Discovering Relationships between Object Categories via Universal Canonical Maps

Natalia Neverova*, Artsiom Sanakoyeu*, Patrick Labatut, David Novotny, Andrea Vedaldi
Facebook AI Research
{nneverova,asanakoy,dnovotny,plabatut,vedaldi}@fb.com

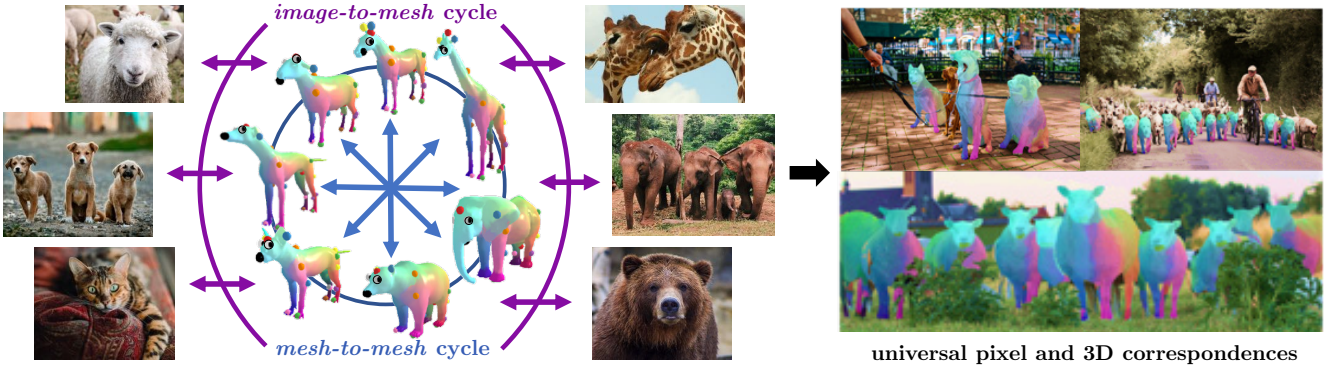


Figure 1: **Inter-category correspondences emerge from dense pose prediction.** Our method discovers high-quality correspondences between different object classes automatically, as a byproduct of learning category-specific dense pose predictors. It does so by enforcing cycle consistency between reference 3D templates as well as by a new type of consistency between images and templates. This allows the model to transfer information between animal classes (*e.g.* the location of the eyes).

Abstract

We tackle the problem of learning the geometry of multiple categories of deformable objects jointly. Recent work has shown that it is possible to learn a unified dense pose predictor for several categories of related objects. However, training such models requires to initialize inter-category correspondences by hand. This is suboptimal and the resulting models fail to maintain correct correspondences as individual categories are learned. In this paper, we show that improved correspondences can be learned automatically as a natural byproduct of learning category-specific dense pose predictors. To do this, we express correspondences between different categories and between images and categories using a unified embedding. Then, we use the latter to enforce two constraints: symmetric inter-category cycle consistency and a new asymmetric image-to-category cycle consistency. Without any manual annotations for the inter-category correspondences, we obtain state-of-the-art alignment results, outperforming dedicated methods for matching 3D shapes. Moreover, the new model is also better at the task of dense pose prediction than prior work.

*Both authors contributed equally to this work.

1. Introduction

Algorithms can nowadays understand well the geometry of *specific object categories* such as humans: we have reliable methods for detecting and segmenting them, extracting their 2D landmarks and dense surface coordinates, as well as reconstructing them in 3D. In principle, these methods can be applied to many other types of objects, such as any kind of animal, from pets to wildlife. In practice, however, doing so is often prohibitively expensive. The main bottleneck is data acquisition, especially for supervised training in 3D, and extensive manual annotation. High-quality 3D human models are bootstrapped using specialized motion capture systems such as domes that are difficult to apply to objects such as wild animals. Annotating 2D geometric primitives such as segments and 2D keypoints can be done manually from raw images, but it is costly and somewhat difficult to do for unfamiliar animal anatomies. Thus, a naïve application of existing high-quality model acquisition techniques cannot trivially scale to learning the massive variety of object types that exist in the world, which include 6.5K mammal species, 7.7M animal species, and around 8.7M natural species overall [9, 34].

The key to scaling is to realize that, while there are indeed millions of different types of objects, these are not independent. For instance, different cat breeds are relatively similar, so a single 'cat' model is likely to work well for all cats, just like a single 'human' model has been shown to work well for many different human body shapes [32]. In fact, useful information can likely be shared among fairly different types of objects, such as all mammals or all animals. The limit is given by the ability of the model to represent diverse information while capturing and eliminating redundancies wherever possible. The hope is that such a model could learn the geometry of different object types with a cost which is sub-linear in their number.

A similar idea was recently pursued in [35] for the task of *dense pose prediction* [17]. Just like 2D pose prediction estimates the location of a small number of distinctive object landmarks, dense pose estimation does so for a continuous set of landmarks, identified as the point of a 3D template of the object (fig. 1). The goal is to learn a *canonical map*, i.e. a function that maps all relevant pixels in an image to the corresponding points in the template, thus identifying them. For supervised learning, correspondences between images and templates are collected manually, using a category-specific template for each example object. As a result, annotations for different object categories are unrelated, which makes it hard to learn a universal, category-agnostic object representation.

In order to address this problem, the authors of [35] establish initial point-to-point correspondences between different category-specific templates using a mix of manual annotations and automated interpolation. However, as we show in the experiments, their approach has two shortcomings. Firstly, their manual correspondence initialization is somewhat arbitrary and thus likely suboptimal. The second problem, which partially arises from the first, is that their initial inter-category correspondences are not *maintained* while the model is trained, and are eventually 'forgotten'.

