An Appearance-and-Structure Fusion Network for Object Viewpoint Estimation

Yueying Kao[1], **Weiming Li**[1], **Zairan Wang**[1], **Dongqing Zou**[1], **Ran He**[2], **Qiang Wang**[1], **Minsu Ahn**[3] and **Sunghoon Hong**[3]
[1] SAIT - China Lab, Samsung Research Institute China - Beijing (SRC-B)
[2] NLPR & CRIPAC, Institute of Automation, Chinese Academy of Sciences
[3] Samsung Advanced Institute of Technology (SAIT)
{yueying.kao, weiming.li, zairan.wang, dongqing.zou, qiang.w, minsu.ahn, ar.sung.hong}@samsung.com, rhe@nlpr.ia.ac.cn

Abstract

Automatic object viewpoint estimation from a single image is an important but challenging problem in machine intelligence community. Although impressive performance has been achieved, current state-of-the-art methods still have difficulty to deal with the visual ambiguity and structure ambiguity in real world images. To tackle these problems, a novel Appearance-and-Structure Fusion network, which we call it ASFnet that estimates viewpoint by fusing both appearance and structure information, is proposed in this paper. The structure information is encoded by precise semantic keypoints and can help address the visual ambiguity. Meanwhile, distinguishable appearance features contribute to overcoming the structure ambiguity. Our ASFnet integrates an appearance path and a structure path to an end-to-end network and allows deep features effectively share supervision from both the two complementary aspects. A convolutional layer is learned to fuse the two path results adaptively. To balance the influence from the two supervision sources, a piecewise loss weight strategy is employed during training. Experimentally, our proposed network outperforms state-of-the-art methods on a public PASCAL 3D+ dataset, which verifies the effectiveness of our method and further corroborates the above proposition.

1 Introduction

Object viewpoint estimation aims at predicting 3D pose (three angles: azimuth, elevation, and in-plane rotation) of an object from a single image. It is important for many machine intelligence applications, such as robotics, augmented reality, surveillance, autonomous driving, and manipulation. Although some existing 3D pose or viewpoint estimation approaches [Su *et al.*, 2015; Tulsiani and Malik, 2015; Zhang *et al.*, 2013; Yang *et al.*, 2014; Elhoseiny *et al.*, 2016; Mahendran *et al.*, 2017; Pavlakos *et al.*, 2017; Wu *et al.*, 2016; Szeto and Corso, 2017; Tekin *et al.*, 2017], especially convolutional neural networks (CNN) based methods have achieved remarkable success, it is still a challenging problem when applying these approaches to various real-word images.

Figure 1: Appearance-based method vs. keypoint-based method. The bars under each image show 360-class confidence of the azimuth angle by different methods and ground truth (GT) respectively, where bright color indicates high confidence. In (a), the appearance based method is confused between its two confidence peaks for the bus front view (about 35 degrees) and bus rear view (about 215 degrees), which have similar appearance. In (b), the car's right-front wheel is occluded and the keypoint based method predicts its position to be the same as the car's right-back wheel by mistake, which leads to wrong viewpoint confidence. By fusing both appearance and keypoint information, our method eliminates the confusion in such cases and yield correct viewpoint estimation near GT.

Most existing viewpoint estimation methods from single images can be generally divided into two categories: appearance based methods and structure or keypoint based methods. However, these two kinds of methods often have difficulty with the visual ambiguity or structure ambiguity. The appearance based methods [Su *et al.*, 2015; Tulsiani and Malik, 2015; Elhoseiny *et al.*, 2016] only rely on appearance information to estimate viewpoint. As a result, these methods sometimes confuse to the similar appearances under different viewpoints. For example, the ambiguous appearance between the front view and rear view of a bus leads to an incorrect estimation result, as shown in Figure 1(a). If coupled with semantic keypoint locations, an accurate viewpoint estimation can be obtained. Instead of using appearance features, local keypoint locations or heatmaps are applied by keypoint based methods [Pavlakos *et al.*, 2017] to estimate viewpoint. However, the keypoint location may present

structure ambiguity in the scenarios of occlusion, truncation or symmetry. For example, the right-back wheel and right-front wheel of a car are predicted in the same location due to the occlusion, which results in viewpoint prediction errors as shown in Figure 1(b). Obviously, the appearance information will contribute to correcting the predicted viewpoint error in this case. The visual feature and structure feature are complementary to each other and these two features can be fused to resolve the issues above.

With these considerations, we aim at leveraging both the appearance information and structure information to improve the viewpoint estimation. Toward this end, an Appearance-and-Structure Fusion network (ASFnet) is proposed as shown in Figure 2. In our network, two different parallel paths are designed to make full use of object appearance information and structure information. One path extracts appearance features by supervision of viewpoint labels (appearance based viewpoint path). The other one uses the structure features by supervision of keypoint labels and then estimates viewpoint from the resultant keypoint heatmaps (structure based viewpoint path). Specifically, as shown in Figure 2, we use the VGG net [Simonyan and Zisserman, 2014] to extract appearance features and generate keypoint heatmaps by integrating the architecture of Feature Pyramid Networks (FPN) [Lin *et al.*, 2017] for its powerful ability of feature extraction at different scales. Then the viewpoint confidences from the two paths are adaptively fused with a convolutional fusion layer to infer the final viewpoint confidences. To balance the supervision effects of viewpoint and keypoint in our network, an adaptive piecewise loss weight strategy is proposed for training the keypoint prediction task.

