Pose Induction for Novel Object Categories

Shubham Tulsiani, João Carreira and Jitendra Malik
University of California, Berkeley
{shubhtuls,carreira,malik}@eecs.berkeley.edu

Abstract

We address the task of predicting pose for objects of unannotated object categories from a small seed set of annotated object classes. We present a generalized classifier that can reliably induce pose given a single instance of a novel category. In case of availability of a large collection of novel instances, our approach then jointly reasons over all instances to improve the initial estimates. We empirically validate the various components of our algorithm and quantitatively show that our method produces reliable pose estimates. We also show qualitative results on a diverse set of classes and further demonstrate the applicability of our system for learning shape models of novel object classes.

1. Introduction

Class-based processing significantly simplifies tasks such as object segmentation [17, 4], reconstruction [6, 21, 38] and, more generally, the propagation of knowledge from class objects we have seen before to those we are seeing for the first time. Looking at the lion in Figure 1 humans can not only easily perceive its shape, but also tell that it is strong and dangerous, get an estimate of its weight and dimensions and even approximate age and gender. We get to know all of this because it is a lion like others we have seen before and that we know many facts about.

Despite its many virtues, class-based processing does not scale well. Learning predictors for all variables of interest – figure-ground segmentation, pose, shape – requires expensive manual annotations to be collected for at least dozens of examples per class and there are millions of classes. Consider again Figure 1 but now look at object A. The underlying structure in our visual world allows us to perceive a rich representation of this object despite encountering it for the first time. We can infer that it is probably hair that covers its surfaces – we have seen plenty of hair-like materials before – and that it has parts and determine their configuration by analogy with our own parts or with other animals. We are able to achieve this remarkable feat by lever-

Figure 1. Inductive pose inference for novel objects. Right : Novel object A. Left : instances from previously seen classes having similar pose as object A.

aging commonalities across object categories via generalizable abstractions – not only can we perceive that all the other animals in Figure 1 are "right-facing", we can also transfer this notion to object A. This type of cross-category knowledge transfer has been successfully demonstrated before for properties such as materials [37, 8], parts [35, 10] and attributes [22, 13].

In this paper we define and attack the problem of predicting object poses across categories – we call this pose induction. The first step of our approach, as highlighted in Figure 2, is to learn a generalizable pose prediction system from the given set of annotated object categories. Our main intuition is that most objects have appearance and shape traits that can be associated with a generalized notion of *pose*. For example, the sentences "I am in front of a car" or "in front of a bus" or "in front of a lion" are clear about where"I" am with respect to those objects. The reason for this may be that there is something generic in the way"frontality" manifests itself visually across different object classes – e.g."fronts" usually exhibit an axis of bilateral symmetry. Pushing this observation further leads to our solution: to align all the objects in a small *seed* set of classes, by endowing them with set of 3D rotations in a consistent reference frame, then training pose predictors that generalize in a meaningful way to novel object classes.

This idea expands the current range of inferences that can be performed in a class-independent manner and allows us to reason about pose for every object without tediously collecting pose annotations. Such pose based reasoning can then inform a system about which directions objects are most likely to move in (usually "front" or "back") and hence

---

Our implementations and trained models are available at https://github.com/shubhtuls/poseInduction

Figure 2. Overview of our approach. We first induce pose hypotheses for novel object instances using a system trained over aligned annotated classes (Section 2). We then reason jointly over all instances of the novel object class to improve our pose predictions (Section 3).

allow it to get out of their way; it can help to identify how to place any object on top a surface in a stable way (by identifying the "bottom" of the object). Ultimately, and the main motivation for this work, it provides important cues about the 3D shape of a novel object and may allow bypassing the existing need for ground truth keypoints in training data for state-of-the-art class-specific object reconstruction systems [21, 38] – we will present a proof of concept for this in Section 4.

Related Work. The problem of generalizing from a few examples [34] was already studied in ancient Greece and has become known as *induction*. Early induction work in computer vision pursued feature sharing between different classes [1, 35]. One-Shot and Zero-Shot learning [14, 26] also represent related areas of research where the task is to learn to predict labels from very few exemplars. Our work differs from these as, in constrast to these approaches, the *few examples* we consider correspond to a small set of annotated object categories. In this sense, our approach is perhaps closer in style to *attributes* [13, 22], which explicitly learn classifiers that are transversal to object classes and can hence be trained on a subset of object classes. Differently, our "attributes" correspond to a dense discretization of the viewpoint manifold that implicitly aligns the shapes of all training object classes. Another relevant recent work, LSDA [18] learns object detectors using a seed set of classes having bounding box annotations. Unlike our work, they leverage available data for a related task (classification) and frame the task as adapting classifiers to object detectors.