In this paper we argue that, if the goal of the alignment is to facilitate learning a multi-category object representation, an optimal alignment should emerge spontaneously as part of the learning process, thus solving the two issues above. Our key contribution is thus a new learning formulation for universal canonical maps that induces *automatically* high-quality intra-category correspondences. The most important outcome is that the learned maps solve the dense pose prediction problem accurately for several object categories while at the same time putting those in correspondence, allowing to transfer information between them.

We base our model on learning a single, universal embedding space to express all required correspondences. Points in the different 3D templates as well as image pixels are mapped to this common space, which allows to compute dense template-to-template and image-to-template correspondences. Differently from [35], the template embeddings in this work are *not* initialized from manually annotated inter-category correspondences. Instead, all embeddings are obtained automatically while learning the canonical maps for individual categories while satisfying certain consistency constraints.

For the constraints, we use simple but effective rules. Apart from the most basic one, which encourages similarity of the embeddings of nearby template points (smoothness), we contribute by introducing two types of cycle-consistency for learning canonical surface mappings: The first one enforces cycle consistency between different 3D templates, which encourages bijective correspondences between them. Additionally, we note that canonical maps, by establishing correspondences from images to templates, are not bijective but injective, and we show that this can be exploited by an asymmetric form of cycle consistency between images and templates. By using the common embedding space, all such constraints are expressed as differentiable loss terms.

Empirically, we demonstrate several advantages of our new approach compared to [35]. We show that our approach finds automatically high-quality correspondences between different object categories *without any manual supervision for this task*. This is compelling because it shows that, as we hypothesized, there is a natural advantage in learning jointly the geometry of different but related object types. In fact, the 3D correspondences we discover in this manner outperform the ones discovered by state-of-the-art 3D shape alignment methods. Finally, our method not only aligns canonical maps, but also improves their quality, resulting in more accurate dense pose prediction than the state of the art.

2. Related work

Human pose estimation. Human pose prediction often starts by detecting 2D landmarks, usually coinciding with the main joints of the body [31, 1, 23, 22]. For this task, early shallow methods [15, 5, 23, 40] have been surpassed by deep convolutional architectures [37, 48, 10]. Sparse landmarks can be replaced by dense ones, identifiable with a reference 3D template of the object. The resulting dense pose prediction problem was pioneered by DensePose [17] using the SMPL [32] mesh as a canonical template. Parsing R-CNN [51] improved the Dense Pose network by extending the popular R-CNN architecture [16]. More recently, Slim DensePose [36] showed that a smaller number of keypoint annotations is sufficient to learn competitive DensePose models, potentially significantly reducing the effort required for learning new non-human categories.

Animal pose estimation. Several works also attempted to estimate the pose of various animal species. Methods such as [52, 44, 25, 26, 38, 49] learned to detect [52, 44], match [25] or reconstruct [26, 38] various birds from the

CUB dataset [49]. 3D reconstruction of the shape of a broader set of animal species has been attempted by Zuffi et al. in [59, 60, 58]. Similar to monocular 3D human mesh recovery models [24, 28, 27] that predict parameters of SMPL, [59, 60, 58] utilize a parametric model of a mesh of an animal body (SMAL) in order to constrain the set of possible animal reconstructions, with further improvements in the work of Biggs et al. [4, 3]. Sanakoyeu et al. [43] transfer DensePose from humans to proximal animal classes without extra labels by a self-training approach.

The work most relevant to ours is Neverova et al. [35], which introduced the idea of continuous surface embeddings (CSE) to tackle the dense pose prediction problem for several animal categories together. They further contributed a dataset of dense pose annotations for various animal species. We improve on [35] by learning more accurate canonical maps that are more consistent across different categories and by not requiring any manual initialization for the correspondences between different object categories. We also contribute with an extended dataset of dense animal poses for experimentation.

Intrinsic 3D shape analysis. Our work is also related to the analysis of the intrinsic properties of 3D shapes, where the fundamental problem is to establish correspondences between different shapes. Non-deep learning methods include embeddings of geodesic distance matrices [13, 6] and various kinds of diffusion geometry [11] descriptors — Heat Kernel Signature [45] and its scale-invariant follow-up [8], Gromov-Hausdorff descriptors [7] and the Wave Kernel Signature [2]. One of the main building blocks of the aforementioned descriptors are the eigenfunctions of the Laplace-Beltrami operator (LBO) [42] that define a smooth basis of a coordinate frame of a mesh surface. Ovsjanikov et al. [39] proposed the functional map (FM) framework that establishes soft correspondences between pairs of shapes by relating the mesh LBO eigenfunctions with a simple linear mapping. The CSE method from [35] uses FMs and ZoomOut [33] to interpolate an initial set of manually-established inter-class correspondences. Differently from them, we only use LBO to express smoothness, but we otherwise consider topological constraints such as bijectivity and injectivity that are more appropriate for establishing non-isometric correspondences, such as between different animal categories.

Cycle consistency. Cycle consistency is a powerful paradigm that has been explored in many different fields of computer vision: pixel-wise image matching [56, 55], image translation [57], or category-specific 3D reconstruction [53]. Given a single input image of an instance of an object category, Kulkarni et al. [30, 29] enforce consistency between a rendered UV map of a 3D template shape of the object category and the learned canonical map, while our method does not require to fit/render the 3D model. In the context of 3D shape analysis, Huang et al. [20] introduced a semi-definite programming formulation that factorized a matrix of all point-to-point matches between pairs of meshes in order to make the matches cycle-consistent. Similarly, Yang et al. [50] use the Sinkhorn regularization (SH) to find the nearest cycle-consistent solution to an initial matrix of noisy point-wise matches. Ren et al. [41] exploit the spectral properties of correspondences and cycle consistence between shape pairs. Our method is inspired by [50] in the sense that we utilize cycle consistency in order to improve our dense pose labels by relating surfaces of different category template shapes.

3. Method

We start by summarizing the continuous surface embedding (CSE) representation of [35] (section 3.1) and then we explain how to extend it to learn high-quality inter-category correspondences automatically (section 3.2).