The proposed method is evaluated on a challenging public PASCAL 3D+ [Xiang *et al.*, 2014] dataset. The experimental results demonstrate that the fusion of appearance and structure information is effective for viewpoint estimation and our method outperforms the state-of-the-art methods [Tulsiani and Malik, 2015; Su *et al.*, 2015; Wu *et al.*, 2016], especially using less training data.

In summary, our contributions are as follows:

1) For object viewpoint estimation from single images, we propose an Appearance-and-Structure Fusion network (ASFnet) with the supervision of viewpoint and keypoint simultaneously. This network is composed of two different parallel paths that can make full use of rich object appearance information and structure information provided by keypoint.

2) An adaptive learning technique is proposed to fuse the two paths with a convolutional layer together with the architecture of Feature Pyramid Networks (FPN) for keypoint prediction. In addition, we present a simple piecewise loss weight strategy for keypoint prediction loss function when training the network.

3) We evaluate our method on the public PASCAL 3D+ dataset, and achieves better performance than state-of-the-art methods in 3D viewpoint estimation.

2 Related Work

This section reviews methods for the two tasks related to our methods: keypoint estimation and viewpoint estimation.

Keypoint Estimation. Keypoint estimation methods for humans are studied widely and have achieved great success with CNN based approaches, such as human pose estimation [Newell *et al.*, 2016; Chen *et al.*, 2017] and face landmark detection [Zhang *et al.*, 2014]. More recently, CNN is also leveraged to predict keypoints of objects [Tulsiani and Malik, 2015; Pavlakos *et al.*, 2017; Wu *et al.*, 2016; Li *et al.*, 2017]. For generic object keypoint prediction, [Tulsiani and Malik, 2015] proposes a fully convolutional network and [Pavlakos *et al.*, 2017] adopts a stacked hourglass architecture. [Wu *et al.*, 2016] and [Li *et al.*, 2017] design their networks to estimate keypoints limited to a specific object category. Furthermore, [Tulsiani and Malik, 2015; Li *et al.*, 2017] also utilize the viewpoint to improve the keypoint prediction. We adapt the architecture of Feature Pyramid Networks (FPN) [Lin *et al.*, 2017] in our network to predict keypoints of generic objects. FPN is powerful for feature extraction at different scales and has achieved significant improvement for object detection, segmentation [Lin *et al.*, 2017] and human pose estimation [Chen *et al.*, 2017] recently.

Viewpoint Estimation. Most viewpoint estimation methods with only single images can be divided into two categories: appearance based methods and keypoint (or structure) based methods. Appearance based methods directly estimate viewpoints of objects from RGB images. In early works [Xiang *et al.*, 2014; Pepik *et al.*, 2012], Deformable Part Models (DPM) [Felzenszwalb *et al.*, 2010] are extended to perform object detection and their viewpoint estimation. Later, CNN based approaches [Su *et al.*, 2015; Tulsiani and Malik, 2015; Elhoseiny *et al.*, 2016; Mahendran *et al.*, 2017; Poirson *et al.*, 2016] are proposed for fine-grained viewpoint estimation and they achieve remarkable progress. Most of these methods focus on augmenting training dataset to improve the performance. For example, in [Tulsiani and Malik, 2015] the training data is augmented with jittered GT bounding boxes. [Su *et al.*, 2015] augments real images by synthesizing millions of high diverse images with 3D models.

Since viewpoint can be calculated by keypoints, keypoint based methods leverage keypoints detected from single images to estimate viewpoint of objects [Pavlakos *et al.*, 2017; Wu *et al.*, 2016]. [Pavlakos *et al.*, 2017] and [Wu *et al.*, 2016] propose to learn or design the projection method from detected keypoint locations or heatmaps to viewpoint. However, detected keypoint locations or heatmaps lack rich object appearance information, and these methods heavily rely on the predicted keypoint quality.

