Occlusion-Net: 2D/3D Occluded Keypoint Localization Using Graph Networks

N. Dinesh Reddy Minh Vo Srinivasa G. Narasimhan
Carnegie Mellon University
{dnarapur,mpvo,srinivas}@cs.cmu.edu

Abstract

We present Occlusion-Net, a framework to predict 2D and 3D locations of occluded keypoints for objects, in a largely self-supervised manner. We use an off-the-shelf detector as input (e.g. MaskRCNN [16]) that is trained only on visible key point annotations. This is the only supervision used in this work. A graph encoder network then explicitly classifies invisible edges and a graph decoder network corrects the occluded keypoint locations from the initial detector. Central to this work is a trifocal tensor loss that provides indirect self-supervision for occluded keypoint locations that are visible in other views of the object. The 2D keypoints are then passed into a 3D graph network that estimates the 3D shape and camera pose using the self-supervised reprojection loss. At test time, Occlusion-Net successfully localizes keypoints in a single view under a diverse set of occlusion settings. We validate our approach on synthetic CAD data as well as a large image set capturing vehicles at many busy city intersections. As an interesting aside, we compare the accuracy of human labels of invisible keypoints against those predicted by the trifocal tensor.

1. Introduction

Virtually any scene has occlusions. Even a scene with a single object exhibits self-occlusions - a camera can only view one side of an object (left or right, front or back), or part of the object is outside the field of view. More complex occlusions occur when one or more objects block part(s) of another object. Understanding and dealing with occlusions is hard due to the large variation in the type, number and extent of occlusions possible in scenes. As such, occlusions are an important reason for failure of many computer vision approaches for object detection [9][14][34][16], tracking [49][5][44][41], reconstruction [20][19] and recognition, even today’s advanced deep learning based ones.

The computer vision community has collectively attempted numerous approaches to deal with occlusions [12].

Figure 1: Accurate 2D keypoint localization under severe occlusion in our CarFusion dataset. Different colors depicts different objects in the scene.

[13][26][35] for decades. Bad predictions due to occlusions are dealt with as noise/outliers in robust estimators. Many methods provide confidence or uncertainty estimates to downstream approaches that need to sort out whether the uncertainty corresponds to occlusion. But it is hard to predict performance as they usually do not take occlusions explicitly into account.

On the other hand, occlusions are explicitly treated as missing parts in model fitting methods [50][40]. These approaches have had better success as they exploit a statistical model of a particular type of object (e.g. car, human, etc.). But much remains to be done. For instance, severe occlusions, such as when a large part of an object is blocked, can result in poor fitting [52]. Further, often these approaches do not explicitly know which parts of an object are missing and attempt to simultaneously estimate the model fit as well as the missing parts.

In this work, we present an approach to explicitly predict 2D and 3D keypoint locations of the occluded parts of an object using graph networks, in a largely self-supervised manner. Our method receives as input, the output of any detector (e.g., the MaskRCNN architecture [16]) that has been trained on a particular category of object with human supervision of only visible keypoints and their types (e.g., front, back, left, right). Implicitly, then, the key points that are not labeled are assumed to be invisible. This is the
Figure 2: Occlusion-net: We illustrate the overall approach to training a network to improve localization of occluded keypoints. The input is a ROI region from any detector, which is passed through multiple convolutional layers to predict the heatmaps with a confidence score. These confidences are passed through a graph encode-decoder network and trained using multi-view trifocal tensor loss for localization of occluded 2D keypoints. The output from the decoder is passed through a 3D encoder to predict the shape basis and the camera orientation. This network is a self-supervised graph network and trained using reprojection loss with respect to the 2D decoder output.

