Single Image Pop-Up from Discriminatively Learned Parts

Menglong Zhu* Xiaowei Zhou* Kostas Daniilidis
Computer and Information Science, University of Pennsylvania
{menglong,xiaowz,kostas}@cis.upenn.edu

Abstract

We introduce a new approach for estimating a fine grained 3D shape and continuous pose of an object from a single image. Given a training set of view exemplars, we learn and select appearance-based discriminative parts which are mapped onto the 3D model through a facility location optimization. The training set of 3D models is summarized into a set of basis shapes from which we can generalize by linear combination. Given a test image, we detect hypotheses for each part. The main challenge is to select from these hypotheses and compute the 3D pose and shape coefficients at the same time. To achieve this, we optimize a function that considers simultaneously the appearance matching of the parts as well as the geometric reprojection error. We apply the alternating direction method of multipliers (ADMM) to minimize the resulting convex function. Our main and novel contribution is the simultaneous solution for part localization and detailed 3D geometry estimation by maximizing both appearance and geometric compatibility with convex relaxation.

1. Introduction

Recovering 3D geometry from 2D imaginary of an object is one of the most fundamental and challenging problems in computer vision. Geometric features were the main representation of objects in the 20th century and have long been used to establish correspondence between vertices and edges of a 3D model and their image projections [14]. Although such representation was successful with geometric invariance it could not cope with the complexity of appearance of 3D object categories in the real world which could only be learned from exemplars.

As soon as massive 2D image exemplars became available on the Internet and through tedious annotation, the computer vision community has harnessed fruitful results as the state of the art in detecting object categories has improved dramatically [8, 11]. More recently, researchers have focused on combining such approaches with 3D geometry to build more powerful object detectors that are also able to provide weak 3D information such as viewpoint [29, 30, 40, 25, 19]. In this paper, we go beyond viewpoint estimation to establishing the actual 3D shape of an object for the sake of fine grained classification or 3D interaction such as grasping and manipulation. Very few efforts have been devoted to such combined estimation of pose and shape from a single image [18, 27].

Recent advances in recognition have opened doors to better understanding of 3D in the wild, but there are three main challenges in the marriage of 2D appearance and 3D geometry: (1) how to learn a representation that captures appearance variation of geometric features across instances and poses, (2) how to establish the 3D shape of an object without exhaustively comparing to all possible instances or when that instance has not been seen before, and (3) how to optimize for appearance and correspondence compatibility as well as 3D shape and pose at the same time, without splitting the problem into subproblems.

In this paper, we propose a novel approach that marries the power of discriminative parts with an explicit 3D geometric representation with the goal to infer 3D shape and continuous pose of an object (or *pop-up*) from a single image. Part descriptors are discriminatively learned in training images. Such parts are centered around projections of 3D landmarks which are given in abundance on the training 3D models. To establish a compact representation we minimize the number of needed landmarks by solving a facility-location problem. To deal with geometric deformation, we summarize the training set of 3D models into a shape dictionary from which we can generalize by linear combination. Given a test image we detect top location hypotheses of each part. The challenge is how to fit best these parts by maximizing the geometric consistency. This entails the selection among the hypotheses of each part and the shape/pose computation. Unlike other approaches which rely on local optimization and initialize pose by DPM-based discretized pose estimation [40, 27], we compute the selection as well as the shape and pose parameters in one step using a convex program solved with the alternating direc-

*These authors contributed equally to this work

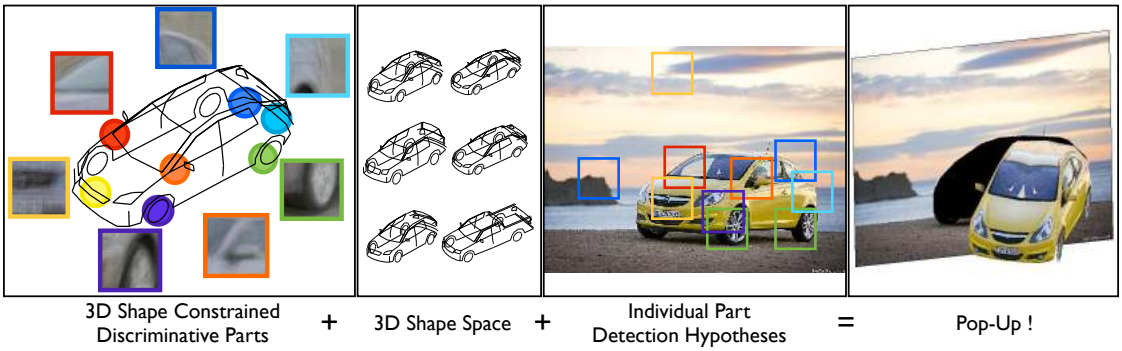

Figure 1: Illustrative summary of our approach: 3D Landmarks on a 3D model are associated with discriminatively learned part descriptors (left). Intra-class shape variation is captured with linear combinations of a sparse shape basis (2nd left). Learned part descriptors produce multiple maximum responses for each part in a testing image (3rd from left). The selection of the part hypotheses, 3D pose and 3D shape are simultaneously estimated and the result is illustrated through a popup (right).

tion method of multipliers (ADMM).