[Szeto and Corso, 2017] proposes a human-in-the-loop network to assist a viewpoint estimation CNN by allowing using a human labeled keypoint at inference time. The network has two streams and takes an image and a keypoint as inputs respectively. Then the CNN features and keypoint generated features in the two streams are directly concatenated for viewpoint estimation. In their method, the keypoint location is assumed to be accurate and must be provided by human at inference time. However, in many real-world scenarios, such human annotations are not available. One may consider using automatic keypoint extractor to replace human efforts. The keypoint prediction on generic

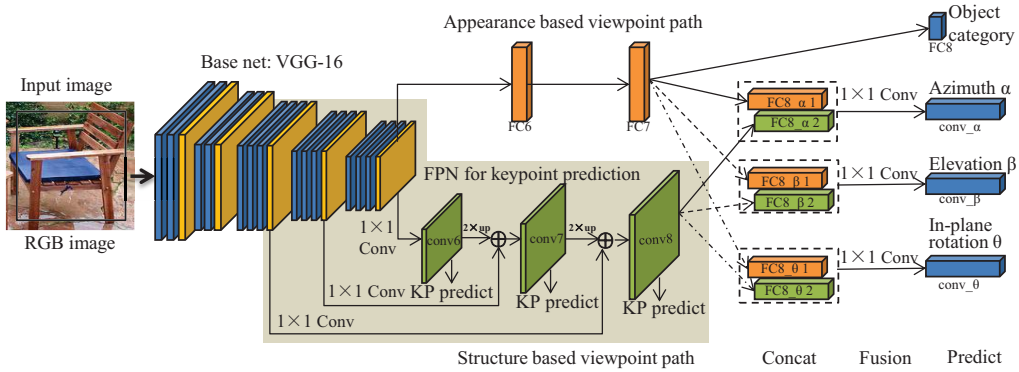

Figure 2: The architecture of our proposed network. (Conv: convolutional layer, FC: fully connected layer, KP: keypoint.)

objects is still a challenging problem [Pavlakos *et al.*, 2017] and cannot be assumed as available by default.

In our paper, we focus on automatic object viewpoint estimation from single images without human interaction. There are several differences between our method and previous works. Compared with appearance based and keypoint based methods, we use both of the appearance and keypoint (structure) information with two parallel paths inside one network. Furthermore, our architectures for keypoint information extraction are different. Compared with [Szeto and Corso, 2017], we simultaneously use the keypoint and viewpoint annotations to supervise the network at training stage. At test time, only an RGB image is used without need of keypoint annotations. In addition, we develop a convolutional fusion layer to automatically learn the fusion of the two kinds of information. Finally, we do not rely on the large amount of augmented training data used in previous methods [Tulsiani and Malik, 2015; Su *et al.*, 2015; Wu *et al.*, 2016; Szeto and Corso, 2017; Pavlakos *et al.*, 2017].

3 Our proposed ASFnet

In this section, we propose an end-to-end network to fuse appearance information and structure information for object viewpoint estimation from single RGB images. Figure 2 illustrates our approach. We begin with the formulation for this problem (Sec. 3.1) and then give the details of our proposed network (Sec. 3.2).

3.1 Problem Formulation

For a single RGB image x as input, our goal is to infer three angles (azimuth α, elevation β, and in-plane rotation θ), the relative viewpoint $V = \{\alpha, \beta, \theta\}$ of an object with respect to a camera. We formulate the viewpoint estimation problem as a fine-grained classification problem, by dividing each angle into N_v bins ($N_v = 360$, similar to previous works [Su *et al.*, 2015; Szeto and Corso, 2017]) respectively.

Viewpoint can be estimated directly from the appearance features in images [Su *et al.*, 2015; Tulsiani and Malik, 2015], and can also be calculated from the predicted keypoints [Pavlakos *et al.*, 2017; Wu *et al.*, 2016]. However, the appearance information lacks the structure information from predicted keypoint. The predicted keypoint information loses the rich appearance information. In this paper, to make full use of the appearance information and structure information, we propose an Appearance-and-Structure Fusion convolutional neural network (ASFnet) with supervision of viewpoint and keypoint labels simultaneously. This can be interpreted as a probabilistic problem. Specifically, the goal is to learn the network parameters W by maximizing the following posterior probability,

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, p(W|x, y), \tag{1}$$

where y is supervisory signals, including object class c, viewpoint labels V and keypoint labels K, $y = \{c, V, K\}$. Due to different appearance of different object categories in the same viewpoint, object classification is included into the network to learn discriminative representation for viewpoint estimation.

3.2 Network Architecture

The architecture of our proposed ASFnet is shown in Figure 2. The base net we use is VGG network [Simonyan and Zisserman, 2014]. Then two parallel paths, appearance based viewpoint path and structure based viewpoint path, are connected to the base net. Each path estimates a confidence vector for each angle. Finally, two confidence vectors are adaptively fused for final estimation. We will introduce the details of our network in the following parts: appearance based viewpoint path, structure based viewpoint path, adaptive fusion learning and objective function.

Appearance based Viewpoint Path. Appearance features from images are useful for viewpoint estimation, which has been shown by some existing state-of-the-art methods [Su *et al.*, 2015; Tulsiani and Malik, 2015]. We set an path in the network to extract appearance representation. The path is from the base net to three fully connected (FC) layers, FC6, FC7, and FC8_x1 for each angle of viewpoint. FC6 and FC7 is the same as those of VGG network. FC8_x1 corresponds to FC8_α1, FC8_β1, FC8_θ1. Then each FC8_x1 and its corresponding GT label are feed into the loss function for viewpoint estimation when training. Here we use the geometric structure aware loss function from [Su *et al.*, 2015], which is an improved softmax loss function. Since the base net is supervised with both of viewpoint and keypoint labels,

both appearance and structure information are learned in the layers of the base net, FC6 and FC7. In contrast with the path of structure based viewpoint path, we call this path appearance based viewpoint path.