Pose estimation is crucial for developing a rich understanding of objects and is therefore an important component of systems for 3D reconstruction [21, 5], recognition [25, 33], robotics [30] and human computer interaction [24, 29]. Traditional approaches to object pose estimation predicted instance pose in context of a corresponding shape model [19]. The task has recently evolved to the prediction of category-level pose, a problem targeted by many recent methods [36, 28, 16]. Motivated by Palmer's experiments which demonstrate common canonical frames for similar categories [27], we reason over cross-category pose - our work can be thought of as a natural extension in the current

paradigm shift of pose prediction from instances/models to categories.

2. Pose Induction for Object Instances

We noted earlier that humans have the ability to infer rich representations, including pose, even for previously unseen object classes. These observations demonstrate the applicability of human inductive learning as a mechanism to infer desired representations for new visual data. We explore the possibility of applications of such ideas to induce the notion of pose for previously unseen object instances. More concretely, we assume pose annotations for some object classes and aim to infer pose for an object instance belonging to a different object category. We describe our formulations and approach below.

2.1. Formulation

Let C denote the set of object categories with available pose annotations. We follow the pose estimation formulation of Tulsiani and Malik [36] who characterize pose via $N_a = 3$ euler angles - azimuth (ϕ), elevation(φ) and cyclo-rotation(ψ). We discretize the space of each angle in N_θ disjoint bins and frame the task of pose prediction as a classification problem to determine the angular bin for each euler angle. Let $\{x_i | i = 1 \ldots N_i\}$ denote the set of annotated instances, each with its object class $c_i \in C$, with pose annotations ($\phi_i, \varphi_i, \psi_i$). The pose induction task is to predict the pose for a novel instance x whose object class $c \notin C$.

2.2. Approach

We examine two different approaches for inducing pose for a novel instance - 1) the baseline approach of explicitly leveraging the inference mechanism for similar object classes and 2) our proposed approach of enforcing the inference mechanism to implicitly leverage similarities between object classes and thereby allowing generalization of inference to novel instances.

Similar Classifier Transfer (SCT). We first describe the baseline approach which infers pose for instances of an unannotated class by explicitly using similarity to some annotated object category and obtaining predictions using a

system trained for a visually similar class. To obtain a pose prediction system for the annotated classes C, we follow the methodology of Tulsiani and Malik [36] and train a VGG net [31] based Convolutional Neural Network (CNN) [15, 23] architecture with $|C| * N_a * N_\theta$ output units in the last layer. Each output unit corresponds to a particular object class, euler angle and angular bin - this CNN system shares most parameters across classes but has some class-specific parameters and disjoint output units. Let $f(x; W_c)$ denote the pose prediction function for image x and class-specific CNN weights W_c, then $f(x_i, W_{c_i})$ computes the probability distribution over angular bins for instance i - the CNN is trained to minimize the softmax loss corresponding to the true pose label $(\phi_i, \varphi_i, \psi_i)$ and $f(x_i, W_{c_i})$.

To predict pose for an instance x with class $c \notin C$, this approach uses the prediction system for a visually similar class c'. We obtain the probability distribution over angular bins for this instance by computing $f(x, W_{c'})$. We then use the most likely hypothesis under this distribution as our pose estimate for the instance x.

Generalized Classifier (GC). To infer properties for a novel instance, our proposed approach is to rely not only on the most similar visual object class, but also on general abstractions from all visual data - seeing a sheep for the first time, one would not just use knowledge of a specific class like cows, but also generic knowledge about four-legged animals. For example, the concept that pose of animals can be determined using generic part representations (head, torso etc.) can be learned if the annotations share a common canonical reference frame across classes and this notion can then be applied to novel related classes. These observations motivate us to consider an alternate approach, termed as Generalized Classifier (GC), where we train a system that exploits consistent visual similarities across object classes that coherently change with the pose label. This approach not only bypasses the need for manually assigning a visually similar class, it can also potentially learn abstractions more generalizable to unseen data and therefore handle novel instances more robustly.

Concretely, we first obtain pose annotations across object classes wrt a common canonical frame (details described in experimental section) and train a category-agnostic pose estimation system. This implicitly enforces the CNN based pose estimation system to exploit similarities across object classes and learn common representations that may be useful to predict pose across object classes. We train a VGG net [31] based CNN architecture with $N_a * N_\theta$ output units in the last layer - the units corresponds to a particular euler angle and angular bin are shared across all classes. Let $f(x; W)$ denote the pose prediction function for image x and CNN weights W, then CNN is trained to minimize the softmax loss corresponding to the true pose label $(\phi_i, \varphi_i, \psi_i)$ and $f(x_i, W)$. To predict pose

for an instance x of an unannotated class c, we just compute $f(x; W)$ - the alignment of all annotated classes to a canonical pose and implicit sharing of abstractions allow this system to generalize well to new object classes.

