Enriching Object Detection with 2D-3D Registration and Continuous Viewpoint Estimation

Christopher Bongsoo Choy[†], Michael Stark[‡], Sam Corbett-Davies[†], Silvio Savarese[†]
[†]Stanford University, [‡]Max Planck Institute for Informatics
{chrischoy, scorbett, ssilvio}@stanford.edu, [‡]stark@mpi-inf.mpg.de

Abstract

A large body of recent work on object detection has focused on exploiting 3D CAD model databases to improve detection performance. Many of these approaches work by aligning exact 3D models to images using templates generated from renderings of the 3D models at a set of discrete viewpoints. However, the training procedures for these approaches are computationally expensive and require gigabytes of memory and storage, while the viewpoint discretization hampers pose estimation performance.

We propose an efficient method for synthesizing templates from 3D models that runs on the fly – that is, it quickly produces detectors for an arbitrary viewpoint of a 3D model without expensive dataset-dependent training or template storage. Given a 3D model and an arbitrary continuous detection viewpoint, our method synthesizes a discriminative template by extracting features from a rendered view of the object and decorrelating spatial dependences among the features. Our decorrelation procedure relies on a gradient-based algorithm that is more numerically stable than standard decomposition-based procedures, and we efficiently search for candidate detections by computing FFT-based template convolutions. Due to the speed of our template synthesis procedure, we are able to perform joint optimization of scale, translation, continuous rotation, and focal length using Metropolis-Hastings algorithm. We provide an efficient GPU implementation of our algorithm, and we validate its performance on 3D Object Classes and PASCAL3D+ datasets.

1. Introduction

Current approaches to object class detection have reached a remarkable level of performance in 2D bounding box localization [4, 5, 22, 13, 7], due to their ability to generalize across differences in object appearance, lighting, and viewpoint. While this generalization ability is beneficial for robustness, it limits the level of detail that these detectors can deliver as an output.

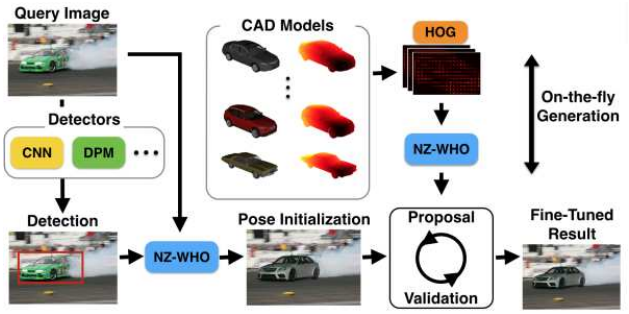


Figure 1: Using a database of 3D CAD models, we generate NZ-WHO templates which can be used to either detect objects directly or enrich the output of an existing detector with high-quality, continuous pose and 3D CAD model exemplar.

As a consequence, there has been increased interest in multi-view object recognition, where viewpoint estimates are provided by detectors in addition to 2D bounding boxes. Several attempts have been made to extend existing detectors to estimate viewpoint along with object class detections, including the implicit shape model [25], the constellation model [24], and the deformable part model [5, 8, 28, 18, 6, 11].

Recently, an even higher level of geometric detail was reached in the form of aligning 3D CAD model instances to real world test images [1, 12, 17, 2, 20]. Interestingly, the problem of matching 3D models to 2D images has been explored since the early days of computer vision [15], but had largely been neglected in recent years in favor of 2D detectors based on robust local features and statistical learning techniques. Now, this problem is being revisited for two main reasons: *(i)* the availability of 3D CAD models for many object classes and *(ii)* the availability of robust image matching techniques.

For (i), recent approaches to 2D-3D matching [1, 12] rely on a collection of 3D exemplar models, which they render from a large number of viewpoints. The resulting artificial images are then used to train exemplar models that can be matched to a real-world image at test time. For

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

(ii), it has been realized that template-based exemplar detectors based on HOG [3] features can be trained analytically, by replacing the standard SVM with an LDA classifier (E-LDA) [10]. The result is a whitened feature representation, termed WHO (Whitened Histogram of Orientations). This development makes it feasible to train hundreds of thousands of mid-level patch detectors for recognition [1]. Unfortunately, the performance of WHO relies crucially on an additional calibration step that equalizes the detection scores of independently trained exemplar models. Since this step involves costly mining of hard negative examples [3, 5] on a validation set, it constitutes the major computational bottleneck of WHO, limiting its scalability.

