Reconstructing PASCAL VOC

Sara Vicente*†
University College London

João Carreira*‡
ISR-Coimbra

Lourdes Agapito
University College London

Jorge Batista
ISR-Coimbra

{s.vicente,l.agapito}@cs.ucl.ac.uk

{joaoluis,batista}@isr.uc.pt

Abstract

We address the problem of populating object category detection datasets with dense, per-object 3D reconstructions, bootstrapped from class labels, ground truth figure-ground segmentations and a small set of keypoint annotations. Our proposed algorithm first estimates camera viewpoint using rigid structure-from-motion, then reconstructs object shapes by optimizing over visual hull proposals guided by loose within-class shape similarity assumptions. The visual hull sampling process attempts to intersect an object's projection cone with the cones of minimal subsets of other similar objects among those pictured from certain vantage points. We show that our method is able to produce convincing per-object 3D reconstructions on one of the most challenging existing object-category detection datasets, PASCAL VOC. Our results may re-stimulate once popular geometry-oriented model-based recognition approaches.

1. Introduction

Formerly a dominant paradigm, model-based recognition was permanently upstaged in the 1990's by a flurry of view-based approaches. The main appeal of view-based approaches was their flexibility: collecting a few example images of the target objects and annotating their bounding boxes or 2D keypoint locations became all the manual labor required to build a recognition system, averting the need for cumbersome manual 3D design and for special instrumentation (3D scanners). This more data-driven approach made it possible to attack harder problems such as category-level object recognition, for which huge datasets have been assembled, such as Pascal VOC [12] and Imagenet [11], that have in the order of 1000 objects per category. It would seem very difficult to assemble datasets as natural and diverse as these if 3D wireframe models had to be acquired or designed and registered with the image for every object instance in it. Model-based recognition holds important advantages, nevertheless [20]: modeling the 3D geometry of an object enables arbitrary viewpoints and occlusion patterns to be rendered and recognized, and it also facilitates higher-level reasoning about interactions between objects and a scene.

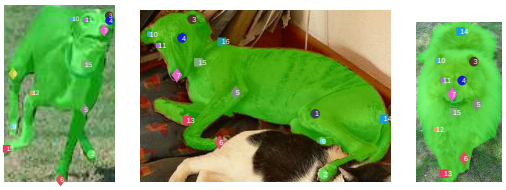

Figure 1. Example inputs of our algorithm, here dog images from the Pascal VOC dataset and their associated figure-ground segmentations and keypoints.

In this paper we attack the 3D model acquisition problem while avoiding any type of special instrumentation or manual model design. We target reconstruction using only annotations available in detection datasets. While our focus is on arguably the most challenging object recognition dataset in the field, PASCAL VOC, our proposed techniques are general and could be applied to any other object detection dataset (e.g. [11]), as long as ground truth figure-ground segmentations and a small number of per-class keypoints are available, as is the case for VOC [14]. These types of annotations can nowadays be easily crowdsourced over Mechanical Turk, as they require only a few clicks per image.

Our approach follows a multiview reconstruction strategy. Unlike settings where multiple calibrated images of the same object are available [15], detection datasets are composed of uncalibrated images of different objects (they are most often assembled from images available on the web). We bypass the problem of establishing point correspondences between different objects, which is still unmanageable with current technology, by relying on a small set of consistent per-class ground truth keypoints, from which scaled orthographic cameras are bootstrapped. We also bypass segmentation, another yet incompletely solved vision problem despite much recent progress [7, 6], and rely on ground truth silhouettes as input to our dense reconstruction

* First two authors contributed equally.
† Now with Anthropics Technology.
‡ Now with the EECS department at UC Berkeley.

engine which is based on a standard visual hull algorithm.

Visual hull computation has been demonstrated to be a simple but powerful reconstruction technique when many diverse views of the same object are available. We adapt it to operate on category detection imagery using a novel formulation we call *imprinted visual hull reconstruction* which we apply within a sampling-based approach: multiple object reconstructions are produced by feeding the visual hull algorithm the reference image and multiple different pairs of other images, among those pictured from vantage points well known to best expose the 3D shape of most objects. Finally, we select the most consistent reconstruction by maximizing intra-category similarity.

Our contributions span different areas of computer vision:

- **For the recognition problem:** a first attempt to semi-automatically augment object detection datasets, here instantiated on PASCAL VOC, with dense per-object 3D geometry and without requiring annotations beyond those readily available online.
- **For the reconstruction problem:** we propose a new data-driven method for class-based 3D reconstruction that relies only on 2D information, such as figure-ground segmentations and a few keypoint annotations.

We have made our full source code and data freely available online[1].

