Zafeiriou, Grigorios G Chrysos, Anastasios Roussos, Evangelos Ververas, Jiankang Deng, and George Trigeorgis. The 3d menpo facial landmark tracking challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2503–2511, 2017.
- [54] 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.
- [55] Wanqing Zhao, Shaobo Zhang, Ziyu Guan, Wei Zhao, Jinye Peng, and Jianping Fan. Learning deep network for detecting 3d object keypoints and 6d poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14134–14142, 2020.
- [56] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4966–4975. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.537. URL <https://doi.org/10.1109/CVPR.2016.537>.
- [57] Yingying Zhu, Dong Huang, Fernando De la Torre, and Simon Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1542–1549. IEEE Computer Society, 2014. doi: 10.1109/CVPR.2014.200. URL <https://doi.org/10.1109/CVPR.2014.200>.