Figure 1 illustrates the outline of our approach. In summary, the major technical contributions are:

- A convex optimization framework for joint landmark localization, fine grained 3D shape and continuous pose estimation from a single image.
- Our convex objective does not require viewpoint or detection initialization.
- An automatic landmark selection method considering both discriminative power in appearance and spatial coverage in geometry.

2. Related Work

The most related work includes the family of methods that estimate an object shape by aligning a deformable shape model to image features. This idea originated from the active shape model (ASM) [4], which was originally proposed for segmentation and tracking based on low-level image features. Cristinacce and Cootes [5] proposed the constrained local models (CLM), which combined ASM with local appearance models for 2D feature localization in face images. Gu and Kanade [16] presented a method to align 3D deformable models to 2D images for 3D face alignment. The similar methods were also proposed for 3D car modeling [18, 40, 19, 27] and human pose estimation [31, 38]. Our method differs in that we use a data-driven approach for discriminative landmark selection and we solve landmark localization and shape reconstruction in a single convex framework, which enables a global solution.

The representation of our model is inspired by recent advances in part-based modeling [8, 33, 17, 22], which models the appearance of object classes with a collection of mid-sized discriminative parts. Our optimization approach is related to the previous work on using convex relaxation techniques for object matching [28, 21, 24]. These methods focused on finding the point-to-point correspondence between an object template and an image in 2D, while our method considers 3D to 2D matching as well as shape variability.

Our paper is also related to recent work on 3D pose estimation which encodes the geometric relations among local parts and achieved continuous pose estimation. Several work leveraged 3D models to warp features or parts into their canonical view [32, 37, 36]. Other work rendered local appearances and depth from 3D models and subsequently encoded in a 3D voting scheme [34, 13, 25]. DPM was further lifted to 3D deformable models [9, 29] to predict continuous viewpoint. Instance models were also used to recover 3D pose of an object [26, 1]. But this line of work focused on pose estimation and either used generic class models or instance-based models. Our approach differs in that we not only provide a detailed shape representation but also consider intra-class variability.

3. Shape Constrained Discriminative Parts

Our proposed method models both 2D appearance variation and 3D shape deformation of an object class. The 2D appearance is modeled as a collection of discriminatively trained parts. Each part is associated with a 3D landmark point on a deformable 3D shape.

Unlike the previous works that manually define landmarks on the shape model, we propose an *automatic* selection scheme: we first learn the appearance models for all points on the 3D model, evaluate their detection performance, and select a subset of them as our part models based on their detection performance in 2D and the spatial coverage in 3D.

3.1. Learning Discriminative Parts

One of the main challenges in object pose estimation rises from the fact that due to perspective transform and self occlusions, even the same 3D position of an object has very different 2D appearances in the image observed from different viewpoints. We tackle this problem by learning a mixture of discriminative part models for each point in the 3D model to capture the variety in appearance. Each part detector consists of a simple but fast HOG detector [6] and a more sophiscated but slow deep classifiers trained with deep Convolutional Neural Net (CNN) [23]. The HOG detectors provide location proposals to deep classifiers. Such design is chosen to balance speed versus accuracy.

Given a training set D, each training image $I_i \in D$ is associated with the 3D points of the object shape $S \in \mathbb{R}^{3\times p}$, their 2D projections $L_i \in \mathbb{R}^{2\times p}$ annotated in the image.

HOG Part Detectors We bootstrap the learning of a discriminative mixture model for each part via clustering whitened HOG (WHO) features [17, 33]. Denote $\phi(L_{ij})$ as the HOG feature of the positive image patch centered at L_{ij} and $\overline{\phi}_{\mathbf{bg}}$ as the mean of background HOG features. We compute the WHO feature as $\Sigma^{-1/2}(\phi(L_{ij}) - \overline{\phi}_{\mathbf{bg}})$, where Σ is the shared covariance matrix computed from all positive and negative features. Then we cluster the WHO features of each part j into m clusters using K-means.

A linear classifier W_{cj} is trained for each cluster c of a part j. We apply linear discriminant analysis due to efficiency in training and limited loss in detection accuracy [17, 12],

$$W_{cj} = \Sigma^{-1}\left(\overline{\phi}(L_{ij}; z_{ij} = c) - \overline{\phi}_{\mathbf{bg}}\right), \tag{1}$$

where $z_{ij} \in \{1, \ldots, m\}$ is the cluster assignment for each feature, and $\overline{\phi}(L_{ij}; z_{ij} = c)$ is the mean feature over all L_{ij} of cluster c. Let $\mathbf{x} = (x, y)$ be the position (x, y) in the image. The response of part j at a given location $\mathbf{x}$ is the max response over all its c components: $score_j(\mathbf{x}) = \max_c\{W_{cj} \cdot \phi(\mathbf{x})\}$.