2.3. Experiments

Pose Annotations and Alignment. We evaluate the performance of our system on PASCAL VOC [11] object categories. We obtain pose annotations for rigid categories via the PASCAL3D+ [39] dataset which annotates instances in PASCAL VOC and Imagenet dataset with their euler angles. The notion of a global viewpoint is challenging to define for various animal categories in PASCAL VOC and we apply SfM-based techniques on ground truth keypoints to obtain the torso pose. We use keypoints annotations provided by Bourdev *et al*. [3] followed by rigid factorization [38] to obtain viewpoint for non-rigid pascal classes. The PASCAL3D+ annotations assume a canonical reference frame across classes - objects are laterally symmetric across X axis and face frontally in the canonical pose. We obtain similarly aligned reference frames for other object classes by aligning the SfM models to adhere to this constraint.

Evaluation Setup. We held out pose annotations for four object classes - bus, dog, motorbike and sheep. We then finetuned the CNN systems, after initializing weights using a pretrained model for Imagenet [9] classification, corresponding to the two approaches described above using pose annotations for the remaining 16 classes obtained via PASCAL3D+ or PASCAL VOC keypoint labels.

To evaluate the performance of our system for rigid objects, we used the Acc_θ metric [36] which measures the fraction of instances whose predicted viewpoint is within a fixed threshold of the correct viewpoint (we use $\theta = \frac{\pi}{6}$). The 'ground-truth' viewpoint obtained for some classes via SfM techniques is often noisy and the above metric which works well for exact annotations needs to be altered. To evaluate the system's performance for these classes, we use an auxiliary task of predicting the 'frontal/left/right/rear-facing' label available in PASCAL VOC for these objects. We use our predicted azimuth for these objects and infer the 'frontal/left/right/rear-facing' label based on the predicted azimuth. We denote the metric that measures accuracy at this auxiliary task as Acc_v.

Results. We report the performance our baseline and proposed approach in Table 1. For the SCT method, we used the weights from car, bicycle, cat and cow prediction systems to predict pose for bus, motorbike, dog and sheep respectively since these correspond to the visually most similar classes with available annotations. We note that the predictions using both approaches are often very close to the actual object pose and are significantly better than chance. We also observe that training a generalized prediction sys-

tem is better than explicitly using a similar class (except for motorbike, where the bicycle class is very similar). This is perhaps because sharing of parameters and output units across classes enables learning shared abstractions that generalize better to novel classes.

	$Acc_{\frac{\pi}{6}}$		Acc_v	
Approach	bus	mbike	dog	sheep
SCT	0.50	**0.58**	**0.75**	0.58
GC	**0.80**	0.55	0.74	**0.78**

Table 1. Performance of our approaches for various novel object classes.

We have described a methodology that aims to provide a richer description, in particular pose, given a single instance belonging to an novel class. We note that though human levels of precision and understanding for novel objects are still far away, the results imply that we can reliably predict pose without requiring training annotations, which is a step in the direction of visual systems capable of dealing with new instances.

Importance of Similar Object Categories. To further gain insight into our prediction system, we focused on the 'bus' object category and trained two additional networks for the GC method by holding out 'car' and 'chair' respectively (in addition to the four held out categories above). In comparison to $Acc_{\frac{\pi}{6}} = 0.80$, the $Acc_{\frac{\pi}{6}}$ measure for bus in these two cases was 0.73 and 0.81 respectively. The observed drop by holding out 'car' confirms our intuition regarding the importance of similar object categories in the seed set.

3. Pose Induction for Object Categories

When reasoning over a single instance of a novel category, any system, including the approaches in Section 2, can only rely on inference and abstractions on previously seen visual data. However, if given at once a collection of instances belonging to the new category, we can infer pose for all instances of the object class under consideration while reasoning jointly over all of their poses. This allows us to go beyond isolated reasoning for each instance and leverage the collection of images to jointly reason over and infer pose for all instances of the object class under consideration. Tackling the problem of inducing pose at a category level is particularly relevant as pose annotations for objects are far more tedious to collect than class labels – there are significantly more datasets with annotated classes than pose. Our method allows us to augment these available datasets with a notion of pose for each object. Our method can also be used in a completely unsupervised setting to infer pose for consistent visual clusters over instances that visual knowledge extraction systems like NEIL [7] automatically discover.

One possible approach to reasoning jointly is to explicitly infer intra-class correspondences, predict relative transformations and augment these with the induced instance predictions to obtain more informed pose estimates for each instance. However, the task of discovering correspondences across instances that differ in both pose and appearance, is a particularly challenging one and has been demonstrated only in limited pose and appearance variability [40, 32]. Our proposed approach provides a simpler but more robust way of leveraging the image collection. We build on the intuition that instances with similar spatial distributions of parts are close on the pose manifold. We define a similarity measure that captures this intuition and encourage similar instances to have similar pose predictions.