In this paper, we propose a novel method for 2D-3D alignment of exemplar CAD models to real-world images that circumvents the need for calibration, and greatly enhances the scalability of WHO. As a result, we can render novel views and train corresponding exemplar models *on-the-fly*, without the need for offline processing. We call these Non-Zero Whitened Histograms of Orientations (NZ-WHO) templates 1. As a by-product, we can formulate the alignment problem as a parameter search in a continuous pose space, consisting of yaw, pitch and roll, which we implement using MCMC sampling.

Our paper makes the following contributions:

First, we present a novel method for training exemplar models from rendered 3D CAD data *on-the-fly*, enabled by a novel variant of WHO, termed Non-Zero Whitened Histograms of Orientations (NZ-WHO), and making efficient use of the specific characteristics of rendered images. To our knowledge, our method constitutes the first attempt to simultaneously render and train exemplar detectors on-the-fly. Second, we demonstrate that our method can enrich the output of an existing object class detector, such as the DPM [5] or the R-CNN [7] with additional 3D information. By applying our method to candidate detections provided by the respective detector, we can augment the original detections with an estimate of 3D continuous pose and a 3D CAD model exemplar. Finally, we give an in-depth experimental study that demonstrates the effectiveness of our approach on a standard benchmark for object detection and viewpoint estimation [27].

2. Related Work

Modern object detectors generalize very well, handling intraclass variability, occlusion, truncation and viewpoint changes [5, 7]. However, this generalization comes at the cost of fine-grained information, including accurate 3D pose and object sub-category recognition. Such methods typically produce bounding box detection hypotheses, with little further information.

Many methods have attempted to move object detection towards richer outputs, especially by jointly performing detection and pose estimation [18, 28, 6, 27, 11, 1, 12]. To achieve this, [28, 11, 6] use 3D representations that deform as viewpoint changes and [18] uses geometric constraints to regularize 2D appearance models.

The methods above perform discrete pose estimation, quantizing the viewing sphere into a number of poses and selecting the best one during inference. Fine-grained pose estimators, in contrast, can infer continuous (or arbitrarily fine-grained) poses. One such method from [29] aligns a 3D deformable part-based wireframe model with input images to accurately predict object poses.

More recently, [1, 12] made progress in joint instance-level object detection and pose estimation. To estimate pose they use synthetic renderings of CAD models to learn discriminative mid-level patches. [1] calibrates these patches on a small set of real images, while [12] presents a method for learning the relative discriminativeness of the patches.

3. Approach Overview

Our method has two modes of operation.

First, it can be run in isolation, as a sliding window object detector similar in spirit to the exemplar SVM [16]. In that case, it can not only provide a detection bounding box but also meta-data such as viewpoint or 3D CAD model exemplar. As we show in our experiments, our method in isolation delivers performance that is on par with state-of-the-art for the task of object class detection and viewpoint estimation while at the same time being much faster to train and requiring no training images (Sect. 6.2).

Second, it can be used to enrich the detections of another object class detector that proposes candidate regions that can then be refined by our method. This mode constitutes the strength of our method, as we will show in our experiments (Sect. 6).

Both modes of operation rely on two steps: on-the-fly generation of exemplar templates, based on NZ-WHO, our novel whitened feature representation (Sect. 4), and pose fine-tuning using MCMC (Sect. 5).

4. On-the-Fly Template Generation

In this section, we describe our approach to generating 3D exemplar-based templates on-the-fly. It is based on 3D CAD model rendering (Sect. 4.1) followed by feature extraction and whitening. Based on the original whitening formulation (Sect. 4.2), we propose three novel extensions that enable the application of whitening to the on-the-fly setting. First, we adapt the whitening to the specific case of rendered images (Sect. 4.3). Second, we show how to speed up the whitening by two orders of magnitude for high-resolution templates (Sect. 4.4). Third, we improve the evaluation of our 3D exemplar template detectors at test time by performing convolutions in the frequency domain (Sect. 4.5).

Figure 2: An example rendering and depth image from renderer.

4.1. Rendering

We use an off-the-shelf rendering engine to generate a realistic rendering and a depth map. The CAD models we used contain texture and material information. These make the rendering more realistic and allow us to transfer natural image statistics to rendered images Fig. 2.

To handle intraclass and viewpoint variability, we used various CAD models and made renderings of these CAD models from different viewpoints. Note that we continuously vary yaw, pitch, roll and the focal length as well so that the final fine tuning stage can produce accurate viewpoint estimations (Sect. 5).