3.1. Continuos surface embeddings

The *continuous surface embedding* (CSE) of [35] allows us to express correspondences between different 3D templates and between templates and images in a homogeneous and differentiable manner. A CSE is a function $e : S \to \mathbb{R}^D$ sending each point $X \in S$ of a 3D surface S to a D-dimensional embedding vector. We assume that the surface S is a mesh with K vertices $S = (X_k)_{1 \le k \le K}$ and we collect the corresponding embedding vectors as the rows of a matrix $E \in \mathbb{R}^{K \times D}$. The matrix E, which we learn from data, can be fairly large, but smoothness[†] can help to reduce its dimensionality. This can be done by considering a smooth functional basis $U \in \mathbb{R}^{K \times Q}$ on the mesh, where $Q \ll D$, and define $E = U\hat{E}$. With this, we can work with the compressed embedding parameterization $\hat{E} \in \mathbb{R}^{Q \times D}$. As in [35], we take U to be the lowest eigenvectors of the Laplace-Beltrami operator (LBO) of the mesh S. While the LBO is often used in the literature as a cue to match near-isometric shapes, our shapes are not at all isometric. For this reason, we use the LBO only to encode a generic notion of smoothness, but *not* as a cue for matching.

Encoding correspondences via CSEs. Embedding vectors can be used to define correspondences between any two sets of objects $A = (a_1, \ldots, a_K)$ and $B = (b_1, \ldots, b_L)$. Namely, given embedding functions $e : A \to \mathbb{R}^D$ and $e : B \to \mathbb{R}^D$, we can compare embedding vectors to send elements of set B to elements of set A probabilistically:

$$p(a_k | b_l, e) = \frac{\exp\left(-\langle e_{a_k}, e_{b_l} \rangle\right)}{\sum_{t=1}^{K} \exp\left(-\langle e_{a_t}, e_{b_l} \rangle\right)}. \quad (1)$$

[†] *I.e.* the fact that nearby vertices should have similar embeddings.

In our case, given two CSEs $E = U\hat{E}$ and $E' = U'\hat{E}'$ for two different meshes S and S', eq. (1) gives us distributions $p(X_k|X'_l, \hat{E}, \hat{E}')$ and $p(X'_l|X_k, \hat{E}, \hat{E}')$, encoding mappings $S' \to S$ and $S \to S'$, respectively. We can also express image-to-mesh and mesh-to-image maps. For this, let I be an image and consider a finite set $\Omega \subset \mathbb{R}^2$ of pixel locations. We use a deep convolutional neural network $e_x = [\Phi(I)]_x$ to compute the embedding vectors for all the pixels $x \in \Omega$. Then, given a mesh S together with its embedding matrix $E = U\hat{E}$, we can use eq. (1) to obtain a distribution $p(X_k|x, \hat{E}, \Phi(I))$ encoding a map $\Omega \to S$ from the pixels to the mesh. Note that the latter is, by definition, a canonical map, and as such it provides a solution to the dense pose prediction task. We can also swap the roles of image and mesh in this expression, obtaining a probability $p(x|X_k, \hat{E}, \Phi(I))$ encoding a reverse map $S \to \Omega$. This map will be useful later.

Finally, we can, in an entirely analogous manner, define image-to-image correspondences $\Omega \to \Omega'$ by comparing embeddings $\Phi_x(I)$ and $\Phi(I')_{x'}$. This is useful to transfer information directly across images, as we demonstrate in the experiments for keypoint transfer.

Working with multiple object categories. The approach above can easily accommodate any number of categories and corresponding templates. Each category $m = 1, \dots, M$ is captured by a mesh and its embedding $(S^m, \hat{E}^m)$. Each mesh can have a different number of vertices $|S^m| = K^m$. Likewise, the LBO basis $U^m \in \mathbb{R}^{Q^m \times D}$ is mesh-specific, including having potentially a different number of basis elements Q^m. Crucially, however, the dimensionality of the embedding space D is the same for all templates, as the embeddings must be comparable.

3.2. Dense pose and emerging correspondences

The use of a common embedding space for templates and images means that all such objects can be put in correspondence by using the method of section 3.1. However, this does not necessarily mean that the correspondences learned by the model are meaningful. In more detail, manual annotations for the dense pose task are of the type (I, m, x, X) where I is an image, m a category, x a pixel, and $X \in S^m$ its corresponding vertex in the category-specific template S^m [17, 35]. By fitting such annotations, the model is encouraged to learn good dense pose predictors for each category, but not necessarily good inter-category correspondences. The latter may emerge because the neural network Φ is shared in full or in part among different categories, which means that similarly-looking images will naturally tend to be embedded in similar ways. However, this is a weak effect. Below, we add several constraints to improve the quality of the emerging correspondences.

Dense pose supervision. Solving the dense pose prediction tasks means learning maps $\Omega \to S^m$ sending the image region Ω that contains an occurrence of the object to the template S^m of the object itself. As noted above, supervision for this task comes in the form of a dataset $\mathcal{D}$ of tuples (I, m, x, X) and is captured by the loss as follows:

$$\mathcal{L}^{\text{sup}} = \frac{1}{|\mathcal{D}|} \sum_{(I,m,x,X)\in\mathcal{D}} \sum_{k=1}^{K^m} d_{S^m}(X_k^m, X) \cdot p(X_k^m|x, \hat{E}^m, \Phi^m(I)). \quad (2)$$

In this expression, d_{S^m} is the geodesic distance on the mesh S^m. This loss is optimized w.r.t. the mesh embeddings and neural networks $(\hat{E}^m, \Phi^m)_{1\le m\le M}$ (where the different networks share most or all of their parameters).