Structure based Viewpoint Path. Viewpoint can be calculated from predicted keypoint loctaions [Pavlakos *et al.*, 2017]. We aim to utilize structure information of keypoints to improve the appearance based viewpoint estimation. Furthermore, the supervision of viewpoint labels on base net can also reduce structure ambiguity in keypoint prediction. Here the architecture of FPN [Lin *et al.*, 2017] is introduced into the network for keypoint prediction. FPN uses a bottom-up pathway and a top-down pathway with lateral connections to construct feature pyramids as powerful feature extraction at different scales. In our network, the path from pool3, pool4 to pool5 layers corresponds to the bottom-up pathway. The path from conv6, conv7, to conv8 corresponds to the top-down pathway. To predict keypoint localization and class, we set the channels of conv6, conv7, conv8 to be equal to the total number of the keypoints of all object classes N_k. $N_k = \sum_{c=1}^{C} |K_c|$, here K_c denotes the set of keypoints of object class c, $|K_c|$ denotes the number of keypoints of class c. The size (width and height) of conv6, conv7 and conv8 is the same as that of layers pool5, pool4 and pool3 respectively. Each feature map of conv6, conv7 and conv8 corresponds to a probability distribution for each keypoint of a specific object class. Thus the keypoint prediction is interpreted as a multi-label classification problem. For network training, we construct a GT label map for each layer (conv6,7,8). Its size is the same as its corresponding layer. The value is set to 1 in the location (h, w) of keypoints in each feature map and others to 0. When training, conv6, 7, 8 and their corresponding GT label maps are feed into each sigmoid-cross-entropy loss function respectively. Then the heatmaps of conv8 are used to predict viewpoints and the conv8 is feeded into a FC layer FC8_x2 for each angle. FC8_x2 refers to FC8_α2, FC8_β2, FC8_θ2. When training, FC8_x2 and its GT label are feed into a geometric structure aware loss function for each angle estimation. This path uses the heatmaps of keypoints. We call it structure based viewpoint path.

Adaptive Fusion Learning. To fuse the predicted viewpoint confidence vectors FC8_x1 and FC8_x2 in the above paths, we utilize an adaptive learning technique for fusion. Traditional fusion methods, such as averaging, maximizing, are hard to design and may not extract appropriate information from different paths. A learning technique can learn the fusion parameters adaptively and is suitable for the network. Specifically, for each angle in our network, two confidence vectors FC8_x1 and FC8_x2 are concatenated as a two-channel confidence map, which then undergoes a 1×1 convolutional layer to generate FC8_x. FC8_x corresponds to the final predicted viewpoint FC8_α, FC8_β, FC8_θ. When training, FC8_x and its GT label are taken as inputs of a geometric structure aware loss function for each angle.

Objective Function. Let $L_v(x, V)$ denote a viewpoint loss function, including three geometric structure aware loss functions for three angles, $L_a(x, \alpha)$ for azimuth, $L_e(x, \beta)$ for elevation, and $L_r(x, \theta)$ for in-plane rotation. $L_v(x, V) = L_a(x, \alpha) + L_e(x, \beta) + L_r(x, \theta)$. Let $L_k(x, K)$ denote a sigmoid cross entropy loss function for keypoint prediction. Let $L_c(x, c)$ denote a softmax loss function for object classification. In our ASFnet, there are three viewpoint loss functions $L_v(x, V)$, three keypoint loss functions $L_k(x, K)$ and an object classification loss functions $L_c(x, c)$. We set $L_{vf}(x, V)$ as the final viewpoint estimation loss, set $L_{va}(x, V)$ as the appearance based viewpoint estimation loss, set $L_{vs}(x, V)$ as the structure based viewpoint estimation loss, set $L_{k6}(x, K)$, $L_{k7}(x, K)$ and $L_{k8}(x, K)$ as the keypoint estimation followed by layers conv6, conv7 and conv8 respectively. The final objective function is

$$L(x, y) = \sum_{u \in U} L_{vs}(x, V) + \sum_{i \in I} f(L_{ki}) L_{ki}(x, K) + L_c(x, c), \quad (2)$$

here $U = \{f, a, s\}$, $I = \{6, 7, 8\}$. $f(L_{ki})$ is our proposed loss weight function.