2. Related Work

3D reconstruction is a core problem in computer vision and has been largely solved in the multiview rigid case [15], when calibration and correspondences can be estimated, as it reduces to a well-understood geometric optimization problem. Here we are interested in class-based reconstruction, where the goal is to reconstruct different objects from the same category, each pictured in a single image.

Model-based reconstruction. Most class-based reconstruction methods are *model-based* and rely on prior information about the 3D shape of the object class. This prior information can be, for example, a handcrafted 3D model, such as a kinematic model for human pose estimation, or a morphable model built from 3D training data. Low-dimensional parametric models, or morphable models, have been used to represent the shape of some object classes. They can be built from 3D scans of different instances of the class, e.g. the face model in [3] and the human body model in [1], or using meshes obtained from 3D shape repositories, such as google sketchup [28]. The trained morphable model can then be used to reconstruct from a single image, usually with some user interaction to initialize the viewpoint [3], for reconstruction from a depth map [10], or for performing single-image detection and pose estimation [28]. In order to partially overcome the need for 3D data, [8] proposes a hybrid method that uses a single 3D shape together with 2D information in order to build a morphable model for classes such as dolphins or pigeons.

Data-driven reconstruction. In this paper we focus on a data-driven method for class-based reconstruction that operates directly on an unordered dataset of 2D images and some associated 2D annotations. To the best of our knowledge, there have only been two previous attempts at tackling the problem in a purely data-driven fashion [27, 21]. These two approaches build upon traditional non-rigid structure from motion methods [5], originally developed for reconstruction from video, and either produce sparse reconstructions [27] or have only been demonstrated on simple classes such as flower petals and clown-fish, while requiring complex manual annotations [21].

Our method differs from the above in two important aspects: (1) we require only a small set of keypoint correspondences across images and these are not the only points we reconstruct; instead we reconstruct dense 3D models of the objects, and (2) we do not build a morphable model for the class. Instead, our aim is to reconstruct every object instance, using "borrowed" shape information from a small number of similar instances seen from different viewpoints. This makes our method applicable to classes with large intra-class variation as those in the VOC dataset.

Dataset augmentation into 3D. The goal of populating detection datasets with 3D annotations has been previously considered for the class *person* [4], using an interactive method to reconstruct a set of body joints. In contrast, we obtain full dense reconstructions for a variety of classes. In a related approach, [22] targeted the problem of automatically bootstrapping 3D scene geometry from 2D annotations on the LabelMe dataset — instead, we focus on objects. Recently and perhaps closest to our approach, Karsch *et al.*[17] experimented with reconstructing VOC objects, using manual curvature annotations on boundaries but computed 2.5D reconstructions while we focus on the full 3D problem.

3. Problem formulation

We assume we are given a set of images depicting different instances of the same object class, which may be very diverse in terms of object scale, location, pose and articulation. We make the small simplification in this paper of not addressing the problem of reconstructing occluded objects, that are marked as such in PASCAL. However, we would like to point out that, in principle, our proposed techniques could handle the case of occlusions. Each object instance n has a corresponding binary mask B_n, a figure-ground segmentation locating the object boundaries in the image, and K_i specific keypoints for each class i, which are on easily identifiable parts of the object, such as "left mirror" for cars or "nose tip" for aeroplanes. Each object instance n is an-

[1] http://www.isr.uc.pt/~joaoluis/carvi

notated with its visible keypoints, i.e. the set (x_k^n, y_k^n) of 2D image coordinates[2].

Our goal in this paper is to generate a dense 3D reconstruction of each of the object instances. It is easy to see that this is a severely underconstrained problem since each image corresponds to a different object instance. Without additional prior knowledge, and if each instance is to be reconstructed independently, an infinite number of reconstructions would be available that could exactly generate the silhouette B_n.

3.1. Our data-driven approach

Instead of relying on single view reconstruction methods and performing reconstruction completely independently for each instance, we leverage the information contained in the collection of images showing objects from the same category, by building upon the assumption that at least some instances of the same class have a similar 3D shape. We propose a feedforward strategy with two phases: first, orthographic cameras for all objects are estimated using both keypoint and silhouette information, then a sampling-based approach employing a novel variation of visual hull reconstruction is used to produce dense per-object 3D reconstructions. These two phases will be explained in the following two sections.

4. Camera viewpoint estimation and refinement

The first step of our algorithm is to estimate the camera viewpoint for each of the instances using a factorization based rigid structure from motion algorithm [19]. Although this might appear to be a suboptimal choice, several non-rigid structure from motion algorithms make use of a similar strategy in viewpoint estimation due to the lack of robustness to noise of specialized non-rigid SFM viewpoint estimates. This particular algorithm assumes that the objects are observed by a scaled orthographic camera and requires point correspondences across the different instances.