Inter-category correspondences. We assume that there exists sensible one-to-one correspondences $S^m \leftrightarrow S^n$ between any pair of templates. In this case, the cycle $S^m \to S^n \to S^m$ should approximate the identity function. We can rewrite the cycle in terms of the probabilistic correspondences described in section 3.1 by marginalizing the intermediate step as follows:

$$p(X_k^m|X_t^m) = \sum_{l=1}^{K^n} p(X_k^m|X_l^n)\, p(X_l^n|X_t^m). \quad (3)$$

While we do not show it for compactness, note that all such probabilities depend on the learned embeddings $\hat{E}^m$ and $\hat{E}^n$. If the cycle is closed correctly, this probability should peak at $X_t^m = X_k^m$, which is captured by the *mess-to-mesh* loss (**m2m**):

$$\mathcal{L}^{mn} = \frac{1}{K^m} \sum_{k=1}^{K^m} \sum_{t=1}^{K^m} d_{S^m}(X_k^m, X_t^m)\, p(X_k^m|X_t^m). \quad (4)$$

To the loss $\mathcal{L}^{mn}$ we also add the symmetric loss $\mathcal{L}^{nm}$. Cycle consistency has been exploited before in many different contexts [30, 47, 50, 57, 54, 19, 21]. Here we use it to guide the discovery of correspondences between different meshes.

Canonical map injectivity. The signal (2) is only given at a sparse set of manually-labelled image pixels. A denser constraint can be obtained by noting that the canonical maps $\Omega \to S^m$ must be *injective*, in the sense that all pixels in the object region Ω should map to different vertices in the mesh S^m. Injectivity means that the canonical map has a left inverse: if a mesh vertex corresponds to at least one image pixel, then this correspondence must be unique. We can thus close the cycle $\Omega \to S^m \to \Omega$, resulting in the *image-to-mesh* loss (**i2m**):

$$\mathcal{L}^{Im} = \frac{1}{|\Omega|} \sum_{x\in\Omega} \sum_{y\in\Omega} d_I(y, x) p(y|x). \quad (5)$$

dataset	init.	train	AP ↓	GErr ↓	GPS ↑
DP-LVIS [35]	ZoomOut	–	25.4	23.7	66.7
		✓	35.1	28.0	68.7
	Random	✓	34.4	34.1	63.7
DP-LVIS v1.0	Random	✓	37.4	20.7	77.1

Table 1: **Baselines (humans & animals).** We train a universal canonical map using DensePose-COCO and animal data and report DensePose performance on animal categories (**AP**), as well as mesh alignment quality for animals and people (**GErr** and **GPS**). The architecture is of [35], combined with multi-class detection. ZoomOut initialization does not result in performance gains on a larger dataset.

where d_I is a distance in image space (e.g. Euclidean) and, similar to eq. (3),

$$p(y|x) = \frac{1}{K^m} \sum_{k=1}^{K^m} p(y|X_k^m)\, p(X_k^m|x). \tag{6}$$

While not shown for compactness, all these probabilities depend on the mesh embedding E^m and the neural network Φ^m that we wish to learn.

Rather than summing eq. (5) on the entire set Ω, we consider a *downsampled* version $\bar{\Omega} \subset \Omega$ with $|\bar{\Omega}| \ll |\Omega|$. This is done for computational efficiency (as there can be a very large number of pixels in certain image regions). Compared to using the full domain, the effect is to slightly relax eq. (5).

Note that, differently from the mesh-to-mesh cycle, this cycle is *not* symmetric: while we can close the chain $\Omega \rightarrow S^m \rightarrow \Omega$, we *cannot* close the chain $S^m \rightarrow \Omega \rightarrow S^m$. The first chain is valid because all pixels in Ω correspond to *a unique point* of the mesh S^m. On the other hand, many of the points in the mesh S^m will *not* have a corresponding image point in Ω for the simple fact that at least part of the object cannot be visible in a given image.

3.3. Overall loss

To summarize, our model is trained by minimizing a combination of the losses of eqs. (2), (4) and (5):

$$\mathcal{L}^{\text{sup}} + \frac{1}{M(M-1)} \sum_{\substack{m,n=1 \\ m \neq n}}^{M} \mathcal{L}^{mn} + \frac{1}{|\mathcal{D}|} \sum_{(I,m)\in\mathcal{D}} \mathcal{L}^{Im}.$$

4. Experiments

After discussing the experimental data and implementation details, we focus on our key contributions: simultaneously discovering high-quality inter-class correspondences while learning category-specific dense pose predictors (sections 4.1 and 4.2). We also show (section 4.3) that learned embeddings in the pixel space allow for effective retrieval of analogous points (body landmarks) on the surfaces of objects belonging to the same or different categories (a task that we call *keypoint transfer*).

Training datasets. Following [35], we use the original people-centric DensePose-COCO [17] for pre-training in all experiments. We also exploit this data for conducting studies on a joint set of animal and human categories.

For the animal classes, we propose an extended version of the DensePose-LVIS data of [35], which originally contained 9 animal categories from the LVIS v0.5 dataset for instance segmentation [18]. Following a recent release of LVIS v1.0, which extended the benchmark to 160k images and 2M instance annotations, we also expand the DensePose annotation pool for the same animal classes and call this benchmark DensePose-LVIS v1.0 (see sup. mat.). The original DensePose-LVIS contains fairly sparse annotations (according to [35] only 18% of the vertices of the animal meshes have at least one ground truth annotation, and each image contains no more than 3 annotated points). While compared to DensePose-LVIS, we have 3.6× annotations, this is still far less than the original DensePose-COCO (which annotates 5 million points and obtained 96% coverage of the SMPL mesh). At the same time, quality of dense labels in DensePose-LVIS v1.0 has been further improved by introducing an additional step of croud-sourced manual verification for all annotations.

Implementation details. Our architecture is similar to the R50 variant of [35], with the only difference being the multi-class setting for object detection (in [35] detection was implemented in a class-agnostic manner, and ground truth class labels were required for inference).