only human supervision used in this work. The detector usually provides an uncertainty of all key point locations. We first show that the distribution of the uncertainties for visible and occluded points overlap significantly, making it hard to predict which key points are occluded at test time. To address this issue, we design an encoder-decoder graph network that first predicts which edges have an occluded node, and then localizes the occluded node in 2D in the decoder. Visible or invisible edge classification is trained using the implicit non-labeled supervision of occluded points. We then train the decoder graph network to localize invisible keypoints using multiple wide-baseline views of objects. Our observation is that while some parts may be missing in one view, they are visible and labeled in another view. But how do we provide supervision for a hidden point location in a view? We use two views where a keypoint is seen (and labeled by humans) and compute the trifocal tensor using camera matrices to predict its location in the view where the keypoint is occluded. We call this the Trifocal tensor loss, which is minimized to correct the 2D keypoint positions from the initial detector. Compared to other approaches that use multiple views [38, 32, 47], our approach explicitly predicts occluded keypoints.

The predicted 2D keypoints (both occluded and visible) are then used in a graph network to estimate the 3D object shape and the camera projection matrix. Similar to previous work [52, 59], we will estimate the parameters of a shape basis computed a priori of the object of interest. The training is performed in a self-supervised way by minimizing the reprojection loss i.e. error between the reprojection and the predicted 2D keypoint locations. We train the entire pipeline, called Occlusion-net, end-to-end with the aforementioned losses.

We evaluate our approach on images of vehicles captured at busy city intersections with numerous types and severity of occlusions. The dataset extends the previous CarFusion dataset [32] to include many more city intersections, where 18 views of the intersection are simultaneously recorded. A MaskRCNN car detector is trained using 100000 cars, with human labeled visible keypoints to produce a strong baseline for our method to compare to and build upon. Our Occlusion-net significantly outperforms (about 10%) this baseline across many metrics and performs well even in the presence of significant occlusions (see Figure 1). As an interesting exercise, we also show a comparison of the trifocal loss against human labeling of the 2D occluded point locations and observe that humans label around 90% of the points to lie within the acceptable range of error. We also evaluate our approach on a large synthetic CAD dataset, showing similar performance benefits and improvements of up to 20% for occluded keypoints. Our network is efficient to train and can localize keypoints in 2D and 3D in real-time (more than 30 fps) at test-time. While we have demonstrated our approach on vehicles, the framework is general and applies to any object category.

2. Related Work

Occlusion Detection: While there has been significant progress in predicting the visible keypoints by using part detectors learned from CNNs [13, 42, 26, 2, 27, 46], most of these methods fail short to precisely localize occluded keypoints. Using synthetic data, Moreno et al. [31] show that such occlusion modeling is crucial. To address this problem, many methods employ active shape models [6] for vehicle detection under occlusion [51, 52, 43]. However, these methods only model self-occlusions and omit often seen occlusions by other objects. Recently, [37, 32] propose a multi-view bootstrapping approach to generate ac-
The relationship between all nodes is encoded in the edge $V$ vertex of the graph as points and nodes \[16\][33] to compute the location of all the keypoints from multiple views to supervise occluded points. We use the heatmap based method, a graph encoder to model the occlusion statistics of the graph, and a graph decoder inferring the 2D locations of the keypoints. Since, we do not know the underlying graph, we use the GNN to predict the latent graph structure. The encoder is modeled as $g(E_{ij}|V) = \text{softmax}(f_{enc}(V))$ where $f_{enc}(V)$ is a GNN acting on the fully connected graph produced from the heatmaps. Given the input graph our encoder computes the following message passing operations to produce the occlusion statistics:

\begin{align}
  h^1_e &= f_{enc}(V_j) \\
  v \rightarrow e : h^1_{(i,j)} &= f^1_e([h^1_i, h^1_j]) \\
  e \rightarrow v : h^2_e &= f_e(\sum_{i \neq j} h^1_{(i,j)}) \\
  v \rightarrow e : h^2_{(i,j)} &= f^2_e([h^2_i, h^2_j])
\end{align}

The de-
Here $E_{ij,p}$ denotes the p-th element of the vector $E_{ij}$. An important thing to observe is the current state is added into Eq. [7] so inherently the model is learning to deform the keypoints i.e predict the difference $\Delta \mathbf{V} = \mathbf{V}^p - \mathbf{V}$. Further in Eq. [7] $\mu$ is the mean location predictor and $\mathcal{N}$ produces the probability of the locations. We only minimize the distance between the predicted and ground truth occluded points in this network using a trifocal tensor loss.