To keep and balance the effects of all tasks in our network in entire training procedure, we propose a piecewise function $f(l)$ to adaptively provide the weight of each keypoint estimation loss function. We find that softmax loss function only considers the value corresponding GT label for each training example, the geometric structure aware loss function considers some values around GT label, while sigmoid cross entropy loss function considers all the values in all the locations of all feature maps. The loss values of different tasks are very different in the training. Thus different loss functions need different loss weights in the network training. In addition, the speed of loss value decent of different tasks through the entire training procedure is also very different. For example, loss of keypoint prediction falls faster than the loss of viewpoint estimation task at the beginning of training. When the network converges, the losses of the two tasks are falling at similar rate. Thus the loss weights for different loss functions need to be adjusted in the entire training procedure to keep the effects of all tasks. To implement it, we propose a simple piecewise function $f(l)$ for the keypoint loss function.

$$f(l) = \begin{cases} a/L_2, & l \geq L_2, \\ a/L_1, & L_1 \leq l < L_2, \\ a, & 0 < l < L_1, \end{cases} \quad (3)$$

here l corresponds to the loss value of keypoint estimation, L_1 and L_2 are two threshold values, a is a constant. In this method, when the loss is very large, the loss weight is very small. When the loss converges, the loss weight becomes larger. In this way, it makes all the tasks affect the network effectively through entire training procedure until the network converges.

4 Experiments

In this section, we present the experimental setup, quantitative and qualitative results on viewpoint estimation.

Dataset and Metrics. Our method is evaluated on 12 object categories of a public PASCAL 3D+ [Xiang *et al.*, 2014] dataset. There are annotations of viewpoints, keypoints, object classes and object bounding boxes in this dataset. This dataset consists of PASCAL VOC 2012 detection train and validation set, and ImageNet images. We use the PASCAL

	aero	bike	boat	bottle	bus	car	chair	table	mbike	sofa	train	tv	mean
$Acc_{\pi/6}$ (Anet)	81.9	83.9	52.9	94.4	92.9	87.1	73.7	60.9	87.6	87.5	79.6	85.6	80.7
$Acc_{\pi/6}$ (ASnet)	85.9	85.6	61.9	94.8	97.4	87.7	78.9	73.9	**90.5**	85.0	79.6	85.1	83.9
$Acc_{\pi/6}$ (ASFnet with average fusion)	82.2	85.6	**62.3**	94.0	97.4	87.1	76.1	78.3	87.6	82.5	82.3	88.7	83.7
$Acc_{\pi/6}$ (ASFnet with max fusion)	85.9	81.4	52.0	**94.8**	96.1	88.1	72.5	73.9	88.3	80.0	81.4	88.3	81.9
$Acc_{\pi/6}$ (ASFnet with adaptive fusion)	**86.6**	**88.1**	58.6	93.3	**98.7**	86.5	78.5	**82.6**	89.8	85.0	84.1	90.1	**85.2**
$Acc_{\pi/6}$ ([Tulsiani and Malik, 2015])	81	77	59	93	98	89	80	62	88	82	80	80	81
$Acc_{\pi/6}$ ([Su *et al.*, 2015])	74	83	52	91	91	88	**86**	73	78	**90**	**86**	**92**	82
$Acc_{\pi/6}$ ([Szeto and Corso, 2017])*	-	-	-	-	96.8	**90.2**	-	-	85.2	-	-	-	
$MedErr$ (Anet)	10.0	12.1	26.6	6.1	2.3	4.1	9.6	23.3	10.0	8.1	4.8	9.5	10.5
$MedErr$ (ASnet)	8.0	10.8	18.6	5.9	2.0	**4.0**	8.4	10.0	9.7	**7.7**	**3.7**	10.4	8.3
$MedErr$ (ASFnet with average fusion)	8.3	**10.7**	20.2	**5.7**	1.9	4.3	8.5	**7.0**	9.8	7.8	4.0	11.0	8.2
$MedErr$ (ASFnet with max fusion)	7.8	12.1	26.8	6.4	2.1	4.1	9.7	8.8	10.4	9.6	4.5	**9.3**	9.3
$MedErr$ (ASFnet with adaptive fusion)	**7.4**	**10.7**	**18.5**	6.1	**1.8**	**4.0**	**8.2**	7.5	**9.0**	8.1	**3.7**	9.7	**7.9**
$MedErr$ ([Tulsiani and Malik, 2015])	13.8	17.7	21.3	12.9	5.8	9.1	14.8	15.2	14.7	13.7	8.7	15.4	13.6
$MedErr$ ([Su *et al.*, 2015])	15.4	14.8	25.6	9.3	3.6	6.0	9.7	10.8	16.7	9.5	6.1	12.6	11.7
$MedErr$ ([Pavlakos *et al.*, 2017])	8.0	13.4	40.7	11.7	2.0	5.5	10.4	-	-	9.6	8.3	32.9	
$MedErr$ ([Mahendran *et al.*, 2017])	13.97	21.07	35.52	8.99	4.08	7.56	21.18	17.74	17.87	12.70	8.22	15.68	15.38
$MedErr$ ([Szeto and Corso, 2017])*	-	-	-	-	2.64	4.98	-	-	11.4	-	-	-	

*This work takes an RGB image and a human localized keypoint as inputs to estimate viewpoint in the test time.