4.2. Whitened Histograms of Orientations (WHO)

Our technique for rendering and generating exemplar template detectors on-the-fly draws from recent work by Hariharan *et al*. [10]. They introduced Whitened Histograms of Orientations (WHO), which uses feature statistics from natural images to create a large number of classifiers analytically using Linear Discriminant Analysis (LDA) rather than training SVM classifiers. The confidence score for data x_i can be defined as $S(x_i) = w_{x_t}^T x_i$ where $w_{x_t} = \Sigma^{-1}(x_t - \mu)$ is an LDA classifier for a template x_t. Since collecting covariance matrices for all possible template shape is intractable, [10] assumed Wide-Sense Stationarity (WSS) of HOG features and generated a covariance Σ from autocovariance Γ collected on a large collection of natural images. In this paper, we further assumed symmetry of the autocovariance. For a 31 dimensional HOG feature x_t at location $t = (u_t, v_t)$, assuming WSS and a symmetric autocovariance, we have

$$\mu = E[x_t] = E[x_\tau] \tag{1}$$

$$\Gamma_{\|t-\tau\|} = E[(x_t - \mu)(x_\tau - \mu)] \tag{2}$$

$$= E[(x_0 - \mu)(x_{\tau - t} - \mu)]. \tag{3}$$

for all $t = (u_t, v_t)$ and $\tau = (u_\tau, v_\tau)$. In practice, we gathered Γ up to $|u_t - u_\tau| \leq 40$ and $|v_t - v_\tau| \leq 40$.

4.3. Whitening Synthesized Templates and Non-Zero WHO

Our first improvement is 'Non-Zero' whitening. When synthesizing detection templates from rendered images, a common problem is how to handle the background. If the model is rendered over a natural image background, gradients in the background will be incorporated into the discriminative template.

Alternatively, if the background is left textureless (see Fig. 2), whitening the resulting HOG template introduces strong negative weights in the textureless region (by subtracting the mean μ), as seen in Fig. 3. This could result in positive matches being suppressed due to spurious background gradients.

NZ-WHO removes these artifacts so that the background has no effect on the template response. Let a vectorized HOG feature of a rendering image $x = [\ x_1^T \quad x_2^T \quad \cdots \quad x_n^T\ ]^T \in \mathcal{R}^{nd}$ where x_i is the ith HOG cell feature, n is the number of HOG cells and d is the dimension of the HOG feature. We create a new vector $\bar{x}$ which contains only the non-zero HOG cell features of x. To be specific, let $I_d \in \mathcal{R}^{d \times d}$ be the identity matrix, $\bar{n}$ be the number of non-zero HOG cells, and a matrix $S \in \mathcal{R}^{\bar{n}d \times nd}$ be the masking matrix that selects non-zero HOG cells. For instance, a template has $n = 3, \bar{n} = 2$, and only the second HOG cell has 0 norm, then

$$S = \begin{bmatrix} I_d & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_d \end{bmatrix} \tag{4}$$

The matrix selects HOG features that correspond to the first and third cells.

Using the selector S, we can define new vectorized HOG features $\bar{x} = Sx, \bar{\mu} = S\mu, \bar{\Sigma} = S\Sigma S^T$. After solving the resulting system (which is now smaller than in the WHO approach) we find

$$\bar{w} = \bar{\Sigma}^{-1}(\bar{x} - \bar{\mu}) \tag{5}$$

To speed up the convolution, we restore zero cells and reshape the vector $\bar{w}$ and compute convolution.

4.4. Fast Whitening using Conjugate Gradient

Synthesizing the LDA template in Eq. 5 requires solving the system of linear equations, $\bar{\Sigma}\bar{w} = (\bar{x} - \bar{\mu})$. In [10], the authors make use of the fact that covariance matrices are symmetric and positive semidefinite to solve the system via the Cholesky decomposition with Gaussian Elimination, which requires $O(n^3)$ time.

The Conjugate Gradient method is an iterative algorithm for solving symmetric positive definite systems which runs in $O(n^2\kappa)$ time, where κ is the condition number of the matrix [23]. This makes Conjugate Gradient faster than decomposition for matrices with small condition numbers relative to their size.

The covariance matrix for HOG templates is typically ill-conditioned [10], but adding a small regularization constant to the diagonal reduces its condition number. We use a

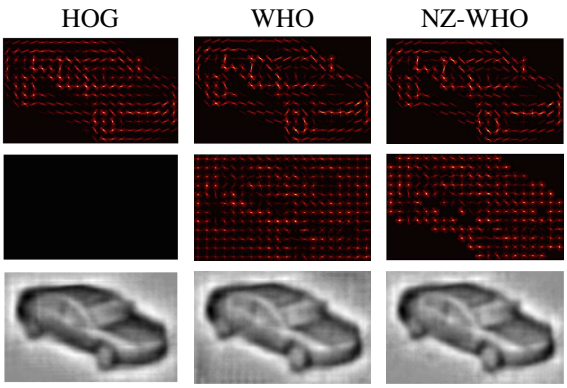


Figure 3: Comparison of HOG, WHO and NZ-WHO. Visualization of positive weights (first row), visualization of negative weights (second row), HOGgles [26] (third row). Note that for WHO, whitening all cells results in strong negative edges on the empty region.

constant of 0.15, which reduces the condition number from 10^{20} to 50, much smaller than the dimension of the matrix (7000).