**Trifocal Tensor Loss.** We exploit multiple views of the object captured “in the wild” to estimate the occluded keypoints. The assumption is that the keypoints occluded in one view are visible in two or more different views. Thus, the trifocal tensor [15] can transfer the locations in the two object captured “in the wild” to estimate the occluded keypoints and 3D shape from a single view on the object in the camera coordinate frame [15]. We assume the principle point of the camera is at the origin. To account for the normalization of the image to a square matrix from the original dimensions, we re-scale the projected 2D points by $s = w/h$, where $w$ and $h$ denote the width and height of the input image (see [22] for further details).

**Keypoint Reprojection Loss:** We train the 3D-Keypoint Graph network in a self-supervised manner using the reprojection loss, i.e. the difference between the projected 3D keypoints and the keypoints computed from the 2D-KGNN:

$$L_{Reproj} = \sum_{j \in k} ||\pi(W_j) - \mathbf{V}^\pi_j||^2$$

The use of the 3D basis shape allows explicit enforcement of 3D symmetry which provides further constraints for the 2D keypoint estimation via the reprojection loss.

**3.2. 3D-Keypoint Graph Neural Network**

Given the graph from the 2D-KGNN decoder, the 3D-keypoint graph neural network encoder predicts a 3D object shape $W$ and the camera projection matrix $\pi$. This encoder takes as input the graph and predicts the 3D location of the all the keypoints using a self-supervised projection loss. Mathematically, this is formulated as $q(\beta, \pi | \mathcal{V}) = f_{enc}(\mathcal{V})$, where, $\beta$ are the deformation coefficients of PCA shape basis of the object and $\pi$ is the camera projection matrix.

**Shape Basis:** We model the shape as a set of 3D keypoints corresponding to the predicted 2D keypoints. We compute the mean shape $b_0$ and n principal shape components $b_j$ and corresponding standard deviations $\sigma_j$, where $1 \leq j \leq n$, using the 3D repository of the object [3] with annotations of 3D keypoints from [26]. Given the shape bases, any set of deformable 3D keypoints can be represented as a linear combination of the $n$ principal components $\beta$ as $W = b_0 + \sum_{k=1}^{n} \beta_k * \sigma_k * b_k$.

**Camera Projection Matrix:** Let $\pi(W)$ be the function that projects a set of 3D keypoints $W$ onto the image coordinates. We use the perspective camera model and describe $\pi$ as a function of the camera focal length $f$, the rotation $q$, represented as quaternion, and translation $t$ of the object in the camera coordinate frame [15]. We assume the principle point of the camera is at the origin. To account for the normalization of the image to a square matrix from the original dimensions, we re-scale the projected 2D points by $s = w/h$, where $w$ and $h$ denote the width and height of the input image (see [22] for further details).

**4. Experimental Results**

We demonstrate the ability of our approach to infer occluded keypoints and 3D shape from a single view on the new and challenging CarFusion dataset. We first describe this dataset in section 4.1. We then perform ablative analysis of the algorithm in Section 4.2. Finally, we show qualitative comparisons against the state of art Mask-RCNN [16] detector in section 4.3. For a fair comparison, we retrain this baseline model on our dataset. In the evaluation metrics, 2D-KGNN refers to the output after the decoder layer and 3D-KGNN refers to the projections of predicted 3D keypoints onto the image.

**4.1. Datasets**

**Car-render Self-occlusion dataset:** We use the 472 cars sampled from shapenet [4] and 3D annotated by [26]. We select 12 keypoints from the annotated 36 keypoints and render them from different viewpoints. The viewpoints are randomly selected on a level 5 Icosahedron, at varying focal lengths and distances from the object. We use 300 synthetic CAD models for training, 72 for validation and 100 for testing. We project the 3D keypoint annotations of the CAD model with visibility, we trace a ray toward the object from a pixel and check if the first intersection is close to the ground truth location to determine visibility.
Figure 3: We analyze the need for a 2D-KGNN encoder. The left image shows the confidence score of the heatmaps from the baseline method (the distribution is colored based on Ground Truth visibility). The right image shows the ROC curve of the predictions from graph encoder and baseline. At 0.1 false positive rate, the baseline returns 0.5 true positive rates compared to 0.8 of the 2D-KGNN.