Using the annotated keypoints we form an observation matrix for each instance:

$$W_n = \begin{bmatrix} x_n^1 & \dots & x_n^K \\ y_n^1 & \dots & y_n^K \end{bmatrix} \quad (1)$$

The SFM algorithm finds the 3D shape S, a $3 \times K$ matrix that can be seen as a rough "mean shape" for the objects of the class, motion matrices M_n and the translation vectors T_n, by minimizing the reprojection error:

$$\sum_{n=1}^{N} \left\| W_n - \begin{bmatrix} M_n & T_n \end{bmatrix} \begin{bmatrix} S \\ \mathbf{1}_{1\times K} \end{bmatrix} \right\|_F^2 \quad (2)$$

[2]These annotations are publicly available for all the 20 classes in the VOC dataset [14].

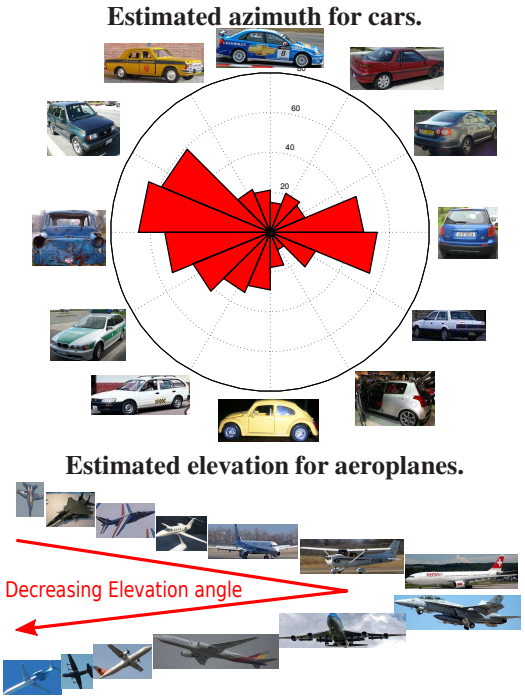

Figure 2. Results of the camera viewpoint estimation. Our method provides useful insight about viewpoint distribution for the different classes in VOC. Here, we show the histogram of different azimuths for "car" and a few samples of estimated elevation angle for "aeroplane". Note that there is significant intra-class variation.

under the constraint that $M_n M_n^T = (\alpha_n)^2 I_{2\times 2}\ \forall n$. This constraint guarantees that matrices M_n correspond to the first two rows of a scaled rotation matrix, that can be easily converted into a full rotation matrix R_n and scale parameter α_n. The SFM algorithm used does not require that all keypoints are visible in all the instances, i.e. it can deal with missing data. Similar to [19], we use an iterative method with power factorization to minimize the reprojection error.

For classes with large intra-class variation or articulation, we manually select a subset of the keypoints to perform rigid SFM. There are two types of classes that follow this behaviour: the class *boat* and animal classes. The class *boat* includes both sailing boats and motor boats and since the sails are not present in the motor boats, we estimate the camera by only considering the keypoints on the hull. For animals, which undergo articulation, different instances may have very different poses. For these classes, we assume that the camera viewpoint is defined with respect to the head and torso and exclude the keypoints corresponding to the limbs or wings when performing rigid SFM. For all classes, for robustness, we double the number of instances by adding left-right flipped versions of each image.

4.1. Silhouette-based camera refinement

To obtain a camera estimate for a particular instance, the SFM algorithm only uses the keypoints visible in that in-

stance. However, if we assume that S is a reasonable approximation of the shape of all the objects in the class, we can refine the camera estimation by imposing extra constraints on the keypoints which are not visible. In particular, all the keypoints even the ones which are not visible, should reproject inside the silhouette. We refine the camera estimate M_n and T_n by fixing the shape S and minimizing an energy function of the form:

$$E(M_n, T_n) = \left\| W_n - [M_n\ T_n] \begin{bmatrix} S \\ \mathbf{1}_{1\times K} \end{bmatrix} \right\|_F^2 + \mathrm{D}\left([M_n\ T_n] \begin{bmatrix} S \\ \mathbf{1}_{1\times K} \end{bmatrix} \right) \tag{3}$$

under the constraint $M_n M_n^T = (\alpha_n)^2 I_{2\times 2}$. The first term of this energy is the reprojection error as in (2) and the second term is defined as:

$$\mathrm{D}\left([M_n\ T_n] \begin{bmatrix} S \\ \mathbf{1}_{1\times K} \end{bmatrix} \right) = \mathrm{D}\left(\begin{bmatrix} u^1 & \dots & u^K \\ v^1 & \dots & v^K \end{bmatrix} \right) = \sum_{k=1}^{K} \mathrm{C}\left(u^k, v^k\right) \tag{4}$$

where $\mathrm{C}(.,.)$ is the Chamfer distance map from the figure-ground segmentation B_n. A point on the mean shape S incurs a penalty if its reprojection, given by (u^k, v^k), is outside the silhouette. To minimize this function, we use gradient descent with a projection step into scaled-Stiefel matrices. A similar projection step is used in [19]. Qualitative results of our camera viewpoint estimation can be seen in fig. 2.