We introduce a latent variable for each training patch, $r_{ij} \in \mathbb{R}^2$ to represent the relative center location to the annotated landmark location L_{ij}. We improve the classifiers learned from (1) by repositioning the patch center in the neighborhood $\Delta(L_{ij})$ of L_{ij} and retrain the classifiers. Note that the latent update procedure is similar to that of DPM [8] with the difference that we do not apply generalized distance transform to filter responses but only consider maximum responses within a local region. The reason is that our model, as will be discussed in Section 3.3, is constrained by the 3D shape space instead of learned 2D deformations. We want, thus, to obtain accurate part localization to estimate the object pose and shape. A 2×2 covariance matrix D_j is estimated for each landmark j from latent variables r_{ij}, to model the uncertainty of the detected landmark position $\mathbf{x}^*_{ij}$ relative to the ground truth.

Method	HOG-SVM	CNN
mAP	0.41	**0.53**

Table 1: Comparison of CNN and HOG-SVM in part localization. Mean average precision (mAP) of localizing the 12 parts of PASCAL3D car category are shown.

Deep Part Classifiers HOG part detections serve as part proposals and are subsequently re-ranked by forwarding through a CNN and applying SVM on the extracted Pool5 layer features. During training, Pool5 features were extracted for both positive and negative patches and an SVM is trained for each part mixture. During our experiments, we observed that 1) fine-tuning from pre-trained AlexNet [23] with part patches of the same object category improves part detection accuracy, 2) Pool5 has better performance than fully connected layers (fc6, fc7) for mid-level patches, 3) training separate classifiers for each part mixture component outperform a combined classifier. We used publicly available deep learning toolbox Caffe [20] in our experiments.

The performance of deep part classifiers is evaluated by comparing against SVM trained HOG filters (HOG-SVM) with hard negative mining. Localization accuracy is measured by the average precision of detecting the part within the close vicinity of the groundtruth location. Table 1 shows performance comparison of CNN and HOG-SVM on the 12 parts of PASCAL3D dataset car category.

3.2. Selecting Discriminative Landmarks

Seeking a compact representation of the object, we try to select only a small subset of discriminative landmarks S_D among all 3D landmarks S. We want the selected landmarks S_D to be both associated with discriminative part models and have a good spatial coverage of the object shape model in 3D. The selection problem is formulated as a **facility location problem**,

$$\begin{aligned} \min_{y_u, x_{uv}} \quad & \sum_u z_u y_u + \lambda \sum_{uv} d_{uv} x_{uv}, \\ \text{s.t.} \quad & \sum_v x_{uv} = 1, \\ & x_{uv} \le y_v, \qquad \forall u, v, \\ & x_{uv}, y_u \in \{0, 1\}, \quad \forall u, v, \end{aligned} \tag{2}$$

where the interpretations of each symbol are presented in Table 2.

The cost z_u for a landmark u should be lower if the associated part model is more discriminative. We model

Symbol	Interpretation
z_u	cost of selecting landmark u
y_u	binary landmark selection variable
d_{uv}	cost of landmark v "serving" u
x_{uv}	binary variable for landmark v "serving" u
λ	trade off between unary costs and binary costs

Table 2: Notations interpretation in (2)

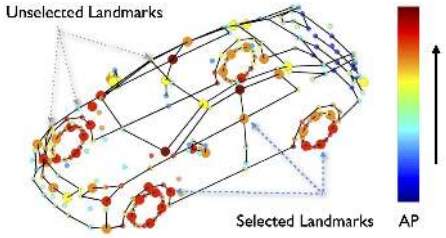

Figure 2: Visualization of the landmark selection optimization result. All 256 landmark points of a car are shown in circle markers. The color of the markers represents the Average Precision(AP) of the landmark part detection on the training set, red means higher AP and blue means lower AP. The size of the landmark represents the selection result, the larger ones are selected via the MIP optimization and the smaller ones are not selected. The red landmarks are preferred since they have higher detection accuracy, but only a subset of red landmarks are selected because they are close in 3D.

the discriminativeness by evaluating the Average Precision (AP) of detecting each landmark in the training set. For any landmark u, we perform detection with the learned part model in the training set S to generate a list of location hypotheses H_u. A hypothesis $h \in H_u$ is considered as true positive if the ground truth location L_{iu} is within a small radius δ. Let the computed AP for a part u be AP_u, we set $z_u = 1 - AP_u$. The costs of "serving" (or suppressing) other landmarks are set to be the euclidean distance between landmarks in 3D, i.e., $d_{uv} = ||S_u - S_v||_2$. The value of λ is set to 1 in our experiments. The minimization problem 2 is a Mixed Integer Programming (MIP) problem, which is known to be NP-hard. But a good approximation solution can be obtained by relaxing the integer constrains to be $x_{uv} \in [0, 1]$, $y_u \in [0, 1]$, solving the relaxed Linear Programming problem, and thresholding the solution. Figure 2 visualizes an example result of MIP optimization for landmark selection.