Table 1: $Acc_{\pi/6}$ (%) and $MedErr$ of different methods for viewpoint estimation with ground truth bounding boxes on PASCAL 3D+ dataset.

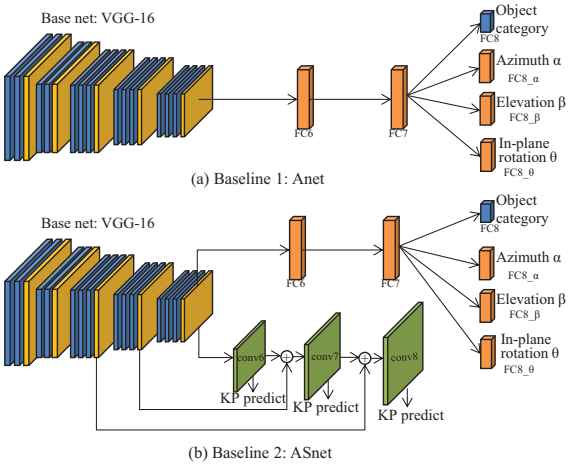

Figure 3: Two baselines for our proposed network.

train set and ImageNet images with GT bounding boxes to train our network. The whole PASCAL validation set is used to evaluate our performance.

To be consistent with previous works [Xiang *et al.*, 2014; Su *et al.*, 2015; Tulsiani and Malik, 2015], we use (Average Viewpoint Precision) AVP, $Acc_{\pi/6}$ and $MedErr$ as the evaluation metrics. AVP [Xiang *et al.*, 2014] is used to evaluate the performance of methods for joint detection and viewpoint estimation. When computing AVP, the result is correct only if both of detection result and viewpoint (only azimuth is considered following previous works [Tulsiani and Malik, 2015; Su *et al.*, 2015; Xiang *et al.*, 2014; Pepik *et al.*, 2012]) are correct. $Acc_{\pi/6}$ (accuracy at $\Delta(R_1, R_2) < \frac{\pi}{6}$) and $MedErr$ are based on the geodesic distance $\Delta(R_1, R_2) = ||log(R_1^T R_2)||_F/\sqrt{2}$ over the manifold of rotation matrix between GT and predicted viewpoint (three angles are all considered). These two metrics [Tulsiani and Malik, 2015] are presented to evaluate the viewpoint estimation performance with GT bounding boxes.

Baselines. We design two baseline networks (shown in Figure 3) to evaluate the effectiveness of our ASFnet. (1) **Baseline 1**: An appearance-only viewpoint estimation network is implemented. We call it Anet, which is similar to the state-of-the-art methods [Su *et al.*, 2015; Tulsiani and Malik, 2015]. Its architecture is same as the appearance based viewpoint path in our ASFnet. (2) **Baseline 2**: Compared to baseline 1, an Anet with added supervision of keypoint is implemented. We call it ASnet. The added architecture of FPN for keypoint prediction is same as the corresponding layers of our ASFnet. Compared with ASFnet, ASnet does not have the fusion part of the ASFnet.

When training the baseline networks and ASFnet, we resize the object images cropped from training set with their GT bounding boxes to $256 \times 256 \times 3$, then randomly extract a $224 \times 224 \times 3$ patch from the resized image or resized mirror image as the input of all networks. In addition to these traditional operations, previous works also augment training data with jittered GT bounding boxes that overlap with annotated bounding box with IoU > 0.7 [Tulsiani and Malik, 2015], or by synthesizing millions of high diverse images with 3D CAD models [Su *et al.*, 2015; Wu *et al.*, 2016; Szeto and Corso, 2017]. All the baseline networks and ASFnet here are only trained with the traditional operations as mentioned above. In the piecewise loss weight function $f(l)$, a is set to 0.5, L_1 to 100, L_2 to 1000 by experience. We initialize the two baseline networks with the trained VGG network on ImageNet classification task. Then the ASnet is used to initialize our ASFnet.

Effect of Structure Information from Keypoints. To evaluate the effectiveness of structure information provided by keypoints for viewpoint estimation quantitatively, we show the $Acc_{\pi/6}$ and $MedErr$ of Anet, ASnet and ASFnet with GT bounding boxes in Table 1. As shown in Table 1, our ASFnet performs best, and ASnet performs better than Anet. By comparing the performance of ASnet and Anet, we can see that the supervision of keypoint is very useful to learn auxiliary structure representations in the base net for viewpoint estimation. By comparing the results of ASFnet and ASnet, we can see that our adaptive fusion for

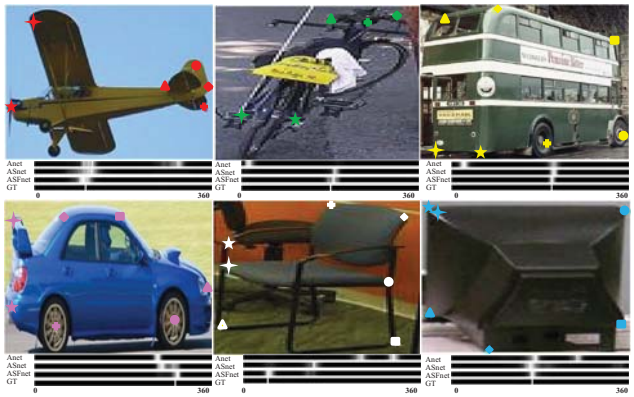

Figure 4: Qualitative comparison results. Azimuth confidences with different methods and ground truth (GT). Our predicted keypoint locations are also shown in the images with different markers and different colors for different object categories.

using the structure information is effective to further improve viewpoint estimation.