This camera refinement step can also be used to estimate the camera viewpoint of a new test image, by initializing M_n to the identity matrix and T_n to the center of the mask. This allows our method to reconstruct a previously unseen image, the only requirement being that the keypoints are marked and the object is segmented.

5. Object reconstruction

After jointly estimating the camera viewpoints for all the instances in each class, we reconstruct the 3D shape of all objects using shape information borrowed from other exemplars in the same class.

5.1. Sampling shape surrogates

In datasets as diverse as VOC, it is reasonable to assume that for every instance n there are at least a few shape surrogates, i. e. other instances of the same class that, despite not corresponding to the same physical object, have a similar 3D shape. Finding shape surrogates is not straightforward, however. When the surrogates have very different viewpoint it is difficult to establish that their 3D shape is similar to the shape of the reference object (e.g. that they are true surrogates). A tension also exists between reconstructing from fewer silhouettes, which may result in a solution with many uncarved voxels, and a large number of silhouettes which may lead to an over-carved or even empty solution, because calibration is not exact and "surrogateness" is only approximate. Here we strike a compromise: we sample groups of three views, where the two surrogates of the reference instance n are selected among those pictured from far apart viewpoints, so as to maximize the number of background voxels carved away (see fig. 3).

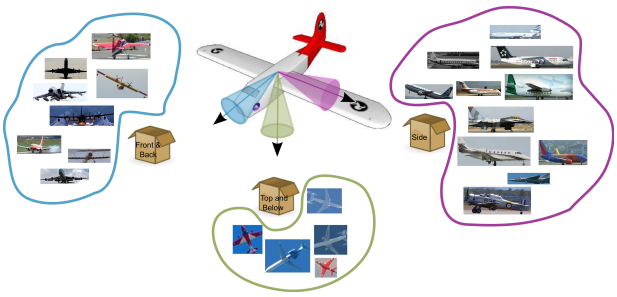

Figure 3. Illustration of our clustering step. Instances which have a viewpoint similar with one of the principal directions are clustered together. We sample from these clusters to generate informative triplets of exemplars for visual hull computation. The process is repeated multiple times for each target object.

Furthermore, when selecting *far apart* viewpoints we took inspiration from technical illustration practices, where the goal is to communicate 3D shape as concisely as possible, and it is common practice to represent the shape by drawing 3D orthographic projections on three orthogonal planes. In a similar vein, we restrict surrogate sampling to be over objects pictured from three orthogonal viewpoints, which we will call principal directions.

Our sampling process has three steps:

(1) Principal direction identification We found empirically that a good set of principal directions is obtained by computing the three principal components of the set of reconstructed keypoints S from rigid structure from motion, using PCA.

(2) Clustering instances around the principal directions Instances that have a viewpoint within a 15° difference to one of the principal directions are clustered together. All other instances are never chosen as surrogate views. An illustration of this clustering step for "aeroplanes" is shown in fig. 3.

(3) Sampling We start by selecting two of the three principal directions, with a probability proportional to the number of instances associated with each. Then, from each of the selected principal directions, we sample one surrogate instance, which together with the reference instance forms a triplet of views.

For three of the classes in the VOC dataset (*bottle*, *dining table* and *potted plant*) the keypoints are view-dependent since the classes have rotational symmetry [14] and 3D registration for all the instances of the class is ambiguous. Furthermore, for these classes, at least one of the principal directions will have no instances associated with it. Instead of sampling surrogate instances, we use the fact that they are

symmetric and synthesize the surrogates from the reference instance by rotating it around the axis of symmetry, every 45 degrees.

5.2. Imprinted visual hull reconstruction

Recovering the approximate shape of an object from silhouettes seen from different camera viewpoints can be done by finding the visual hull of the shape [18], the reconstruction with maximum volume among all of those that reproject inside all the different silhouettes. Visual hull reconstruction is a frequent first step in multi-view stereo [23], providing an initial shape that is then refined using photo-consistency. Existing visual hull methods assume that the different silhouettes project from the same physical 3D object [13]. This is in contrast with our scenario where images of different objects are considered. Visual hull reconstruction is known to be sensitive to errors in the segmentation and in the viewpoint estimate and it is clear that such sources of noise are very present in our framework, and can lead to overcarving if handled naively.