3.3. 3D Shape Model

We start our description by explaining how we would estimate the shape of an object if 2D part - 3D landmark correspondences were known. We represent a 3D object model as a linear combination of a few basis shapes to constrain the shape variability. This assumption has been widely used in various shape-related problems such as object segmentation [4], nonrigid structure from motion [3] and single image-based shape recovery [16, 40]. We use a weak-perspective model, which is a good approximation when the depth of the object is smaller than the distance from the camera. With these two assumptions, the 2D shape $P \in \mathbb{R}^{2\times p}$ can be described by

$$P = R\sum_{i=1}^{k} c_i B_i + \mathbf{t}\mathbf{1}^T, \tag{3}$$

where $B_i \in \mathbb{R}^{3\times p}$ denotes the i-th basis shape, $R \in \mathbb{R}^{2\times 3}$ represents the first two rows of camera rotation, and $\mathbf{t} \in \mathbb{R}^2$ is the translation vector. In model inference, the reprojection error is minimized to find the optimal parameters.

However, the model in (3) is bilinear in R and c_is yielding a nonconvex problem. In order to have a linear representation, we use the method proposed in [39], which assumes that the unknown shape is a linear combination of scalable and rotatable basis shapes:

$$P = \sum_{i=1}^{k} T_i B_i + \mathbf{t}\mathbf{1}^T, \tag{4}$$

where $T_i \in \mathbb{R}^{2\times 3}$ denotes the first two rows of a similarity transformation matrix. In order to enforce T_i to be orthogonal, the spectral norms of T_is are minimized during model inference. The spectral norm is the largest singular value of a matrix, and minimizing it enforces the two singular values of T_i to be equal, which yields an orthogonal matrix [39]. After T_is are estimated, the third rows of T_is can be recovered from the orthogonality and then the estimated 2D shape can be lifted to 3D.

4. Model Inference

Finally, we obtain global geometry-constrained local-part models, in which the unknowns are the 2D part locations as well as the 3D pose and shape. In model inference, we maximize the detector responses over the part locations while minimizing the geometric reprojection error.

4.1. Objective Function

We try to locate a part by finding its correspondence in a set of hypotheses given by the trained detector. The cost without geometric constraints is

$$f_{score}(\mathbf{x}_1, \cdots, \mathbf{x}_p) = -\sum_{j=1}^{p} \mathbf{r}_j^T \mathbf{x}_j, \tag{5}$$

where $\mathbf{x}_j \in \{0, 1\}^l$ is the selection vector and $\mathbf{r}_j \in \mathbb{R}^l$ is the vector of the detection scores for all hypotheses for the j-th part.

Geometric consistency is imposed by minimizing the following reprojection error:

$$f_{geom}(\mathbf{x}_1, \cdots, \mathbf{x}_p, T_1, \cdots, T_k, \mathbf{t}) = \frac{1}{2} \sum_{j=1}^{p} \left\| D_j^{-\frac{1}{2}} \left(L_j^T \mathbf{x}_j - \left[\sum_{i=1}^{k} T_i B_i \right]_j - \mathbf{t} \right) \right\|^2, \quad (6)$$

where we concatenate the 2D locations of hypotheses for part j in $L_j \in \mathbb{R}^{l \times 2}$ and denote the covariance estimated in training as D_j.

As introduced in Section 3.3, we add the following regularizer to enforce the orthogonality of T_i:

$$f_{reg}(T_1, \cdots, T_k) = \sum_{i=1}^{k} \|T_i\|_2, \quad (7)$$

where we use $\|T_i\|_2$ to represent the spectral norm of T_i, i.e., the largest singular value.

To simplify the computation, we relax the binary constraint on $\mathbf{x}_i$ and allow it to be a soft-assignment vector $\mathbf{x}_i \in \mathcal{A}$, where $\mathcal{A} = \{\mathbf{x} \in [0,1]^l | \ \sum_{i=1}^{l} x_i = 1.\}$.

Finally, the objective function reads

$$\min_{\overline{X}, \overline{T}, \mathbf{t}} \ f_{geom}(\overline{X}, \overline{T}, \mathbf{t}) + \lambda_1 f_{score}(\overline{X}) + \lambda_2 f_{reg}(\overline{T}), \quad (8)$$
$$\text{s.t.} \quad \mathbf{x}_j \in \mathcal{A}, \ \forall j = 1:p,$$

where $\overline{X}$ and $\overline{T}$ represent the unions of $\mathbf{x}_1, \cdots, \mathbf{x}_p$ and $T_1, \cdots, T_k$, respectively. After solving (8), we recover the 3D shape S and pose $\theta = (R, \mathbf{t})$ from T_is, as introduced in Section 3.3.