To qualitatively demonstrate the benefits of structure information for viewpoint estimation, we show some example images with predicted viewpoints in Figure 4. It shows that the appearance only method Anet sometime has two peaks in the viewpoint confidence due to visual ambiguity and is confused. The keypoint information of ASnet and ASFnet can reduce the ambiguity effectively.

Effect of Adaptive Fusion. Traditional fusion techniques often use averaging, maximizing, etc. To verify the effect of our adaptive fusion method, we replace the 1×1 convolutional layer in ASFnet with averaging or maximizing the values from two paths. We call these fusion networks as ASFnet with average fusion and ASFnet with max fusion. As shown in Table 1, traditional fusion methods contribute little for improvement. Our adaptive fusion performs much better.

Effect of Piecewise Loss Weight. We propose a piecewise loss weight strategy for keypoint prediction loss function when training our ASnet and ASFnet. To demonstrate the effect of the piecewise loss weight, we implement an ASnet with a fixed loss weight for each keypoint loss function when training. We choose three best loss weights for L_{k6}, L_{k7}, L_{k8} by experiments. The training and testing errors of the two ASnets with different loss weight strategies for viewpoint estimation and keypoint prediction respectively are shown in Figure 5. We can see that piecewise loss weight is beneficial to both of the two tasks by keeping and balancing the effects of all tasks. It makes the network learn more powerful features for viewpoint estimation.

Comparison to State-of-the-art Methods. We compare with the viewpoint estimation methods that use only a single image as input. In all the comparisons, our training data augmentation does not use the CAD synthesized images [Su *et al.*, 2015; Wu *et al.*, 2016; Szeto and Corso, 2017; Pavlakos *et al.*, 2017] nor the jittered bounding boxes [Tulsiani and Malik, 2015]. Note that results of some compared methods are only available with one or two metrics on some object categories. Firstly, we evaluate the performance of different methods with GT bounding boxes in Table 1. Our network

Methods	mAVP				AVP (4V)	
	4V	8V	16V	24V	chair	sofa
[Xiang *et al.*, 2014]	19.5	18.7	15.6	12.1	6.8	5.1
[Pepik *et al.*, 2012]	23.8	21.5	17.3	13.6	6.1	11.8
[Poirson *et al.*, 2016]	50.7	45.1	33.6	28.8	11.3	40.6
[Wu *et al.*, 2016]	-	-	-	-	23.1	45.8
[Su *et al.*, 2015]	39.7	32.9	24.2	19.8	15.7	28.4
Ours+detection-1	46.5	41.0	32.1	27.4	17.3	34.7
[Tulsiani and Malik, 2015]	49.1	44.5	36.0	31.1	25.1	43.8
Ours+detection-2	**52.1**	**45.5**	**37.2**	**31.4**	**26.4**	**54.2**

Table 2: Joint object detection and viewpoint estimation. We show mean AVPs (mAVP) of different methods on 12 categories of the PASCAL 3D+ and AVPs (4 bins) on two categories.

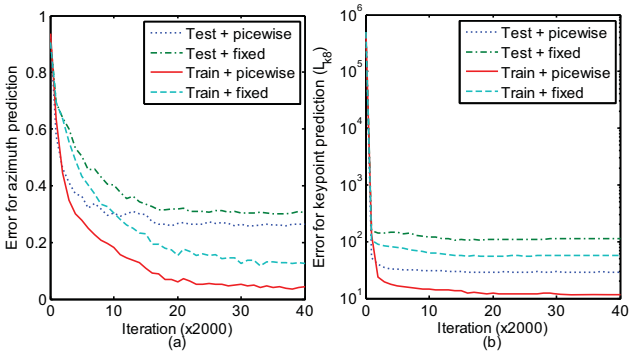

Figure 5: Fixed loss weight vs. Piecewise loss weight. (a) shows the error (1- accuracy at $\pi/6$) of azimuth angle in viewpoint estimation (b) shows the error (L2 distance between predicted and GT) for keypoint estimation. In both training and test set, network using the piecewise loss weight achieves lower error with faster convergence.