A clear inefficiency of using the standard visual hull algorithm in our setting is that there is no guarantee that the visual hull is silhouette-consistent with the reference instance n, i.e. that for all the foreground pixels in the mask B_n there will be an active voxel reprojecting on them. This happens because the algorithm trusts equally all silhouettes. Here we propose a variation of the original formulation that does not have this problem, that we denote *imprinted visual hull reconstruction*. We will use a volumetric representation of shape and formulate imprinted visual hull reconstruction as a binary labelling problem. Let T be the set of instances corresponding to a sampled triplet and $\mathcal{V}$ be a set of voxels. The goal is to find a binary labelling $L = \{l_v : v \in \mathcal{V}, l_v \in \{0,1\}\}$ such that $l_v = 1$ if voxel v is inside the shape, and $l_v = 0$ otherwise. Let $C_m(.)$ be a signed distance function to the camera cone of instance m, so that $C_m(v) < 0$ if voxel v is inside the camera cone, and $\bar{C}(v) = \max_{m \in T} C_m(v)$ the largest signed distance value over all the cameras, for each voxel v. Visual hull reconstruction can be formulated as the minimization of the energy:

$$E(L) = \sum_{v \in \mathcal{V}} l_v \bar{C}(v) \tag{5}$$

To enforce silhouette consistency with the reference mask B_n (imprinting), we need to guarantee that all the rays cast from the foreground pixels of B_n intersect with an interior voxel. Let R_p be the set of voxels that intersect with the ray corresponding to pixel p. Imprinting is then enforced by minimizing energy (5) under the following constraints:

$$\sum_{v \in R_p} l_v \geq 1 \quad \forall\, p \in \text{Foreground}(B_n). \tag{6}$$

Similar constraints have been previously used for multi-view stereo [9], where they were enforced equally for all the images. Energy (5) can be minimized exactly under constraint (6), by simply setting $l_v^* = 1$ if and only if $\bar{C}(v) < 0$ or if $\exists p, v = \arg\min_{u \in R_p} \bar{C}(u)$. We choose to formulate our reconstruction algorithm as a labelling problem, to motivate future extensions such as adding pairwise constraints between voxels or connectivity priors [26].

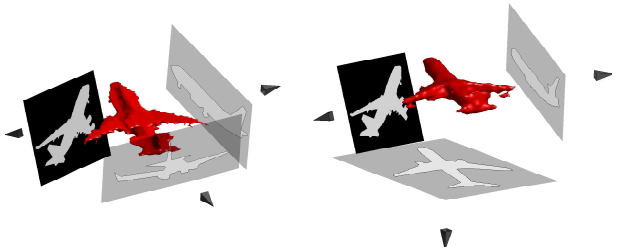

Figure 4. Illustration of the imprinted visual hull reconstruction method, for two different triplets corresponding to the same reference instance (in black). The reconstructions are obtained by intersecting the three instances shown and their left-right flipped versions.

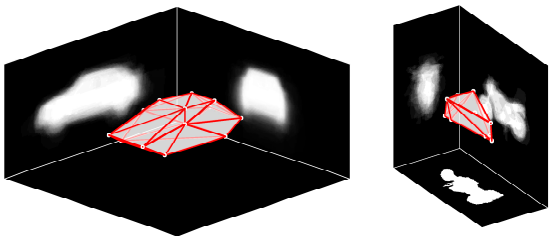

Figure 5. Average mask for each of the principal directions for the car and motorbike classes, as well as the convex hull of the 3D keypoints obtained with SFM. These average masks are used when ranking the reconstructions for a single instance. Note that for the class car, there is no instance associated with the top-bottom axis and for motorbike there is only one instance.

5.3. Reconstruction selection

Once all reconstruction proposals have been computed based on different sampled triplets, the final step is to choose the best reconstruction for the reference instance. Here we propose a simple selection criterion reflecting a natural assumption: reconstructions should be similar to the average shape of their object class. Our selection procedure first converts the voxel-based reconstructions into meshes, then computes an average mask for each of the principal directions. This is done by aligning all the masks of the instances in that bin and averaging them. Afterwards, the reconstruction is projected into a plane perpendicular with the principal direction and the difference between this projection and the average mask associated with that direction is measured. The final score is the sum of the three differences, one for each direction. The average masks for each principal direction for two classes are shown in fig. 5.