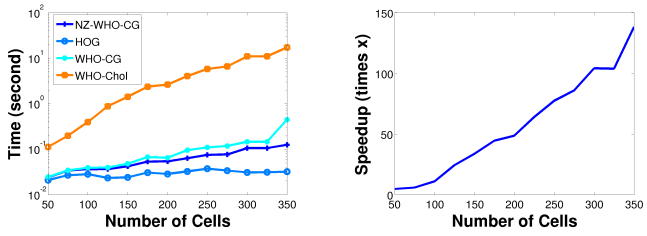
As a result, a GPU implementation of conjugate gradient converges in 60 ms when using 250 HOG cells on Nvidia GTX660, two orders of magnitude faster than using Cholesky factorization with Gaussian Elimination.

We report the real time analysis of whitening using decomposition and Conjugate Gradient methods in Fig. 4. (a) compares the absolute runtime of the different methods while (b) gives the obtained speedup. We see that whitening the HOG template takes several seconds for realistic template sizes of several hundred cells, but only 60 ms using the Conjugate Gradient method. If we use NZ-WHO, we can gain extra speed up since we only whiten non-zero cells.

In addition, since the iterative Conjugate Gradient method directly tries to reduce the residual (the norm of $y - Ax$ for $Ax = y$), it is more numerically stable than Cholesky decomposition with Gaussian Elimination. In Fig. 5 we vary the number of cells in a template and show that the residual of NZ-WHO is smaller than that of WHO.

4.5. High Resolution Templates and FFT-based Convolution

We generate high resolution templates with more than 250 HOG cells to capture details of an object to give accurate 2D-3D matching. These large templates cause computational burden when computing convolution. Though good for accurately determining model and pose, these large templates slow down the convolution since computation time scales linearly with the number of HOG cells in the template. To overcome this, we used FFT-based GPU convolution [19]. Briefly, for length n signal and length m filter, naive convolution takes $O(nm)$ time whereas FFT-based convolution takes $O\left((n+m)\log(n+m)\right)$ time. For large m (high resolution templates), we can gain computational advantage.



(a) (b)

Figure 4: (a) Runtime analysis of whitening. HOG means feature extraction time, WHO-Chol uses our implementation of [10] and WHO-CG uses iterative Conjugate Gradient. NZ-WHO-CG (Ours) uses only non-zero cells and Conjugate Gradient. (b) final speedup of NZ-WHO vs. WHO.

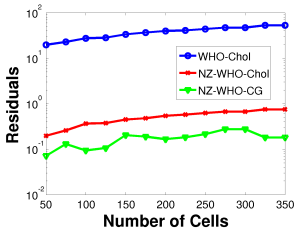


Figure 5: Residuals of different method

5. Pose Fine-Tuning via MCMC

The NZ-WHO template matching method we have presented (Sect. 4.3) makes template generation and evaluation computationally inexpensive. This means that we can use a hypothesize-and-test scheme to efficiently explore the continuous parameter space to find the best object pose, scale, 3D CAD model type and camera focal length. In particular, we propose to implement this parameter search as a Markov Chain Monte Carlo (MCMC) procedure based on the Metropolis-Hastings algorithm.

Probabilistic formulation. We parameterize the continuous parameter space as $\theta = [v, m, f]$, where v is the 3D rotation of the CAD model, m is the discrete CAD model index, and f is the focal length.

We model the probability of an object with the parameter θ in the test image $\mathcal{I}$ as a distribution in the exponential family, and let

$$P(\theta|\mathcal{I}) \sim e^{\max_s w(\theta) * \mathcal{T}_s(\mathcal{I})}, \tag{6}$$

where $\max_s w(\theta) * T_s(\mathcal{I})$ is the maximum convolution score of NZ-WHO template $w(\theta)$ with image features $T_s(\mathcal{I})$ for all scale s, as defined in Sect. 4.3.

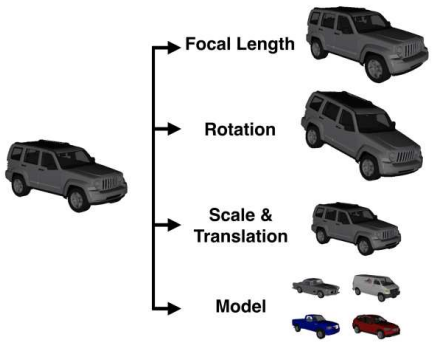


Figure 6: Search space that our fine-tuning stage handles (Sect. 5).