We pre-train on the DensePose-COCO dataset for 130k iterations (following the standard s1x schedule). All animal models are then trained on DensePose-LVIS v1.0 for 16k iterations, with a 10x drop of learning rate after 12k and 14k iterations, with the rest of hyperparameters being identical to [35]. For experiments on the joint set of human and animal categories, we train our models for 80k iterations (with a learning rate decrease after 60k and 70k iterations).

Evaluation metrics. The quality of learned *dense pose predictions* is evaluated by a standard set of **AP/AR** metrics [17] (higher is better). We also estimate the quality of *inter-class mesh alignment* by computing the Geodesic Error (**GErr**, lower is better) between the predicted and the ground truth vertices along the surface of the target mesh, given a set of manually annotated semantic keypoints. For this purpose, all vertex coordinates in each mesh are normalised to have the maximum of geodesic distance $\mathbf{d}_{max} = 2.27$ (analogously to [17, 35]). Finally, we report Geodesic Point Similarity (**GPS**, as in [17], higher is better) as an alternative indicator of the quality of cross-category mesh alignment.

Figure 2: **Qualitative results by the full model m2m+i2m-all.** The `dog` 3D model serves as a common reference for all classes. This setup is significantly more challenging than in [35], where each class was visualised with its own 3D reference.

method	GErr ↓	GPS ↑	AP	AP_{50}	AP_{75}	AP_M	AP_L	AR	AR_{50}	AR_{75}	AR_M	AR_L
our baseline	14.05	84.33	37.5	67.8	36.4	35.1	41.9	51.5	78.8	53.2	41.7	55.1
w/ **m2m**	11.96	87.35	38.2	68.5	36.4	36.6	42.6	52.0	79.7	52.6	42.9	55.6
w/ **i2m**	12.67	85.13	38.1	68.7	36.2	35.5	42.4	52.0	79.6	53.2	42.7	55.6
w/ **i2m**-all	11.74	87.48	38.3	**68.9**	36.3	35.7	42.5	52.3	**79.9**	53.5	42.8	55.7
w/ **m2m+i2m**	11.37	88.14	38.3	68.7	36.6	36.7	42.5	52.3	79.7	53.7	43.0	55.7
w/ **m2m+i2m**-all	**10.90**	**88.85**	**38.5**	68.7	**37.1**	**37.5**	**42.6**	**52.5**	79.7	**54.3**	**43.8**	**55.9**

Table 2: **Ablation of cycle-consistency loss terms (animals only): i2m** corresponds to comparing the image to the target mesh given the ground truth class label, while **i2m**-all matches all object instances to all meshes in a cross-category regime.

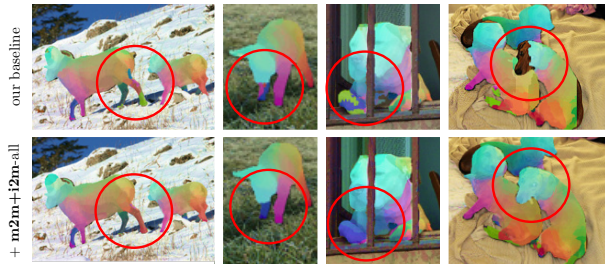


Figure 3: **Effect of the m2m+i2m-all term:** improved local consistency and smoothness in dense pose predictions (see outlined regions), as well as more accurate instance masks.

4.1. Inter-class alignment and dense pose prediction

Compared to prior work such as [35], our most important contribution is to discover automatically effective inter-category correspondences, without manual input for this task, while simultaneously learning high-quality canonical maps for each of the individual animal object categories.

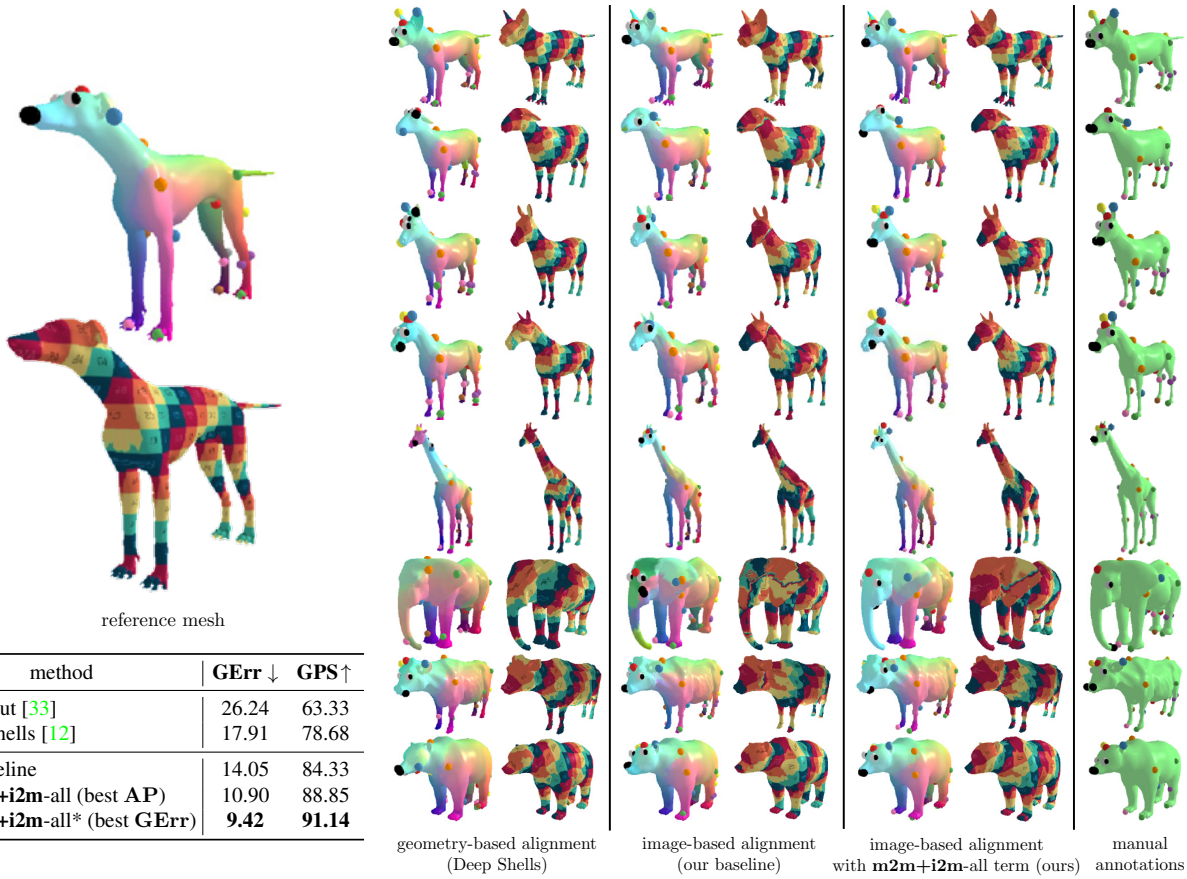
In order to conduct a fair comparison, we start by re-running the baseline of [35] using the embeddings and the DensePose-LVIS v1.0 data (table 1). We also compare using two different initializations for the embedding of the different 3D canonical shapes: random and ZoomOut. The latter follows [35], obtaining an initial set of inter-class 3D mesh correspondences from sparse manual annotations interpolated using the ZoomOut technique [33].