4.2. Optimization

The problem in (8) is convex since f_{score} is a linear term, f_{geom} is the sum of squares of linear terms, and f_{reg} is the sum of norms of unknown variables. We use the alternating direction method of multipliers (ADMM) [2] to solve the convex problem in (8). Since f_{reg} is nondifferentiable, which is not straightforward to optimize, we introduce an auxiliary variable Z and reformulate the problem as follows:

$$\min_{\overline{X}, \overline{T}, \mathbf{t}, \overline{Z}} \ f_{geom}(\overline{X}, \overline{T}, \mathbf{t}) + \lambda_1 f_{score}(\overline{X}) + \lambda_2 f_{reg}(\overline{Z}),$$
$$\text{s.t.} \quad \overline{T} = \overline{Z}, \quad \mathbf{x}_j \in \mathcal{A}, \ \forall j = 1:p. \quad (9)$$

The corresponding augmented Lagrangian is:

$$\mathcal{L} = f_{geom}(\overline{X}, \overline{T}, \mathbf{t}) + \lambda_1 f_{score}(\overline{X}) + \lambda_2 f_{reg}(\overline{Z}) + \langle Y, \overline{T} - \overline{Z} \rangle + \frac{\rho}{2} \|\overline{T} - \overline{Z}\|_F^2. \quad (10)$$

The ADMM algorithm iteratively updates variables by the following steps to find the stationary point of (10):

$$\mathbf{t} \leftarrow \arg\min_{\mathbf{t}} \mathcal{L}, \quad (11)$$
$$\overline{X} \leftarrow \arg\min_{\overline{X}} \mathcal{L}, \quad (12)$$
$$\overline{T} \leftarrow \arg\min_{\overline{T}} \mathcal{L}, \quad (13)$$
$$\overline{Z} \leftarrow \arg\min_{\overline{Z}} \mathcal{L}, \quad (14)$$
$$Y \leftarrow \rho(\overline{T} - \overline{Z}). \quad (15)$$

It can be shown that (11), (12) and (13) are all quadratical programming problems, which have closed-form solutions or can be solved efficiently using existing convex solvers. (14) is a spectral-norm regularized proximal problem, which also admits a closed-form solution [39].

4.3. Visibility Estimation

In model inference, only visible landmarks should be considered. To estimate the unknown visibility, we adopt the following strategy. We first assume that all landmarks are visible and solve our model in (8) to obtain a rough estimate of the viewpoint. Since the landmark visibility of a car only depends on the aspect graph, the roughly estimated viewpoint can give us a good estimate of the landmark visibility. We observed that our model could reliably estimate the coarse view by assuming the full visibility, which might be attributed to the global optimization. After obtaining the visibility, we solve our model again by only considering the visible landmarks. The full shape can be reconstructed by the linear combination of full meshes of basis shapes after the coefficients are estimated.

4.4. Successive Refinement

The relaxation of binary selection vectors $\mathbf{x}_j$s in (8) may yield inaccurate localization, since it allows the landmark to be located inside the convex hull of the hypotheses. To improve the precision, we apply the following scheme: we solve our model in (8) repeatedly, and in each iteration we define a trust region based on the previous result for each landmark and merely keep the hypotheses inside the trust region as the input to fit the model again. We use three iterations. We can start from a large trust region to achieve global fitting and gradually decrease the trust region size in each iteration to reject outliers and improve localization. This successive refinement scheme has been widely-used for feature matching [24, 21].

5. Experiments

In this section, we evaluate our method (PopUp) in terms of both shape and pose estimation accuracy. The experiments are carried out on the Fine Grained 3D Car dataset (FG3DCar) [27] and PASCAL3D [35]. Both datasets have landmark locations in the image and pose annotation for 3D objects.

Method		meanAPD (SL)	meanAPD
PopUp	Mean shape	16.5	20.6
PopUp	Class mean	15.4	18.9
PopUp	Shape space	**14.6**	**17.7**
FG3D	Class mean	-	18.1
FG3D	Shape space	-	20.3

Table 3: Model fitting error of PopUp versus FG3D in terms of mean APD in pixels evaluated on 52 selected landmarks (SL) and 64 landmarks provided in the dataset.

5.1. FG3D Car Dataset

FG3DCar dataset consists of 300 images with 30 different car models of 6 car types under different viewing angles. Each car instance is associated a shape model of 256 3D landmark points and their projected 2D locations annotated in the image as well as 3D pose annotation. We perform the following evaluations: First, we compare the accuracy of pose and shape estimation to the iterative model fitting method of [27] (FG3D) in terms of 2D landmark projection error. Second, we compare the coarse viewpoint estimation error to viewpoint-DPM (VDPM) [15, 35]. In addition, since our viewpoint estimation is continuous, we also show the angular errors comparing to the groundtruth annotation. Through out the experiments, we follow the same training-testing split as [27].