Inference. We approximate the MAP solution for θ by drawing samples from the distribution $P(\theta|\mathcal{I})$, using the Metropolis-Hastings algorithm. Specifically, we use a variant that changes only a single component of the parameter vector θ at a time, termed Single Component Metropolis-Hastings [9].

This algorithm changes the current state θ to a new state θ^+ based on the acceptance probability

$$A(\theta \rightarrow \theta^+) = min\left(1, \frac{P(\theta^+|\mathcal{I})g(\theta^+ \rightarrow \theta)}{P(\theta|\mathcal{I})g(\theta \rightarrow \theta^+)}\right) \quad (7)$$

We define 3 different types of moves that can alter the state (Fig. 6), *(i)* changing the focal length f, *(ii)* one of the rotational pose parameters v_i, and *(iii)* CAD model index m. For (i) and (ii), we use Gaussian proposal distributions, and a uniform distribution over model changes for (iii). We implicitly compute all possible translation and scale by convolving $w(\theta^+)$ with a HOG pyramid.

$$g(\theta \rightarrow \theta(v_i^+)) \sim \mathcal{N}(\theta_{v_i}, \sigma_v) \quad \text{for } i \in \{1,2,3\} \quad (8)$$

$$g(\theta \rightarrow \theta(f^+)) \sim \mathcal{N}(\theta_f, \sigma_f) \quad (9)$$

$$g(\theta \rightarrow \theta(m^+)) \sim (1-c)\delta(\theta_m) + c\,\text{Unif}(1, M) \quad (10)$$

where g is the proposal distribution, and $m \in \{1, \cdots, M\}$.

In practice, we run this algorithm for 20 iterations, keeping the sample with the highest probability as our approximate estimate of the MAP solution. We set $\sigma_v = 5°$, $\sigma_f = 1$ and $c = 0.1$ for all experiments.

Initialization. Since the objective function is non-convex, we need a good initialization in order to increase the probability that we find a solution close to the global optima. We achieve this by first running a discrete, pre-trained version of our algorithm (i.e., an ensemble of NZ-WHO templates) in order to get promising candidate 2D bounding boxes and poses to start from. We then initialize θ for each of these candidates.

6. Experiments

In this section, we give an experimental evaluation of our approach, highlighting three different aspects. First, we verify that our NZ-WHO method delivers performance that is at least on par with the original WHO formulation [10] in terms of accuracy, while at the same time resulting in large computational savings (Sect. 6.1). Second, we demonstrate that our method can be used for multi-view object class detection in isolation. It can be applied in a sliding window fashion and deliver 2D bounding box as well as viewpoint information. Our method is competitive with the state-of-the-art in this case (Sect. 6.2). Finally, we show that our method can be used to complement the detections provided by an existing object class detector, such as DPM [5] or RCNN [7]. In this case, we show a considerable performance improvement compared to previous work in the task of joint object class detection and VP estimation (Sect. 6.3).

Setup. We use established benchmark datasets to validate our approach, namely the 3D Object Classes dataset [21], and PASCAL3D+ [27], a recently proposed extension of Pascal VOC'12 [4] that provides additional annotations in the form of aligned 3D CAD models. In both cases, we use the test data provided by the respective datasets, but train our models entirely from rendered 3D CAD model images.

6.1. WHO Variants

To validate our approach, we run an ensemble of NZ-WHO templates as a bank of detectors and compare its performance with other WHO variants, notable WHO [10] and the original non-whitened HOG [3]. In addition, we evaluate WHO-CG and WHO-CG-Z: WHO-CG uses iterative Conjugate Gradient method to generate WHO. WHO-CG-Z whitens the whole template, but zeros out textureless region. NZ-WHO-CG is the NZ-WHO which whitens only non-zero cells using iterative Conjugate Gradients (our method).

Tab. 1 gives the corresponding results for the various methods on a subset of the 3D Object Classes car dataset [21] (all images corresponding to one particular car instance, seen from different viewpoints), reporting 3 quantities: detection performance in average precision (AP), pose estimation in mean precision in pose estimation (MPPE), and the respective runtime. For all methods, the table gives the results with and without calibration, using the calibration method of [1]. This calibration learns affine transformation of the detection confidence. For each method, we generated templates for 1 CAD model exemplar rendered from 24 azimuth and 4 elevation angles.