We observe a 2.3pp AP gain in DensePose performance on the new dataset (AP 35.1 → 37.4). While ZoomOut initialization of animal mesh embeddings provided a clear advantage for DensePose in a lower data regime (AP 25.4 → 35.1), the quality of mesh alignment worsens as the network diverges from its initialization point (GErr 23.7 → 28.0). The dynamic on animal-only categories is similar. On DensePose-LVIS v1.0 the automatic alignment learned from random initialization is already of better quality than ZoomOut (20.7 GErr), and the difference in DensePose performance is no longer observed. Note that the latter is already a confirmation of our key hypothesis that good inter-category correspondences should spontaneously emerge by jointly modelling them.

In table 2 we report results on the animals-only benchmark, including assessing the contributions of the **m2m** and **i2m** regularisers. Both **m2m** and **i2m**-all terms significantly improve mesh alignment (GErr 14.05 → 12.67, 11.74, respectively) and also contribute to the dense pose performance (AP 37.5 → 38.1, 38.3, respectively). Their combination yeilds best results (GErr 14.05 → 10.90, AP 37.5 → 38.5) and fixes certain typical errors, as shown in fig. 3. Fig. 2 shows qualitative results.

4.2. Further inter-class alignment analysis

We compare the quality of inter-class mesh alignment produced by our networks with state-of-the-art methods exploiting 3D geometry: namely, ZoomOut [33] (initialized with the same manually keypoints, as we use for evaluation) and Deep Sheels [12] (unsupervised). Qualitative and quantitative results on animal classes are shown in



method	GErr ↓	GPS↑
ZoomOut [33]	26.24	63.33
Deep Shells [12]	17.91	78.68
our baseline	14.05	84.33
+ **m2m+i2m**-all (best **AP**)	10.90	88.85
+ **m2m+i2m**-all* (best **GErr**)	**9.42**	**91.14**

Figure 4: **3D mesh alignment: 9 animals.** *number obtained with a 10x increased weight of the **m2m+i2m**-all term.

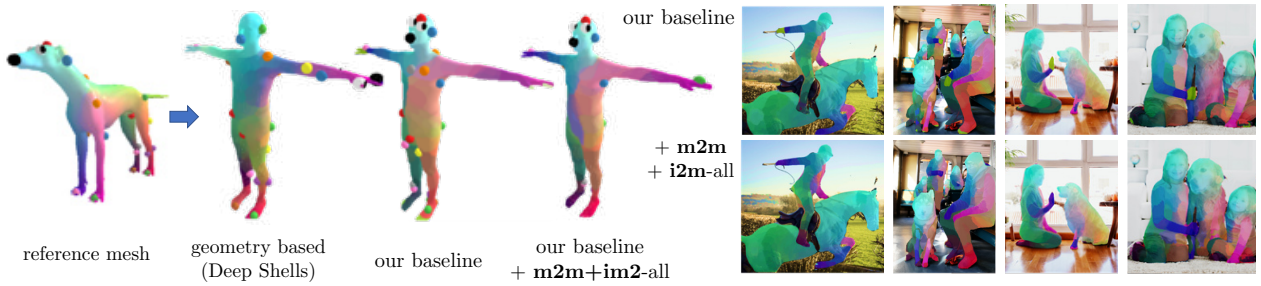


Figure 5: **3D model alignment: `dog-human`.** The **m2m+i2m**-all term is critical for aligning dissimilar categories in 3D.

fig. 4. Our method demonstrates consistently stronger performance across all categories, and rather successfully handles transfer between highly dissimilar categories, such as `dog-giraffe` and `dog-elephant`, where state-of-the-art geometry-based methods tend to fail (GErr ZoomOut: 26.24, Deep Shells: 17.91, and our best result: 9.42).

An extreme case of `human-dog` alignment is shown in fig. 5: our method produces meaningful correspondences in the 3D space (on the left) and consistent cross-category predictions in the pixel space (on the bottom right, shown using the DOG 3D mesh as a reference for visualization).