We learn a mixture of discriminative part models of three components for each of 256 landmark points as described in Section 3. The Average Precision (AP) of the landmark detection is evaluated on the training set. We count a detection as true positive only if the detected location is close to the annotated location. We optimize the landmark selection with unary cost as 1 - AP of each landmark and pairwise cost as the average pairwise 3D distance over all the 3D models. 52 out of 256 landmark points are selected with MIP optimization while FG3DCar provides 62 manually selected landmark. To build the shape models, we learned a dictionary consisting of 10 basis shapes from the 3D models provided in the FG3DCar dataset.

Note that, unlike FG3D, our method does not need an external object detector to initialize either the location and scale in the image or coarse landmark locations. We perform pose and shape estimation on the original image with background clutter.

3D Shape Estimation 3D Shape estimation accuracy is evaluated in terms of meanAPD which is the average landmark projection error in pixels over the landmarks and the test instances. In the following experiments, we investigate the effect of using different 3D shapes on the model fitting error. We compare three setups with different basis shapes: only the mean shape, class-mean shapes and the learned shape space (10 basis shapes). The middle column of Table 3 shows the fitting error on selected discriminative landmarks. The fitting error decreases when we use the shape space instead of the mean shape or the class mean, which validates the use of shape space to express intra-class shape variation.

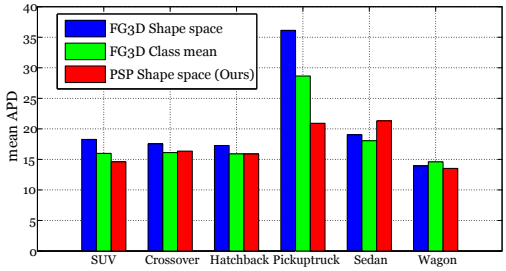

Figure 3: Car type specific meanAPD of PopUp versus FG3D with mean prior and class prior. Comparing to the FG3D method, our method achieves lower meanAPD on most car types. For the type of pickup truck, our method significantly outperforms FG3D.

	Accuracy	
Method	40° per view	20° per view
VDPM	82.7%	71.3%
PopUp	89.3%	84.7%

Table 4: Coarse viewpoint estimation accuracy versus VDPM evaluated on the FG3DCar dataset. Accuracies are compared with two discretization schemes, 20 degrees per coarse viewpoint and 40 degrees per coarse viewpoint.

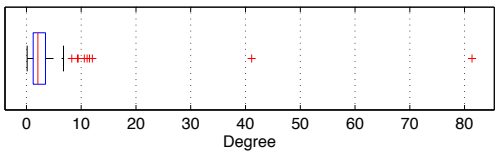

Figure 4: Continuous viewpoint (azimuth) error comparing to the groundtruth on all 150 test images in the FG3DCar dataset. The mean error is 3.4 in degrees.

Since the selected discriminative landmarks are not identical to the landmarks provided in the FG3DCar dataset, we also compare the meanAPD on the landmarks provided in the dataset. Our method outperforms FG3D using the shape space without knowing the class type. Note that, their detectors are trained on the manually selected 64 landmarks provided in the dataset while our detectors are trained on the 52 automatic selected discriminative landmarks.

Although our objective is to optimize the projection error on the discriminative landmarks, the fitting error on the dataset provided landmarks is also minimized. This shows

Method	Views	bicycle	bus	car	mbike
PopUp (ours)	4	42.6	49.3	29.8	39.9
	8	33.2	36.7	27.4	24.4
	16	16.9	40.7	21.4	16.6
	24	13.0	31.5	16.0	11.3
VDPM[35]	4	41.7	26.1	20.2	30.4
	8	36.5	35.5	23.5	25.1
	16	18.4	46.9	18.1	16.1
	24	14.3	39.2	13.7	10.1
Ghodrati et al.[10]	4	34.4	50.7	28.9	29.4
	8	27.6	50.3	26.6	24.7
	16	18.0	42.9	19.6	15.9
	24	12.6	40.2	15.9	13.2
DPM-3D[30]	4	43.9	50.7	36.9	31.8
	8	40.3	50.3	36.6	32.0
	16	22.9	42.9	29.6	16.7
	24	16.7	42.1	24.6	10.5

Table 5: Average (discrete) Viewpoint Accuracy on four categories of PASCAL3D dataset.

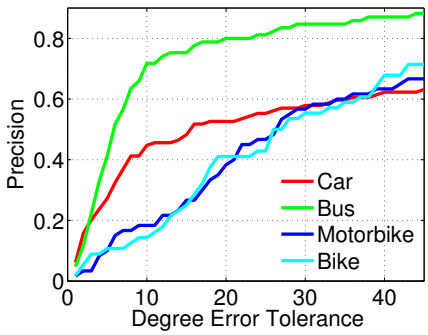

Figure 5: Precision recall curves for continuous viewpoint estimation on four categories of PASCAL3D, occluded instances are excluded. The horizontal axis is the tolerance of viewpoint error to count a prediction as correct, in the range of $[0, \pi/4]$. The vertical axis shows the precision.

the effectiveness of the landmark selection process. The error is reported on the same scale as FG3D. Figure 3 shows the per class 3D model fitting error. Our method outperforms FG3D on most class types with particular success on the pickup trucks.