Results. In Tab. 1, we observe that, on average, calibration indeed improves performance in terms of AP, sometimes drastically (e.g., from 54.4 to 92.8 for WHO-CG-Z), as observed in prior work [1]. At the same time, calibration

is computationally expensive, resulting in generation time that are two orders of magnitudes larger than without calibration. Strikingly, our method NZ-WHO-CG achieves the second best AP of 90.0 while completing in 79 ms vs. 8.5 s when using calibration.

Methods (AP/MPPE)	before calib.	synth. time	after calib. [1]	calib. time
HOG[3]	72.3 / 65.0	31ms	60.4 / 50.2	8.7 sec
WHO[10]	82.1 / **85.4**	6162ms	84.4 / 83.0	15.4 sec
WHO-CG	81.7 / 84.9	104ms	83.7 / **87.3**	8.3 sec
WHO-CG-Z	54.4 / 65.1	103ms	**92.8** / 86.7	8.7 sec
NZ-WHO-CG *(ours)*	**90.0** / 82.8	79ms	90.3 / 86.8	8.5 sec

Table 1: Average Precision (AP), Mean Precision in Pose Estimation (MPPE) [14] of variants of WHO on 3D Object Classes cars [21], and their corresponding synthesis and calibration time per template. Please see text for details.

6.2. 2D-3D Matching as an Object Detector

In this section, we evaluate the performance of our method in isolation, for the task of object class detection and viewpoint estimation, on the 3D Object Classes dataset [21], for the categories *car* and *bicycle*.

To that end, we create an ensemble of NZ-WHO templates (Sect. 4) using 9 different CAD models and a total of 192 different viewpoints: 4 elevation angles, 24 azimuth angles, and 2 focal lengths. We run the entire ensemble exhaustively over each test image in a sliding window fashion.

Performance is measured in Average Precision (AP) for object detection and Mean Precision in Pose Estimation (MPPE) [14]. Pose estimation is here understood as a discrete problem in which the predicted azimuth angle is binned into a set of 8 discrete viewpoint classes.

Results. Tab. 2 gives the corresponding results, comparing our method to two recent state-of-the-art baselines, the aspect layout model (ALM [28]), and the DPM-VOC+VP [18]. Fig. 7 gives qualitative results.

We observe that our model performs on par with the state-of-the-art methods in terms of AP (99.8) for cars. It performs slightly worse on bicycles than DPM-VOC+VP (93.0 vs. 98.8), but on par with the ALM (93.0). In viewpoint estimation, our model performs slightly worse than both methods (91.7 vs. 93.4 and 97.5 for cars, and 90.9 vs. 91.4 and 97.5 for bicycles, respectively).

This result is encouraging, since our approach reaches a level of performance that is on par to current state-of-the-art while at the same time being much faster to train. It takes merely a few minutes to train, while both ALM [28] and DPM-VOC+VP [18] are complex models that optimize non-convex objective functions during training, which is only made tractable by resorting to delayed constraint generation in the form of hard negative mining, and can easily take a day on a single machine. In addition, our approach uses only rendered images, avoiding the need for real-world training data with costly bounding box and viewpoint annotations.

AP/MPPE	Ours	ALM[28]	DPM-VOC+VP[18]
car	**99.8** / 91.7	98.4 / 93.4	**99.8 / 97.5**
bicycle	93.0 / 90.9	93.0 / 91.4	**98.8 / 97.5**

Table 2: Average Precision (AP) and Mean Precision in Pose Estimation (MPPE) on 3D Object Classes [21] cars.

Figure 7: Detection results on 3D Object Classes [21]. Original image (left) and detection result overlaid on top (right).

6.3. Enriching Existing Detections

In this section, we use our method to enrich the detections provided by existing, high-performance object detectors with additional output in the form a 3D pose, focal length, and 3D CAD model exemplar shape.

To show such ability, we evaluate our method on the PASCAL3D+ dataset [27]. This dataset augments PASCAL 2012 images with high quality viewpoint annotations thus is ideal to measure pose estimation. The dataset proposes a new metric called Average Viewpoint Precision (AVP) where it measures the area under viewpoint precision and detection recall curve. The viewpoint is measured by azimuth similarity. If the distance between predicted azimuth and ground truth azimuth is below a certain threshold, the viewpoint is correct. The baseline methods V-DPM [27] and DPM-VOC+VP [18] reported on the PASCAL3D+ dataset are variants of DPM [5] where each component of DPM accounts for azimuth. Thus V-DPM and DPM-VOC+VP provide discrete azimuths only whereas our method provides 3D viewpoint (yaw, pitch, roll), CAD model instance (model index, rendering, depth) and focal length.