4.3. Keypoint transfer

To evaluate learned transferability of surface embeddings within and across training categories, as well as their ability to generalize to new animal classes, we look at the problem of *keypoint transfer*. As per section 3.1, we can in fact use our learned embeddings to establish correspondences between a source image I and a target image I' directly and use it to transfer keypoints. To do this, in the target image, for each type of landmark annotated as "visible", we take the nearest neighbor of the embedding of the pixel in the source image, corresponding to the same landmark.

method	target class	supervision			animal category					mean
		mask	points	3D mesh	HORSE	COW	SHEEP	CAT	DOG	
Rigid-CSM [30]	single	✓	✗	✓	31.2	26.3	24.7	–	–	–
Dense-Equi [46]	single	✓	✗	✗	23.3	20.9	19.6	–	–	–
A-CSM [29]	single	✓	✗	✓	32.9	26.3	28.6	–	–	–
Rigid-CSM + keyp. [30]	single	✓	✓*	✓	42.1	28.5	31.5	–	–	–
A-CSM + keyp. [29]	single	✓	✓*	✓	44.6	29.2	39.0	–	–	–
our baseline	multi	✓	✓	✗	58.1	49.9	43.9	41.6	41.9	47.1
w/ **m2m**	multi	✓	✓	✗	57.1	49.5	45.1	40.0	42.5	46.8
w/ **i2m**	multi	✓	✓	✗	59.0	51.1	46.2	45.9	45.7	49.7
w/ **i2m**-all	multi	✓	✓	✗	**59.2**	**51.5**	**46.3**	**46.5**	**45.9**	**49.9**

Table 3: **Keypoint transfer on PASCAL VOC, within each of training animal categories**. PCK-Transfer metric, higher is better. * – supervision on the same set of keypoints that are used for evaluation, as opposed to random sampling in DensePose.

	method	HORSE	COW	SHEEP	CAT	DOG	mean
(I)	our baseline	47.7	45.7	43.5	41.8	40.0	43.8
	w/ **m2m**	47.6	45.0	45.0	41.4	40.5	43.9
	w/ **i2m**	**49.5**	**47.4**	**47.0**	**44.4**	**44.1**	**46.5**
(II)	our baseline	52.0	49.1	43.0	34.6	42.1	44.2
	w/ **i2m**	**54.6**	**49.5**	**44.7**	**37.7**	**43.7**	**46.0**

Table 4: **Keypoint transfer on PASCAL VOC:** (I) across training categories, (II) within new animal categories not observed during training for dense correspondences (only boxes and masks are provided during training to ensure robust detection). PCK-Transfer metric, higher is better.

Figure 6: **Keypoint transfer on PASCAL VOC.** One experiment – one column. Green marks indicate categories included in training, red marks – a new, test only category.

Evaluation metric. Following [29], we evaluate performance on this task by estimating the Percentage of Correct Keypoint transfers (PCK-Transfer). The transfer is said to be correct if the target landmark is localized within distance $0.1 \cdot \max(h, w)$ from the annotated location, where h, w are the height and the width of the predicted bounding box. Prior to that, we match predicted object instances to ground truth objects by estimating the bounding box IoU. Objects that are not retrieved are excluded from evaluation. We report performance on animal categories from PASCAL VOC [14], overlapping with animal categories in DensePose-LVIS v1.0 (`horse`, `cow`, `sheep`, `cat`, `dog`).

Experimental protocol. We evaluate the ability of our model to perform keypoint transfer in three distinct settings: (a) within each category observed at training time (Tab. 3); (b) across training categories (Tab. 4, I); (c) within new animal categories (zero shot) not observed at training time. For (c), dense correspondences for one class are removed from the training set, and only bounding boxes and object instance masks are provided as supervision (Tab. 4, II).

Discussion. As shown in tables 3 and 4 and fig. 6, the learned embeddings work very well to transfer keypoints between known as well as unknown animal classes, demonstrating once more the power of generalization that comes from joint modelling. For this experiment, the **i2m** regularization term significantly improves the results (*e.g.* **m2m** vs **i2m** PCK: 46.8 → 49.9 in table 4). This might be expected since **m2m** works on the alignment between 3D templates, whereas **i2m** works at the image level, which is more relevant for this experiment. Note that methods [30, 29] reproject a 3D mesh template onto the images to establish correspondences, which we do not do.

5. Conclusions

We have introduced a method to learn high-quality dense pose predictors for multiple object categories while discovering automatically semantic correspondences between them. The method represents correspondences via a unified embedding and network predictor while enforcing reasonable topological consistency constraints. Our encouraging results indicate that joint modelling is not only just viable, but has significant positive effects on performance and scalability of such neural dense predictors.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter V. Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proc. CVPR*, 2014.

[2] Mathieu Aubry, Ulrich Schlickewei, and Daniel Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1626–1633, 2011.

[3] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop. *arXiv preprint arXiv:2007.11110*, 2020.

[4] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures Great and SMAL: Recovering the Shape and Motion of Animals from Video. In *Asian Conference on Computer Vision (ACCV)*, pages 3–19, 2018.

[5] Lubomir D. Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *Proc. ICCV*, 2009.

[6] Alexander M. Bronstein, Michael M. Bronstein, and Ron Kimmel. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences (PNAS)*, 103(5):1168–1172, 2006.

[7] Alexander M. Bronstein, Michael M. Bronstein, Ron Kimmel, Mona Mahmoudi, and Guillermo Sapiro. A Gromov-Hausdorff framework with Diffusion Geometry for Topologically-Robust Non-rigid Shape Matching. *International Journal of Computer Vision*, 89(2–3):266–286, 2010.

[8] Michael M. Bronstein and Iasonas Kokkinos. Scale-invariant heat kernel signatures for non-rigid shape recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1704 – 1711, 2010.

[9] Connor J Burgin, Jocelyn P Colella, Philip L Kahn, and Nathan S Upham. How many species of mammals are there? *Journal of Mammalogy*, 99(1):1–14, 02 2018.

[10] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proc. CVPR*, 2017.

[11] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[12] Marvin Eisenberger, Aysim Toker, Laura Leal-Taixe, and Daniel Cremers. Deep shells: Unsupervised shape correspondence with optimal transport. *arXiv preprint*, 2020.