Algorithm 1 Joint Pose Induction

INITIALIZATION
    for i in test instances **do**
        Predict pose distribution $F(x_i; W)$ (Section 2)
        Compute K pose hypotheses and likelihood scores
        $\{(R_{ik}, \beta_{ik})|k \in \{1, .., K\}\}$ using $F(x_i; W)$
        Compute similar instances N_i using F_i (eq 1)
        $z_i \leftarrow \underset{k}{\mathrm{argmax}}\,\beta_{ik}$
    end for

POSE REFINEMENT
    $\forall i$, Update z_i (eq 6) until convergence

3.1. Approach

We first obtain multiple pose hypotheses for each instance by obtaining a diverse set of modes from the distribution predicted by the system described in Section 2 . We then frame the joint pose prediction task as that of selecting a hypothesis for each instance while taking into consideration the prediction confidence score as well as pose consistency with similar instances. We describe our formulation in detail below.

Instance Similarity. For each instance i, we obtain a set of instances N_i whose feature representations are similar to instance i. Our feature representation for an instance is motivated by the observation that each channel in a higher-layer of a CNN can be reasoned as encoding a spatial likelihood of abstract parts. Let $C_i(x, y, k)$ denote the instance's convolutional feature response for channel k at location (x, y), our feature representation F_i is as follows.

$$F_i(\cdot, \cdot, k) = \frac{\sigma(C_i(\cdot, \cdot, k))}{\|\sigma(C_i(\cdot, \cdot, k))\|_1} \qquad (1)$$

The above, where $\sigma(\cdot)$ represents a sigmoid function, encodes each instance via the normalized spatial likelihood of
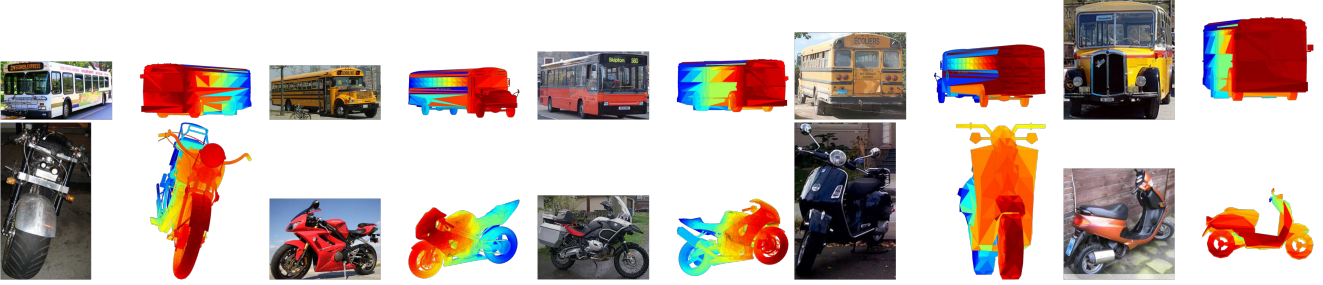
Figure 3. Viewpoint predictions for unoccluded groundtruth instances using our full system ('GC+sim'). The columns show 15th, 30th, 45th, 60th and 75th percentile instances respectively in terms of the error. We visualize the predictions by rendering a 3D model using our predicted viewpoint.

these 'parts'. We use histogram intersection over these representations as a similarity measure between two instances and obtain the set of neighbors N_i for each instance.

Unaries. For each instance i, we obtain K distinct pose hypotheses $\{R_{ik}|k \in \{1,..,K\}\}$ along with the corresponding log-likelihood scores β_{ik}. By $Z_i \in \{1,..,K\}$, we denote the random variable which corresponds to the pose hypothesis we select for instance i. The log-likelihood scores for each pose hypothesis act as the unary likelihood terms.

$$P_u(Z_i = z_i) \propto e^{\beta_{iz_i}} \qquad (2)$$

Pose Consistency. Let $\Delta(R_1, R_2) = \frac{\|log(R_1^T R_2)\|_F}{\sqrt{2}}$ denote the geodesic distance between rotation matrices R_1, R_2 and $\mathcal{I}$ denote the indicator function. We model the consistency likelihood term as the fraction of instances in N_i with a similar pose.

$$P_c(Z_i = z_i) \propto \frac{\sum\limits_{j \in N_i} \mathcal{I}(\Delta(R_{iz_i}, R_{jz_j}) < \delta)}{|N_i|} \qquad (3)$$