Setup. We use detection bounding boxes provided by both object detectors and use our method to perform fine-grained viewpoint estimation. We use detection bounding boxes from two different methods: DPM-VOC+VP [18] and R-CNN [7], both in their variants trained on 8 viewpoint categories, since these perform best in terms of AP. For both

Figure 8: Effect of fine tuning. (left) original image, (middle) initial detection, (right) continuous fine tuning using Single-Component Metropolis Hastings

cases, we use the original score for plotting precision-recall curves (meaning that we can not improve over their AP). We revert back to a default viewpoint prediction (0° azimuth) in case the confidence of our method falls below a threshold.

For both cases, we compare the performance of a discrete incarnation of our method (Sect. 4) and our full model, including the fine-tuning based on MCMC (Sect. 5).

Quantitative results are given in Tab. 3 in terms of AP and AVP and in Fig. 10 in terms of precision-recall plots and viewpoint confusion matrices. Example outputs of our method applied to candidate object detections are given in Fig. 9, and the effect of fine-tuning is visualized in Fig. 8. More qualitative results can be found in the supplemental material.

Results. In Tab. 3, we make two main observations. First, we see that adding our method to DPM-VOC+VP consistently improves performance in terms of AVP, for both car and bicycle, and across all viewpoint bins. This is already the case for our discrete method: the improvement ranges from 0.3 for bicycle-8v to 5.8 for car-4v. Using the R-CNN as the base detector increases AVP even more, in particular for the bicycle class: for bicycle-4v, our discrete method improves over the corresponding DPM-VOC+VP result by 12.6.

The second observation is that the fine-tuning based on MCMC can indeed improve pose estimation performance slightly, e.g., by 1.1 for bicycle-16v and using the DPM-VOC+VP as the base detector, or by 1.1 for bicycle-16v when using the R-CNN as the base detector. In Fig. 8, we visualize the effect of fine-tuning qualitatively for two different test images. In both cases, the fine-tuned pose is a better visual match to the true 3D pose.

Robustness. While the R-CNN detector [7] is highly robust to variations in object appearance and even occlusion and truncation, the resulting bounding box detections vary largely in the object portions that they contain, which provides a major challenge to any method that uses these detections as an input for further processing, such as ours. In order to accommodate this variability, we add a considerable

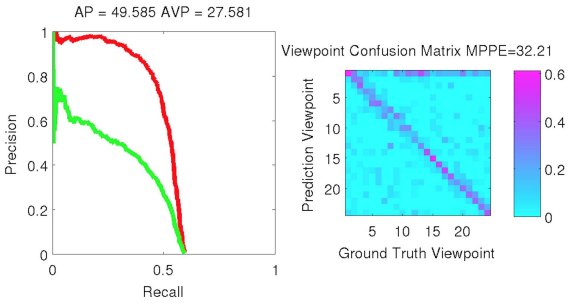


Figure 10: Average Precision (AP)(red) and Average Viewpoint Precision (AVP)(green) and viewpoint confusion table on PASCAL 12 car validation set using R-CNN + Ours (full) for 24 views. All four viewpoint discretizations are available on the supplementary paper.

context region around the proposed bounding boxes before running our method. We assume that the object can be arbitrarily truncated by the bounding box and search all plausible scales and translations. This can be efficiently computed using FFT-based convolution (Section. 4.5).

Fig. 11 visualizes example outputs of our method when starting from different proposed R-CNN bounding boxes. As can be seen from the figure, although the input bounding boxes (cyan) are often irregular and contain truncated objects, our method reliably generates a reasonable prediction of pose, translation, scale and CAD model (magenta bounding boxes enclose the output of our system).

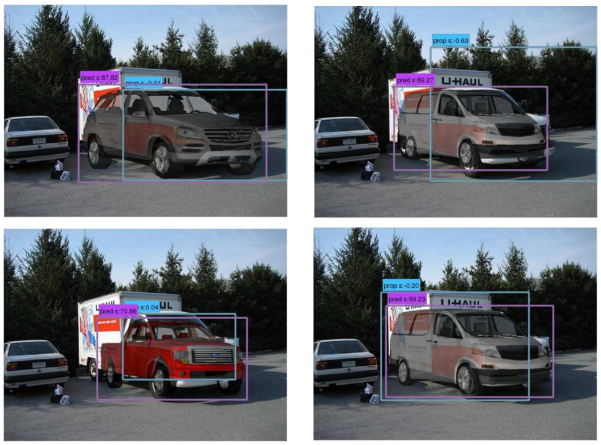


Figure 11: Robustness of our method against irregular and truncated R-CNN detection proposals (cyan).