Viewpoint Estimation We compare PopUp to VDPM in discrete viewpoint estimation accuracy. For VDPM we train two sets of baseline VDPM with coarse viewpoints (azimuth) of every 20 degrees and every 40 degrees for each view. Each component of VDPM corresponds to a viewpoint label. During inference, the viewpoint of the test car instance is predicted as the training viewpoint of the max scoring component. For PopUp, the estimated continuous viewpoint is discretized in the same way as VDPM. Table 4 shows the comparison of the two methods. In both two cases, PopUp outperforms VDPM. We further analyze the estimation error of PopUp by looking at continuous viewpoint estimation error and show that the majority error is introduced by discretization. We compare our estimation to the ground-truth viewpoint (azimuth) and report the absolute angular value in Figure 4. The mean error over the whole test set is only 3.4 in degree.

In addition to the quantitative evaluations, we show qualitative results on the test images from FG3DCar in Figure 6, where we project the 3D model wireframe with the estimated pose and shape on to the image. We also show the textured model rendered at novel views.

5.2. PASCAL3D Dataset

The PASCAL3D dataset augments a subset of PASCAL dataset [7] with 3D models and pose annotations. PASCAL3D consists of images captured under various natural conditions. Occlusions and various object sizes cast great challenges to 3D estimation. We validate our method on four categories of PASCAL3D (test set): bicycle, bus, car and motorbike. Both discrete and continuous viewpoint accuracies are evaluated. Part Detectors are trained on the PASCAL3D training set. We use the provided landmarks and 3D models.

For discrete viewpoint accuracy, we compare Average Viewpoint Accuracy to recently reported state-of-art results on the benchmark [35, 10, 30]. We use VDPM as base detector and estimate viewpoint within each detection hypothesis and quantize our continuous viewpoint output into discrete bins. Table 5 shows the accuracy of viewpoint prediction with different quantization of the azimuth angle, namely 4, 8, 16 and 24 views. Our results are comparable on different categories. While our model-based method performs well on larger objects, statistical learning based approaches as [30] have advantages on small and heavily occluded instances in terms of viewpoint prediction.

We evaluated the continuous viewpoint accuracy on non-occluded instances within groundtruth bounding boxes. Figure 5 shows the precision-recall curves for four categories as the viewpoint error tolerance changes within $[0, \pi/4]$. We can observe that for bus and car the precision increases quickly as the angular tolerance increases from 0 to 10 degrees, meaning that the majority angular errors are less than 10 degrees. Bus and car outperform bicycle and motorbike with our method because their landmarks have larger appearance variation. Figure 7 shows qualitative results with estimated visible landmarks reprojected.

We break down the running time of our system on a 3.3Ghz Intel i7 CPU and an Nvidia TitanZ GPU as the following. Estimating a single object instance in a PASCAL3D image (500x300 pixels) requires: 0.08 seconds building HOG pyramid; 1.41 seconds in filters convolution; 3.76 seconds in CNN classification and 1.52 seconds in ADMM.

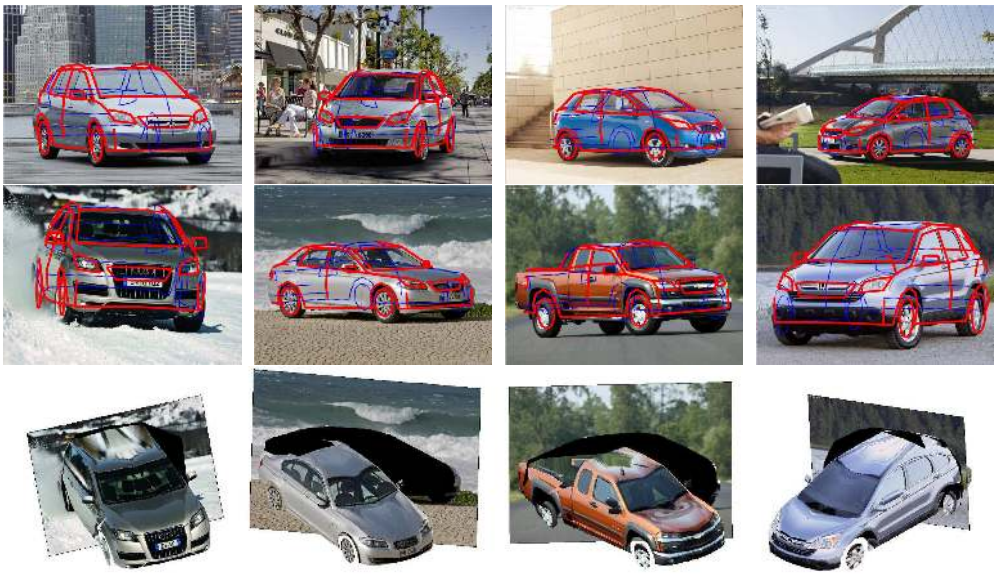

Figure 6: Example 3D estimation results from FG3D Car are shown. In the first two rows, the 3D wire frame of the car model is projected on the image with estimated pose and shape. Red solid lines represent visible wire frames and blue dotted lines represent invisible wire frames. In the last row, the textured 3D reconstructions of the cars in the second row are rendered at novel viewpoints. (We use symmetry to texture the invisible faces).