While this formulation encourages similar pose estimates for neighbors, it is biased towards more 'popular' pose estimates (if the dataset has more front facing bikes, it is more likely to find neighbors for the corresponding pose hypothesis). Motivated by the recent work of Isola *et al*. [20], who use Pointwise Mutual Information [12] (PMI) to counter similar biases, we normalize by the likelihood of randomly finding similar pose estimates for neighbors to yield -

$$P_c(Z_i = z_i) \propto \frac{\sum\limits_{j \in N_i} \mathcal{I}(\Delta(R_{iz_i}, R_{jz_j}) < \delta)}{\sum\limits_{j} \mathcal{I}(\Delta(R_{iz_i}, R_{jz_j}) < \delta)} \qquad (4)$$

Formulation. P_u favors the pose hypotheses that are scored higher by the instance pose induction system and P_c, weighted by a factor of λ, leads to a higher joint probability if predicted pose in consistent with pose for similar

instances. We finally combine these two likelihood terms to model the likelihood for the pose hypotheses for a given instance.

$$P(Z_i = z_i) \propto P_u(z_i)P_c(z_i)^{\lambda} \qquad (5)$$

Inference. We aim to infer the MAP estimates z_i^* for all instances to give us a pose prediction via joint reasoning over all instances. We use iterative updates and at each step, we condition on all the unknown variables except a particular Z_i; the update for assignment z_i as follows -

$$z_i = \operatorname*{argmax}_k (\beta_{ik} + \lambda log(\sum\limits_{j \in N(i)} \mathcal{I}(\Delta(R_{ik}, R_{jz_j}) < \delta)) - \lambda log(\sum\limits_{j} \mathcal{I}(\Delta(R_{ik}, R_{jz_j}) < \delta))) \quad (6)$$

Our overall method, as summarized in Algorithm 1, computes pose estimates for every instance of a novel object class given a large collection of instances.

3.2. Experiments

The aim of the experiments is twofold - 1) to demonstrate the benefits of jointly reasoning over all instances of the class and 2) to show that a spatial feature representation capturing abstract parts, as defined in eq 1, yields better performance than alternatives for improving pose estimates. We follow the experimental setup previously described in Section 2.3 and build on the 'GC' approach. Our method using spatial features (from Conv5 of VGG net) is denoted as 'GC+c5' and the alternate similarity representation using fc7 features from VGG net is denoted as 'GC+fc7'. We visualize the performance of our system in Figure 3 where the columns show $15^{th} - 75^{th}$ percentile instances, when sorted in terms of error. We observe that the predictions are accurate even around the $60^{th} - 75^{th}$ percentile regime.