[13] Asi Elad (Elbaz) and Ron Kimmel. On Bending Invariant Signatures for Surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1285–1295, 2003.

[14] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1), 2015.

[15] Pedro F. Felzenszwalb, David A. McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. CVPR*, 2008.

[16] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013.

[17] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. In *Proc. CVPR*, 2018.

[18] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proc. CVPR*, 2019.

[19] Qi-Xing Huang and Leonidas J. Guibas. Consistent shape maps via semidefinite programming. *Eurographics Symposium on Geometry Processing*, 32(5), 2013.

[20] Qi-Xing Huang and Leonidas Guibas. Consistent shape maps via semidefinite programming. In *Computer Graphics Forum*, volume 32, pages 177–186. Wiley Online Library, 2013.

[21] Daniel Huber. *Automatic Three-dimensional Modeling from Reality*. PhD thesis, Carnegie Mellon University, 2002.

[22] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[23] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proc. CVPR*, 2011.

[24] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc. CVPR*, 2018.

[25] Angjoo Kanazawa, David W. Jacobs, and Manmohan Chandraker. WarpNet: Weakly supervised matching for single-view reconstruction. In *Proc. CVPR*, 2016.

[26] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proc. ECCV*, 2018.

[27] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[28] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proc. CVPR*, 2019.

[29] Nilesh Kulkarni, Abhinav Gupta, David F Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 452–461, 2020.

[30] Nilesh Kulkarni, Abhinav Gupta, and Shubham Tulsiani. Canonical surface mapping via geometric cycle consistency. In *Proc. ICCV*, 2019.

[31] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. ECCV*, 2014.

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. on Graphics (TOG)*, 2015.

[33] Simone Melzi, Jing Ren, Emanuele Rodolà, Abhishek Sharma, Peter Wonka, and Maks Ovsjanikov. Zoomout: Spectral upsampling for efficient shape correspondence. *ACM Transactions on Graphics (TOG)*, 38(6):155, 2019.

[34] Camilo Mora, Derek P. Tittensor, Sina Adl, Alastair G. B. Simpson, and Boris Worm. How many species are there on earth and in the ocean? *PLOS Biology*, 9(8), 2011.

[35] Natalia Neverova, David Novotný, and Andrea Vedaldi. Continuous surface embeddings. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[36] Natalia Neverova, James Thewlis, Riza Alp Güler, Iasonas Kokkinos, and Andrea Vedaldi. Slim DensePose: Thrifty learning from sparse annotations and motion cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[37] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proc. ECCV*, 2016.

[38] David Novotný, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3DPO: Canonical 3d pose networks for non-rigid structure from motion. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019.

[39] Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas J. Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Trans. Graph.*, 31(4), 2012.

[40] Deva Ramanan. Learning to parse images of articulated bodies. In *Proc. NeurIPS*, 2006.

[41] Jing Ren, Simone Melzi, Maks Ovsjanikov, and Peter Wonka. Maptree: recovering multiple solutions in the space of maps. *ACM Transactions on Graphics (TOG)*, 39(6):1–17, 2020.

[42] Raif M. Rustamov. Laplace-Beltrami eigenfunctions for deformation invariant shape representation. In *Symposium on Geometry Processing*, pages 225–233, 2007.

[43] Artsiom Sanakoyeu, Vasil Khalidov, Maureen S. McCarthy, Andrea Vedaldi, and Natalia Neverova. Transferring dense pose to proximal animal classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[44] Saurabh Singh, Derek Hoiem, and David A. Forsyth. Learning to Localize Little Landmarks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 260–269, 2016.

[45] Jian Sun, Maks Ovsjanikov, and Leonidas J. Guibas. A Concise and Provably Informative Multi-Scale Signature Based on Heat Diffusion. *Computer Graphics Forum*, 28(5):1383–1392, 2009.

[46] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

[47] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *Proc. CVPR*, 2019.

[48] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proc. CVPR*, 2016.

[49] P. Welinder, S. Branson, T. Mita, C. Wah, and F. Schroff. Caltech-ucsd birds 200. Technical report, 2010.

[50] Lei Yang, Wenxi Liu, Zhiming Cui, Nenglun Chen, and Wenping Wang. Mapping in a cycle: Sinkhorn regularized unsupervised learning for point cloud shapes. In *Proc. ECCV*, 2020.

[51] Lu Yang, Qing Song, Zhihui Wang, and Ming Jiang. Parsing r-cnn for instance-level human analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 364–373, 2019.

[52] Ning Zhang, Jeff Donahue, Ross B. Girshick, and Trevor Darrell. Part-based R-CNNs for fine-grained category detection. In *Proc. ECCV*, 2014.

[53] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 117–126, 2016.

[54] Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qi-Xing Huang, and Alexei A. Efros. Learning dense correspondence via 3D-Guided cycle consistency. In *Proc. CVPR*, 2016.

[55] Tinghui Zhou, Yong Jae Lee, Stella X. Yu, and Alexei A. Efros. FlowWeb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *Proc. CVPR*, 2015.

[56] Xiaowei Zhou, Menglong Zhu, and Kostas Daniilidis. Multi-image matching via fast alternating minimization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4032–4040, 2015.

[57] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. ICCV*, 2017.

[58] Silvia Zuffi, Angjoo Kanazawa, Tanya Y. Berger-Wolf, and Michael J. Black. Three-D Safari: Learning to Estimate Zebra Pose, Shape, and Texture from Images "In the Wild". In *International Conference on Computer Vision (ICCV)*, pages 5358–5367, 2019.

[59] Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, and Michael J. Black. 3D Menagerie: Modeling the 3D Shape and Pose of Animals. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5524–5532, 2017.

[60] Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, and Michael J. Black:. Lions and Tigers and Bears: Capturing Non-Rigid, 3D, Articulated Shape from Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3955–3963, 2018.