6. Experiments

We consider the subset of 9,087 fully visible objects in 5,363 images from the 20,775 objects and 10,803 images available in the PASCAL VOC 2012 training data. We use the publicly available keypoints and figure-ground segmentations [14]. VOC has 20 classes, including highly articulated ones (dogs, cats, people), vehicles (cars, trains, bicycles) and indoor objects (dining tables, potted plants) in realistic images drawn from FLICKR. Amongst these, fewer than 1% have focal lengths in their EXIF metadata, which we ignored.

We reconstructed all the objects and show two example outputs from each class in fig. 7. A much larger subset of our reconstructions can be found in the supplemental material. We observe that surprisingly accurate reconstructions are obtained for most classes, with some apparent difficulties for "dining table", "sofa" and "train". The problems with "dining table" can be explained by there being only 13 exemplars marked as unoccluded, which makes camera viewpoint estimation frail. "Sofa" has a strong concavity which makes visual-hull reconstruction hard and would benefit from stereo-based post-processing, which we leave for future work. "Train" is a very difficult class to reconstruct in general: different trains may have a different number of carriages, there are strong perspective effects and it is articulated. Finally, sometimes our reconstructions of animals have either fewer or more limbs than in the image, and certain reconstructions have disconnected components.

In all experiments, we sampled 20 reconstructions of each reference object instance and we found our algorithm to be very efficient: it took just 7 hours to reconstruct VOC on a 12-core computer, with the camera refinement algorithm taking around 5 hours.

6.1. Synthetic test data

We also performed a quantitative evaluation on synthetic test images with similar segmentations and keypoints as those in VOC. To make results as representative of performance on real data as possible, we reconstruct using only surrogate shapes from VOC. We downloaded 10 meshes for each category from the web, then manually annotated keypoints consistent with those of [14] in 3D and rendered them using 5 different cameras, sampled from the ones estimated on VOC for that class. This resulted in 50 synthetic images per class, each with associated segmentation and visible keypoints, for a total of 1000 test examples. More details can be found in the supplemental material.

We measure the distortion between a reconstruction and a ground truth 3D mesh using the root mean squared error between the two meshes [2]. We normalize scale using the diagonal length of the bounding box of the ground truth 3D model, such that the error is a percentage of this length, and report the average error over all the objects in each category. Table 1 demonstrates the benefits of the different components of our proposed methodology. Since no other existing class reconstruction technique scales to such a large and diverse dataset using simple 2D annotations we compare to two simple baselines: an inflation technique originally proposed for silhouette based single-view reconstruction [25] and a multi-view baseline based on our rigid SFM. Our method is significantly better for most classes, and a visual comparison of resulting reconstructions obtained is available in the supplementary material. Fig. 6 suggests large gains of our simple ranking approach over random selection but also that there is much to improve with the addition of more advanced features.

	Full	-CRef	-SImp	[25]	SFMc
aeroplane	**3.58**	4.94	3.95	9.64	5.79
bicycle	4.30	**3.26**	4.75	10.51	6.56
bird	9.98	10.92	10.34	**8.76**	12.01
boat	**5.91**	6.78	6.05	8.81	6.52
bottle	8.09	10.77	8.53	**6.25**	12.13
bus	6.45	**6.10**	6.49	11.02	7.34
car	**3.04**	6.33	3.10	11.07	3.22
cat	**6.98**	7.57	7.49	11.39	9.61
chair	**5.36**	5.73	6.06	8.13	7.37
cow	5.44	**5.24**	5.83	9.17	7.50
diningtable	8.97	12.57	14.30	**8.67**	9.52
dog	**7.08**	8.38	7.19	11.61	9.91
horse	**6.05**	7.05	6.38	6.90	7.41
motorbike	**4.12**	4.24	4.16	9.24	5.32
person	**7.35**	7.95	7.55	9.14	19.46
pottedplant	7.72	8.15	7.99	**7.58**	17.86
sheep	7.18	**7.15**	7.66	8.77	7.16
sofa	6.11	6.24	6.31	8.06	**5.75**
train	**15.73**	20.55	16.19	17.01	17.47
tv/monitor	9.73	10.45	10.28	**9.67**	10.08
Mean	**6.96**	8.01	7.53	9.57	9.40

Table 1. Root mean square error between reconstructed and ground truth 3D models. Lowest errors are displayed in bold. We compare our full model (Full), with severed versions without our proposed camera refinement process (-CRef) and reference silhouette imprinting (-SImp). As baselines we consider a recent single view silhouette-based reconstruction method [25] and the convex hull of the 3D points returned by our rigid structure from motion component (SFMc).