We see that the results in Table 2 clearly support our two main hypotheses - that given multiple instances of a novel category, jointly reasoning over all of them improves the
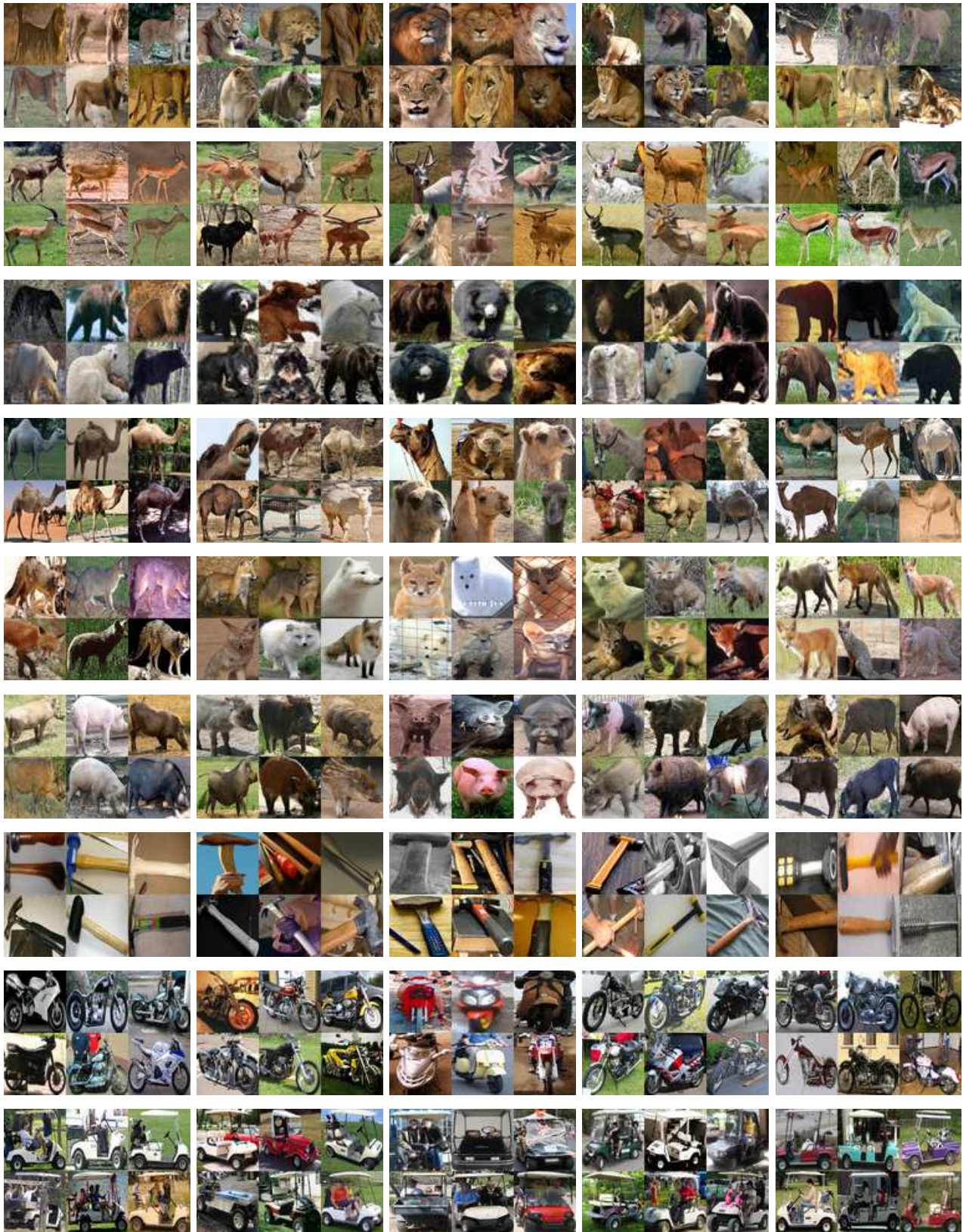
Figure 4. Viewpoint predictions for novel object classes without any pose annotations. The columns show randomly selected instances whose azimuth is predicted to be around $\frac{-\pi}{2}$ (right-facing), $\frac{-\pi}{4}$, 0(front-facing), $\frac{\pi}{4}$, $\frac{\pi}{2}$ (left-facing) respectively.
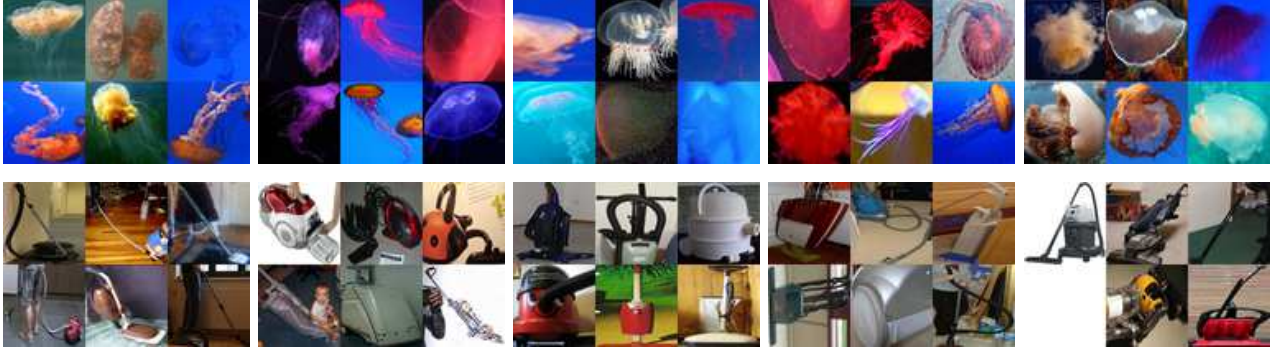
Figure 5. Failure modes. Our method is unable to induce pose for object classes which drastically differ from the annotated seed set. The columns show randomly selected instances whose azimuth is predicted to be around $\frac{-\pi}{2}$(right-facing), $\frac{-\pi}{4}$, 0(front-facing), $\frac{\pi}{4}$, $\frac{\pi}{2}$(left-facing) respectively.

	$Acc_{\frac{\pi}{6}}$		Acc_v	
Approach	bus	mbike	dog	sheep
GC	0.80	0.55	**0.74**	0.77
GC+fc7	0.76	0.51	0.73	0.75
GC+c5	**0.86**	**0.60**	**0.74**	**0.79**

Table 2. Joint reasoning for Pose Induction.

	$Acc_{\frac{\pi}{6}}$		Acc_v	
Setting	bus	mbike	dog	sheep
All	0.86	0.60	0.74	0.79
Confident	0.97	0.76	0.89	0.90

Table 3. Performance for Confident Predictions.

induced pose estimates and that the feature representation we described further improves performance.