CarFusion dataset: To model a wide range of real occlusions, we collect an extensive dataset captured simultaneously by multiple mobile cameras at 60fps at 5 crowded traffic intersections (extending previous work [32]). This extended dataset consists of 2.5 million images out of which 53000 images were sampled at uniform intervals from each video sequence. Approximately, 100000 cars detected in these images were annotated with 12 keypoints each. Each annotation contains the visible and occluded keypoint locations on the car. We do not use the occluded keypoints for training the Occlusion-Net. We selected four annotated intersections to train the network while using one intersection to test it, which split the annotation data into 36000 images for training and 17000 for testing. We further compute 90-10 train validation split on the training data to validate our training algorithm. The dataset was completely captured “in the wild” and contains numerous types and severity of occlusions.

Preprocessing: Computing the trifocal loss requires the virtual camera poses in the object frame. For every image, the virtual pose is estimated by solving a PnP [23] between the visible keypoints and the 3D points computed from [32].

4.2. Quantitative Evaluation

We compare our approach with other state-of-the-art keypoint detection networks. We use the PCK metric [47] to analyze both the 2D and the 3D occluded keypoint locations. According to the PCK metric, a keypoint is considered correct if it lies within the radius $\alpha L$ of the ground truth. Here $L$ is defined as the maximum of length and width of the bounding box and $0 < \alpha < 1$. To evaluate the 3D reconstruction, we project the reconstructed keypoints into their respective views and compute the 2D PCK error.

Occlusion Prediction: We demonstrate that the confidence scores computed using MaskRCNN is insufficient to predict occlusions. The left image in Fig 5 shows the distributions of confidence scores of occluded and visible points. These distributions overlap significantly making it hard to distinguish occluded points from visible points. In contrast, by modeling a graph network to exploit relative locations of the keypoints, we observe a significant boost in the accuracy of occlusion prediction as seen from the right image in figure 3. We observe an AUC of 0.83 with MaskRCNN, whereas 2D-KGNN gives an AUC of 0.95.

Evaluations of visible points: We show evaluation of our network with respect to existing visible keypoint estimation methods. Both 3D-KITTI[24] and PASCAL3D+ [45] datasets have annotations only for visible keypoints and do not contain occluded point annotations or multiple views to directly evaluate our method. The 2D keypoint predictions in [24] are evaluated only on visible keypoints and the 3D model is evaluated by fitting only visible keypoints on objects that are not truncated or occluded by other objects (“Full” in their table). Our model has not been trained on either of these datasets or the CAD dataset from [24]. Table 1 compares our method against those on the annotated 2D visible points in 3D-KITTI. Table 1 also shows the evaluation against the ground truth 3D model for the “Full” (unoccluded) case - the only case mentioned in [24]. We observe that our approach outperforms the other methods for two categories i.e. Truncation and oth-Occlusion. This can be attributed to the fact that our dataset models a range of occlusion types and severity.

Importance of 3D-KGNN: The 3D pose computed is useful for traffic analysis (speed, flow) and understanding/visualizing activity at busy city intersections. 3D-KGNN can also be used to find correspondence across views for multi-view reconstruction, especially when there are very few views available and the keypoints may be occluded. Figure 4 demonstrates that 3D-KGNN finds significantly more inliers for multiview correspondence compared to 2D-KGNN or MaskRCNN.
Human Annotation vs Geometric Prediction: The CarFusion dataset has annotated keypoints for occluded points as well as the visible points across multiple views. Thus, as an interesting aside, we evaluate the accuracy of hand-labeled occluded points with respect to those obtained using the trifocal tensor, as shown in Figure 4. We observe that at $\alpha = 0.1$, nearly 90% of the hand-labeled keypoints lie within the region of the geometrically consistent keypoints.