7. Conclusion

We have proposed a novel data-driven methodology for bootstrapping 3D reconstructions of all objects in detection datasets, based on a small set of commonly available annotations, namely figure-ground segmentations and a small set of keypoints. Our approach is the first to target class-based 3D reconstruction on a challenging detection dataset, PASCAL VOC, and is demonstrated to achieve very promising performance. It produces recognizable 3D shapes for

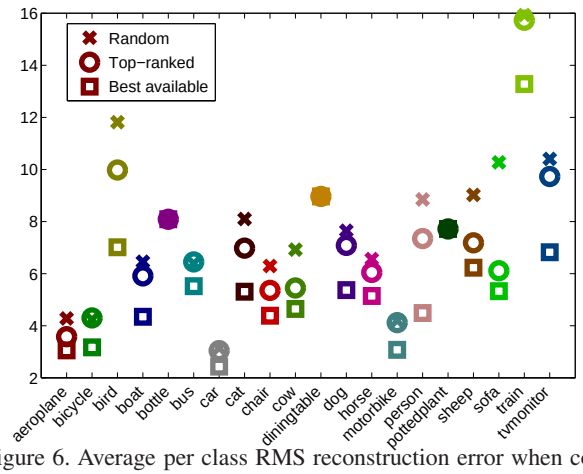

Figure 6. Average per class RMS reconstruction error when considering the top ranked, randomly selected and best available reconstructions for each individual object. "Random" and "Best available" represent, respectively, lower and upper bounds on ranking performance.

most categories, handling widely different objects such as animals, vehicles and indoor furniture using the same integrated framework. We believe this paper contributes to the recently renewed interest in 3D modeling in recognition (eg. [16, 24]) and that it will facilitate progress in this direction since it provides the first semi-automatic solution to 3D model acquisition of detection data, which has possibly been the main obstacle to any previous attempts to 3D model-based recognition. As future work we plan to develop more advanced features to rank reconstructions, better surrogate shape sampling approaches and more constrained forms of our imprinted visual hull optimization.

Acknowledgments. This work was supported by FCT grants PTDC/EEA-CRO/122812/2010 and SFRH/BPD/84194/2012 and by the European Research Council under the ERC Starting Grant agreement 204871-HUMANIS.

References

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. In *ACM Trans. Graph.*, 2005. 2

[2] N. Aspert, D. Santa-Cruz, and T. Ebrahimi. Mesh: Measuring errors between surfaces using the hausdorff distance. In *ICME*, 2002. 6

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999. 2

[4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 2

[5] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 2000. 2

[6] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. 1

[7] J. Carreira, F. Li, and C. Sminchisescu. Object Recognition by Sequential Figure-Ground Ranking. *IJCV*, 2012. 1

[8] T. J. Cashman and A. W. Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. *TPAMI*, 2013. 2

[9] D. Cremers and K. Kolev. Multiview stereo and silhouette consistency via convex functionals over convex domains. *TPAMI*, 2011. 5

[10] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. Reid. Dense reconstruction using 3d object shape priors. In *CVPR*, 2013. 2

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 1

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 1

[13] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3d structure with a statistical image-based shape model. In *CVPR*, 2003. 5

[14] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 1, 3, 4, 6

[15] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. 2004. 1, 2

[16] D. Hoiem and S. Savarese. *Representations and techniques for 3D object recognition and scene interpretation*, volume 15. Morgan & Claypool Publishers, 2011. 7

[17] K. Karsch, Z. Liao, J. Rock, J. T. Barron, and D. Hoiem. Boundary cues for 3d object shape recovery. In *CVPR*, 2013. 2

[18] A. Laurentini. The visual hull concept for silhouette-based image understanding. *TPAMI*, 1994. 5

[19] M. Marques and J. P. Costeira. Estimating 3D shape from degenerate sequences with missing data. *CVIU*, 2008. 3, 4

[20] J. L. Mundy. Object recognition in the geometric era: A retrospective. In *Toward Category-Level Object Recognition*, 2006. 1

[21] M. Prasad, A. Fitzgibbon, A. Zisserman, and L. Van Gool. Finding nemo: Deformable object class modelling using curve matching. In *CVPR*, 2010. 2

[22] B. C. Russell and A. Torralba. Building a database of 3d scenes from user annotations. In *CVPR*, 2009. 2

[23] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*. IEEE, 2006. 5

[24] M. Sun, H. Su, S. Savarese, and L. Fei-Fei. A multi-view probabilistic model for 3d object classes. In *CVPR*, June 2009. 7

[25] N. R. Twarog, M. F. Tappen, and E. H. Adelson. Playing with puffball: simple scale-invariant inflation for use in vision and graphics. In *ACM Symp. on Applied Perception*, 2012. 6