An additional result that we show in Table 3 is that if we rank the predictions by confidence (eq. 5) and take the top third confident predictions, error rates are significantly reduced. This means that the pose induction system has the desirable property of having low confidence when it fails. As we demonstrate later, for various applications *e.g.* shape model learning, we might only need accurate pose estimates for a subset of instances and this result allows us to automatically find that subset by selecting the top few in terms of confidence.

3.3. Qualitative Results

The evaluation setup so far has focused on PASCAL VOC object classes because of readily available annotations to measure performance. However, the aim of our method is to be able to infer pose annotations for any novel object category. We can qualitatively demonstrate the applicability of our approach for diverse classes using the Imagenet ob-

ject classes. Figure 4 shows the predictions of our method for several classes for which we do not use any pose annotations (we use randomly selected instances from the top third, in terms of prediction confidence, to visualize the predictions in Figure 4). It is clear that the system performs well on animals in general as well as for other classes related to the initial training set (eg. golfcart, motorbike). While we can often infer a meaningful representation of pose even for some classes rather different from the initial training classes e.g. hammer, object categories which differ drastically from the annotated seed set (eg. jellyfish, vacuum cleaner) are the principal failure modes as illustrated in Figure 5.

4. Shape Modelling for Novel Object Classes

Acquiring shape models for generic object categories is an integral component of perceiving scenes with a rich 3D representation. The conventional approach to acquiring shape models includes leveraging human experts to build 3D CAD models of various shapes. This approach, however, cannot scale to a large number of classes while capturing the wildly different shapes in each object class. Learning based approaches which also allow shape deformations [2] provide an alternative solution but typically rely on some 3D initialization [6]. Kar *et al.* [21] recently showed that these models can be learned using annotations for only object silhouettes and a set of keypoints. These requirements, while an improvement over previous approaches, are still prohibitive for deploying similar approaches on a large scale. Enabling such approaches to learn shape models in the wild - given nothing but a set of instances, is an important endeavor as it would allow us to scale shape model acquisition to a large set of objects.

We take a step towards this goal using our pose induction system - we demonstrate that it is possible to learn shape models for a novel object category using just object silhouette annotations. We build on the formulation by Kar *et al.*
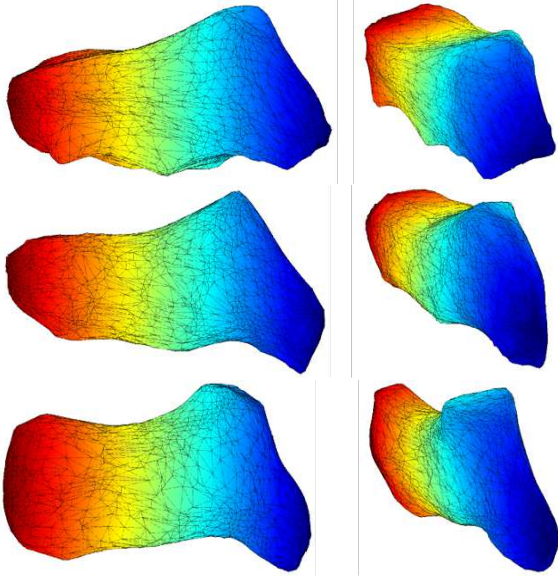
Figure 6. Mean shape models learnt for motorbike using a) top : all pose induction estimates b) mid : most confident pose induction estimates c) bottom : ground-truth keypoint annotations.

[21] and note that they mainly used keypoint annotations to estimate camera projection parameters and that these can be initialized using our induced pose as well. We briefly review their formulation and describe our modifications that allow us to learn shape models without keypoint annotations.

Formulation. Let $P_i = (R_i, c_i, t_i)$ represent the projection parameters (rotation, scale and translation) for the ith instance. Kar *et al.* obtain these using the annotated keypoints and we instead initialize the scale, translation parameters using bounding box scale, location and the rotation using our induced pose. Their shape model $M = (\overline{S}, V)$ consists of a mean shape $\overline{S}$ and linear deformation bases $V = \{V_1, ., V_K\}$. The energies used in their formulation enforce that the shape for an instance is consistent with its silhouette (E_s, E_c), shapes are locally consistent (E_l), normals vary smoothly (E_n) and the deformation parameters are small ($\|\alpha_{ik} V_k\|_F^2$) (they also use a keypoint based energy E_{kp} which we ignore). We refer the reader to [21] for details regarding the optimization and formulations of shape energies. While Kar *et al.* only optimize over shape model and deformation parameters, we note that since our projection parameters are noisy, we should also refine them to minimize the energy. Therefore, we minimize the objective mentioned in eq. 7 over the shape model, deformation parameters as well as projection parameters (initialized using the induced pose) to learn shape models of a novel object class using just silhouette annotations.