Accuracy Analysis: Figure 5 depicts the change in accuracy with respect to Alpha on Car-render dataset. We show four different plots with different occlusion configuration, ranging from 3 (very less occluded) to 9 (highly occluded) invisible points out of 12 keypoints in total. We observe that our method outperforms the baseline method in all configurations for occluded keypoints. At $\alpha = 0.1$ we observe a boost of 22% for 3 invisible points and 10% for 9 invisible points. Figure 6 shows the change in accuracy with respect to number of occlusions for Car-render dataset. We plot the graph for two different value of $\alpha$ and observe that 2D graph method is more stable with increasing occlusion compared to the 3D-KGNN. We show similar accuracy vs. alpha plots on CarFusion dataset in Figure 8. We observe that with increasing occlusions our method shows higher accuracy improvement compared to the baseline MaskRCNN. At $\alpha = 0.1$ we nearly gain a boost of at least 6% in all the occlusion categories and nearly 12% boost for 5 occluded points. Figure 9 depicts the change in accuracy with increasing number of occluded points on CarFusion dataset. For the case of 4 invisible points configuration, our approach is nearly 25% higher compared to the baseline. To conclude we observe that the accuracy of KGNN on occluded points is higher than using the baseline method.

Robustness Analysis: We analyze the effect of adding error to input locations of the graph to analyze the robustness of the learned model. Figure 10 shows the accuracy with respect to different Gaussian error added to the input graph. We observe that 3D-KGNN is more stable with increasing error while 2D-KGNN performs well for highly occluded points but falls steeply with increasing error in input.

4.3. Qualitative Evaluation

In this section, we analyze the visual improvements of our method across different categories of occlusion. Figure 11 depicts the visual results of the algorithm in different occlusion situations. We demonstrate results on four occlusion types namely, self-occlusion, vehicle occluding car, other objects occluding car, and truncation where the car is partially visible. The first column depicts the output from the MaskRCNN keypoints. The color is coded blue because the output from heatmaps does not give statistics about the occlusion categories of the keypoints. The other column shows ablation results on our approach. The results demonstrate
that predicting occluded keypoints as a heatmap generate large errors in localization while learning a graph based latent space improves the location of the occluded keypoints with respect to the visible points. Specifically, in high occlusion scenarios, graph-based methods show large improvement visually compared to MaskRCNN. We further show the results of our method on multiple cars simultaneously in Figure 7. Our method performs accurate occluded keypoint localization on very challenging occluded cars.

5. Conclusion

We presented a novel graph based architecture to predict the 2D and 3D locations of occluded keypoints. Since supervision for 2D occluded keypoints is challenging, we computed the error using labeled visible keypoints from different views. We proposed a self-supervised network to lift the 3D structure of the keypoints from the 2D keypoints. We demonstrated our approach on synthetic CAD data as well as a large image set capturing vehicles at many busy city intersections and improve localization accuracy (about 10%) with respect to the baseline detection algorithm.

Acknowledgements

This work was funded in parts by Heinz Endowments, US DOT RITA (University Transportation Center and Mobility 21 Center), NSF #CNS-1446601 and DARPA REVEAL Phase 2 contract HR0011-16-C-0025.
Figure 11: Qualitative evaluation of the 2D/3D keypoint localization for different occlusion categories of cars from the CarFusion dataset. The initial detector was trained using the MaskRCNN on the visible 2D keypoints. We use our self-supervised 2D-KGNN and 3D-GNN to localize keypoints from a single view. 2D reprojections of the 3D keypoints are shown in third column. The second and third columns show clear improvement in the localization of the occluded keypoints with respect to the baseline MaskRCNN. The canonical 3D views computed using 3D-KGNN are shown in the last column. The ground truth is obtained by applying trifocal tensor on the human labeled visible points to estimate the invisible points. Green represents visible edges and red represents occluded edges.
References