Figure 7: Examples of landmark localization results from different categories of PASCAL3D are shown in the first two rows. Visible 3D landmarks are projected back to the image. The yellow dots are groundtruth locations and the green dots are the estimation. The last row shows example pop-up results of different object classes.

6. Conclusion

We proposed a novel approach for estimating the pose and the shape of a 3D object from a single image. Our approach is based on a collection of automatically-selected and discriminatively-trained 2D parts with a 3D shape-space model to represent the geometric relation. In model inference, we simultaneously localized the parts, estimated the pose, and recovered the 3D shape by solving a convex program with ADMM.

Acknowledgement Financial support through the following grants: NSF-DGE-0966142, NSF-IIS-1317788, NSF-IIP-1439681, NSF-IIS-1426840, ARL RCTA W911NF-10-2-0016, ONR N000141310778 is gratefully acknowledged.

References

[1] M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014. 2

[2] S. Boyd. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010. 5

[3] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, 2000. 4

[4] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models – their training and application. *CVIU*, 61(1):38–59, 1995. 2, 4

[5] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, 2006. 2

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 3

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, June 2010. 7

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010. 1, 2, 3

[9] S. Fidler, S. Dickinson, and R. Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. In *NIPS*, 2012. 2

[10] A. Ghodrati, M. Pedersoli, and T. Tuytelaars. Is 2d information enough for viewpoint estimation. In *BMVC*, 2014. 7

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[12] R. Girshick and J. Malik. Training deformable part models with decorrelated features. In *ICCV*, 2013. 3

[13] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and pose estimation. In *ICCV*, 2011. 2

[14] W. Grimson. *Object recognition by computer: The role of geometric constraints*. The MIT Press, 1990. 1

[15] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, 2010. 6

[16] L. Gu and T. Kanade. 3D alignment of face in a single image. In *CVPR*, 2006. 2, 4

[17] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *ECCV*. 2012. 2, 3

[18] M. Hejrati and D. Ramanan. Analyzing 3d objects in cluttered images. In *NIPS*, 2012. 1, 2

[19] W. Hu and S. Zhu. Learning 3d object templates by quantizing geometry and appearance spaces. *PAMI*, 2014. 1, 2

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 3

[21] H. Jiang, S. X. Yu, and D. R. Martin. Linear scale and rotation invariant matching. *PAMI*, 33(7):1339–1355, 2011. 2, 5

[22] I. Kokkinos. Rapid deformable object detection using dual-tree branch-and-bound. In *NIPS*, 2011. 2

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3

[24] H. Li, J. Huang, S. Zhang, and X. Huang. Optimal object matching via convexification and composition. In *ICCV*, 2011. 2, 5

[25] J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3d feature maps. In *CVPR*, 2008. 1, 2

[26] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing ikea objects: Fine pose estimation. In *ICCV*, 2013. 2

[27] Y.-L. Lin, V. I. Morariu, W. Hsu, and L. S. Davis. Jointly optimizing 3d model fitting and fine-grained classification. In *ECCV*, 2014. 1, 2, 5, 6

[28] J. Maciel and J. P. Costeira. A global solution to sparse correspondence problems. *PAMI*, 25(2):187–199, 2003. 2

[29] B. Pepik, P. Gehler, M. Stark, and B. Schiele. 3d2pm–3d deformable part models. In *ECCV*, 2012. 1, 2

[30] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3D geometry to deformable part models. In *CVPR*, 2012. 1, 7

[31] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *ECCV*, 2012. 2

[32] S. Savarese and F.-F. Li. 3d generic object categorization, localization and pose estimation. In *ICCV*, 2007. 2

[33] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 2, 3

[34] M. Sun, G. Bradski, B.-X. Xu, and S. Savarese. Depth-encoded hough voting for joint object detection and shape recovery. In *ECCV*, 2010. 2

[35] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014. 5, 6, 7

[36] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. In *CVPR*, 2012. 2

[37] P. Yan, S. M. Khan, and M. Shah. 3d model based object class detection in an arbitrary view. In *ICCV*, 2007. 2

[38] F. Zhou and F. De la Torre. Spatio-temporal matching for human detection in video. In *ECCV*, 2014. 2

[39] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis. 3d shape estimation from 2d landmarks: a convex relaxation approach. *CVPR*, 2015. 4, 5

[40] M. Z. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3d representations for object recognition and modeling. *PAMI*, 35(11):2608–2623, 2013. 1, 2, 4