outperforms the state-of-the-art methods [Tulsiani and Malik, 2015; Su *et al.*, 2015] on the whole PASCAL 3D+ validation set. Although the work [Szeto and Corso, 2017] uses an RGB image and a human localized keypoint as inputs, we also list their results in Table 1 just for reference. Then, we evaluate our method on the joint detection and viewpoint estimation task as shown in Table 2. AVPs are shown for four cases that the 360-degree views are discretized to 4, 8, 16, 24 bins respectively as in previous works. We use the detection results, detection-1 provided by [Su *et al.*, 2015] and detection-2 provided by [Tulsiani and Malik, 2015] respectively to evaluate our ASFnet. The method [Wu *et al.*, 2016] is limited to some specific categories and is only evaluated on two categories, sofa and chair, for 4 bins. It is shown that our viewpoint estimation method (with less training data) outperforms the state-of-the-art methods [Tulsiani and Malik, 2015; Su *et al.*, 2015; Wu *et al.*, 2016; Poirson *et al.*, 2016]. These results on the three metrics demonstrate that fusing appearance and structure information leads to significant improvement for viewpoint estimation.

Compared with previous methods, our method works much better for the error cases due to the problem of visual ambiguity and structure ambiguity as shown in Figure 1 and 4. However, as for other sources of errors in existing methods [Tulsiani and Malik, 2015], such as very small objects, our method may encounter difficulty too. In these cases, it is difficult to recover useful features since most visual information are lost due to low resolution.

5 Conclusion

Current state-of-the-art methods for viewpoint estimation have achieved remarkable progress by exploring deep learning models, but they still suffer from visual ambiguity and structure ambiguity issues. This paper addresses the issue of how to effectively fuse the appearance and structure information to reduce such ambiguity for viewpoint estimation. An appearance and structure fusion network is proposed with two different parallel paths and an adaptive fusing technique. The two paths fuse supervision from viewpoint labels on appearance extraction and from keypoint labels on structure information extraction respectively. Experiments demonstrate that our fusion based method outperforms the state-of-the-art methods on the challenging Pascal3D+ dataset. We wish this work would inspire more efforts in the research community to investigate fusing relevant supervision to further improve the performance of this important task.

References

[Chen *et al.*, 2017] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. *arXiv preprint arXiv:1711.07319*, 2017.

[Elhoseiny *et al.*, 2016] Mohamed Elhoseiny, Tarek El-Gaaly, Amr Bakry, and Ahmed Elgammal. A comparative analysis and study of multiview cnn models for joint object categorization and pose estimation. In *International Conference on Machine Learning*, pages 888–897, 2016.

[Felzenszwalb *et al.*, 2010] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[Li *et al.*, 2017] Chi Li, M Zeeshan Zia, Quoc-Huy Tran, Xiang Yu, Gregory D Hager, and Manmohan Chandraker. Deep supervision with shape concepts for occlusion-aware 3d object parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 388–397, July 2017.

[Lin *et al.*, 2017] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 936 – 944, 2017.

[Mahendran *et al.*, 2017] Siddharth Mahendran, Haider Ali, and René Vidal. 3d pose regression using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 494–495, 2017.

[Newell *et al.*, 2016] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499, 2016.

[Pavlakos *et al.*, 2017] Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G Derpanis, and Kostas Daniilidis. 6-dof object pose from semantic keypoints. In *International Conference on Robotics and Automation*, pages 2011 – 2018, 2017.

[Pepik *et al.*, 2012] Bojan Pepik, Michael Stark, Peter Gehler, and Bernt Schiele. Teaching 3d geometry to deformable part models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3362–3369, 2012.

[Poirson *et al.*, 2016] Patrick Poirson, Phil Ammirato, Cheng-Yang Fu, Wei Liu, Jana Kosecka, and Alexander C Berg. Fast single shot detection and pose estimation. In *International Conference on 3D Vision*, pages 676–684, 2016.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Su *et al.*, 2015] Hao Su, Charles R. Qi, Yangyan Li, and Leonidas J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *IEEE International Conference on Computer Vision*, pages 2686–2694, 2015.

[Szeto and Corso, 2017] Ryan Szeto and Jason J. Corso. Click here: Human-localized keypoints as guidance for viewpoint estimation. In *IEEE International Conference on Computer Vision*, pages 1604–1613, 2017.

[Tekin *et al.*, 2017] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. *arXiv preprint arXiv:1711.08848*, 2017.

[Tulsiani and Malik, 2015] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1519, 2015.

[Wu *et al.*, 2016] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. Single image 3d interpreter network. In *European Conference on Computer Vision*, pages 365–382, 2016.

[Xiang *et al.*, 2014] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision*, pages 75–82, 2014.

[Yang *et al.*, 2014] Linjie Yang, Jianzhuang Liu, and Xiaoou Tang. Object detection and viewpoint estimation with auto-masking neural network. In *European conference on computer vision*, pages 441–455, 2014.

[Zhang *et al.*, 2013] Haopeng Zhang, Tarek El-Gaaly, Ahmed Elgammal, and Zhiguo Jiang. Joint object and pose recognition using homeomorphic manifold analysis. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 1012–1019, 2013.

[Zhang *et al.*, 2014] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108, 2014.