7. Conclusion

We have proposed a method for generating 3D CAD model exemplar templates on-the-fly, based on a novel variant of WHO features (NZ-WHO). It circumvents the need for calibration, is computationally efficient, and allows it to be run in an on-the-fly setting. As a result, we can use our method to enrich existing object detections with additional

Figure 9: Examples of enriched bounding boxes. Given R-CNN [7] detection bounding boxes, our method predicted 2D-3D matching reasonably. The first column shows bounding box candidates produced by R-CNN detection. Subsequent columns show the output of our method given the bounding box from R-CNN detection candidates. Blue boxes are R-CNN output and purple boxes are the tightest bounding box enclosing predicted CAD model.

AP/AVP	V-DPM [27]	DPM-VOC+VP [18]	[18] + Ours (discrete)	[18] + Ours (full)	R-CNN + Ours (discrete)	R-CNN + Ours (full)
car-4v	37.2 / 20.2	45.6 / 36.9	47.6 / 42.7	47.6 / **42.7**	49.6 / 41.5	49.6 / 41.5
car-8v	37.3 / 23.5	47.6 / 36.6	47.6 / **39.8**	47.6 / 39.5	49.6 / 38.0	49.6 / 39.0
car-16v	36.6 / 18.1	46.0 / 29.6	47.6 / 32.7	47.6 / 33.0	49.6 / 34.0	49.6 / **34.3**
car-24v	36.3 / 13.7	42.1 / 24.6	47.6 / 27.4	47.6 / 27.4	49.6 / 27.0	49.6 / **27.6**
bicycle-4v	45.2 / 41.7	46.9 / 43.9	48.1 / 47.6	48.1 / 46.6	61.7 / 56.5	61.7 / **56.7**
bicycle-8v	47.3 / 36.7	48.1 / 40.3	48.1 / 40.6	48.1 / 40.6	61.7 / 48.9	61.7 / **49.2**
bicycle-16v	46.5 / 18.4	45.6 / 22.9	48.1 / 26.2	48.1 / 27.3	61.7 / 34.7	61.7 / **35.8**
bicycle-24v	44.4 / 14.3	45.9 / 16.7	48.1 / 21.5	48.1 / 20.9	61.7 / **27.0**	61.7 / 23.9

Table 3: Average Precision (AP) and Average Viewpoint Precision (AVP) on PASCAL3D+ [27]. For combined methods (* + Ours), we use bounding boxes from * and augment viewpoint using our method.

information such as precise 3D pose and 3D CAD model exemplar. Combined with an R-CNN detector, we achieve state-of-the-art results in joint detection and VP estimation.

Acknowledgement We acknowledge the support of NSF CAREER grant (N1054127), Ford-Stanford Innovation Alliance Award, DARPA, Fulbright New Zealand, Korea Foundation for Advanced Studies and the Max Planck Center for Visual Computing & Communication.

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] T. Chen, Z. Zhu, A. Shamir, S.-M. Hu, and D. Cohen-Or. 3sweepp: Extracting editable objects from a single photo. *ACM Trans. Graph.*, 2013.

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

[4] 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.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010.

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

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

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

[9] H. Haario, E. Saksman, and J. Tamminen. Component-wise adaptation for high dimensional mcmc. *Computational Statistics*, 20(2):265–273, 2005.

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

[11] M. Hejrati and D. Ramanan. Analysis by synthesis: 3d object recognition by object reconstruction. In *CVPR*, 2014.

[12] A. K. J. Lim and A. Torralba. Fpm: Fine pose parts-based model with 3d cad models. In *ECCV*, 2014.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *NIPS*. 2012.

[14] R. Lopez-Sastre, T. Tuytelaars, and S. Savarese. Deformable part models revisited: A performance evaluation for object category pose estimation. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 2011.

[15] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 1987.

[16] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[17] A. E. Natasha Kholgade, Tomas Simon and Y. Sheikh. 3d object manipulation in a single photograph using stock 3d models. *ACM Transactions on Computer Graphics*, 2014.

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

[19] V. Podlozhnyuk. FFT-based 2d convolution, June 2007.

[20] K. Rematas, T. Ritschel, M. Fritz, and T. Tuytelaars. Image-based synthesis and re-synthesis of viewpoints guided by 3d models. In *CVPR*, 2014.

[21] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *ICCV*, 2007.

[22] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, 2013.

[23] J. R. Shewchuk. An introduction to the conjugate gradient method without the agonizing pain, 1994.

[24] M. Stark, M. Goesele, and B. Schiele. Back to the future: Learning shape models from 3D CAD data. In *BMVC*, 2010.

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

[26] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. *ICCV*, 2013.

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

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

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