[26] S. Vicente, V. Kolmogorov, and C. Rother. Graph cut based image segmentation with connectivity priors. *CVPR*, 2008. 5

[27] S. Zhu, L. Zhang, and B. Smith. Model evolution: An incremental approach to non-rigid structure from motion. In *CVPR*, 2010. 2

[28] M. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3d representations for object recognition and modeling. *TPAMI*, 2013. 2

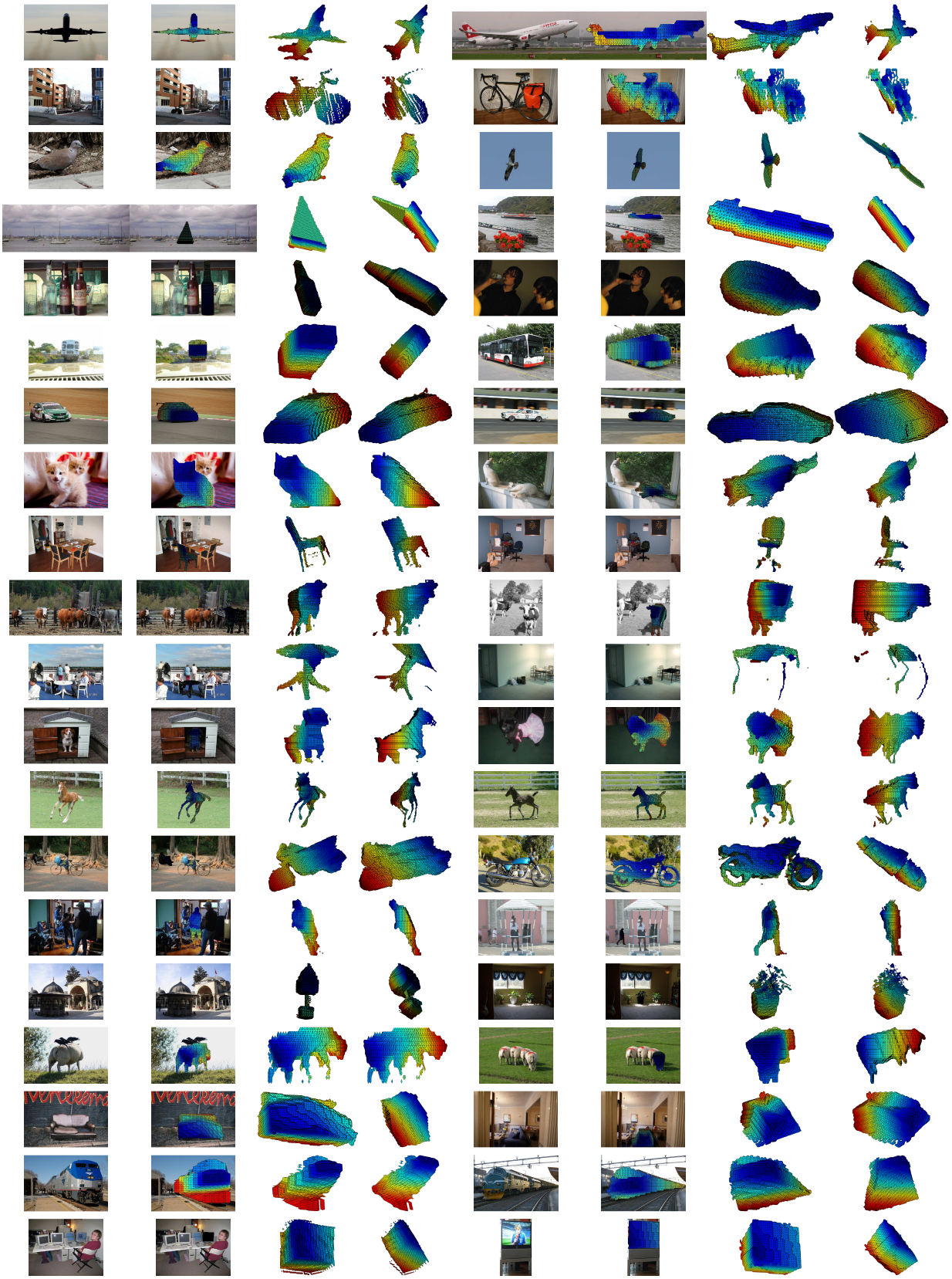

Figure 7. Examples of our reconstructions for all 20 PASCAL VOC categories. For each object we show the original image, the original image with the reconstruction overlaid and two different viewpoint of our reconstruction. Blue is closer to the camera, red is farther (best seen in color). For most classes, our reconstructions convey the overall shape of the object, which is a remarkable achievement given the limited information used as input and the large amount of intra-class variation.