$$\min_{\overline{S}, V, \alpha, P} E_l(\overline{S}, V) + \sum_i (E_s^i + E_c^i + E_n^i + \sum_k (\|\alpha_{ik} V_k\|_F^2))$$

$$\text{subject to: } S^i = \overline{S} + \sum_k \alpha_{ik} V_k$$

$$(7)$$

Results. We use the unoccluded instances of the class motorbike to demonstrate the applicability of our pose induction system for shape learning. Since we are interested in learning a shape model for the class, we can ignore some object instances for which we are uncertain regarding pose. As shown in table 3, we can use the subset of most confident pose estimates to get a higher level of precision. Figure 6 shows that our model learnt without any keypoint annotation is quite similar to the model learnt by Kar *et al.* using full annotations and that using the subset of instances with confident pose induction predictions substantially improves shape models. The learnt model demonstrates that our pose induction system makes it is feasible to learn shape models for novel object classes without requiring keypoint annotations. This not only qualitatively verifies the reliability of our pose induction estimates, it also signifies an important step towards automatically learning shape representations from images.

5. Conclusion

We have presented a system which leverages available pose annotations for a small set of seed classes and can induce pose for a novel object class. We have empirically shown that the system performs well given a single instance of a novel class and that this performance is significantly improved if we reason jointly over multiple instances of that class, when available. We have also shown that our pose induction system enables learning shape representations for object classes without any keypoint/3D annotations required by previous methods. Our qualitative results on Imagenet further demonstrate that this approach generalizes to a large and diverse set of object classes.

References

[1] E. Bart and S. Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In *CVPR*, 2005. 2

[2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 7

[3] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010. 3

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

[5] J. Carreira, A. Kar, S. Tulsiani, and J. Malik. Virtual view networks for object reconstruction. In *CVPR*, 2015. 2

[6] T. Cashman and A. Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. *PAMI*, 2013. 1, 7

[7] X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting Visual Knowledge from Web Data. In *ICCV*, 2013. 4

[8] M. Cimpoi, S. Maji, and A. Vedaldi. Deep filter banks for texture recognition and segmentation. In *CVPR*, 2015. 1

[9] 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. 3

[10] I. Endres, V. Srikumar, M.-W. Chang, and D. Hoiem. Learning shared body plans. In *CVPR*, 2012. 1

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 3

[12] R. Fano. *Transmission of Information: A Statistical Theory of Communications*. The MIT Press, 1961. 5

[13] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 2

[14] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 2006. 2

[15] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 1980. 3

[16] A. Ghodrati, M. Pedersoli, and T. Tuytelaars. Is 2d information enough for viewpoint estimation? In *BMVC*, 2014. 2

[17] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 1

[18] J. Hoffman, S. Guadarrama, E. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. LSDA: Large scale detection through adaptation. In *NIPS*, 2014. 2

[19] D. P. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *IJCV*, 1990. 2

[20] P. Isola, D. Zoran, D. Krishnan, and E. H. Adelson. Crisp boundary detection using pointwise mutual information. In *ECCV*, 2014. 5

[21] A. Kar, S. Tulsiani, J. Carreira, and J. Malik. Category-specific object reconstruction from a single image. In *CVPR*, 2015. 1, 2, 7, 8

[22] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2

[23] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to hand-written zip code recognition. In *Neural Computation*, 1989. 3

[24] V. Lepetit, J. Pilet, and P. Fua. Point matching as a classification problem for fast and robust object pose estimation. In *CVPR*, 2004. 2

[25] M. Osadchy, Y. L. Cun, and M. L. Miller. Synergistic face detection and pose estimation with energy-based models. *JMRL*, 2007. 2

[26] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009. 2

[27] S. E. Palmer. *Vision science: Photons to phenomenology*. MIT press Cambridge, MA, 1999. 2

[28] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In *CVPR*, 2012. 2

[29] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 2013. 2

[30] D. A. Simon, M. Hebert, and T. Kanade. Real-time 3-d pose estimation using a high-speed range sensor. In *IEEE International Conference on Robotics and Automation*, 1994. 2

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3

[32] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 4

[33] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 2

[34] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 2011. 2

[35] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. *PAMI*, 2007. 1, 2

[36] S. Tulsiani and J. Malik. Viewpoints and keypoints. In *CVPR*, 2015. 2, 3

[37] M. Varma and A. Zisserman. A statistical approach to material classification using image patch exemplars. *PAMI*, 2009. 1

[38] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc. In *CVPR*, 2014. 1, 2, 3

[39] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014. 3

[40] T. Zhou, Y. J. Lee, S. X. Yu, and A. A. Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *CVPR*, 2015. 4