Adaptation of Synthetic Data for Coarse-to-Fine Viewpoint Refinement

Pau Panareda Busto[1,2]
pau.panareda-busto@airbus.com
Joerg Liebelt[1]
joerg.liebelt@airbus.com
Juergen Gall[2]
gall@iai.uni-bonn.de

[1] Airbus Group Innovations
Munich, Germany

[2] Computer Vision Group
Institute of Computer Science III
University of Bonn, Germany

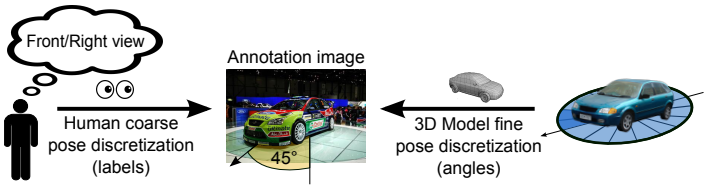

Figure 1: Humans are perfect for annotating coarse viewpoints of objects in real images, but fail to estimate pose accurately at a fine level. 3D graphic models can be used to synthesize data at very accurate fine angles, but it is time-consuming to model all appearance variations present in real images. We therefore propose to leverage the abilities of humans of estimating coarse viewpoints and the pose accuracy of synthetic data.

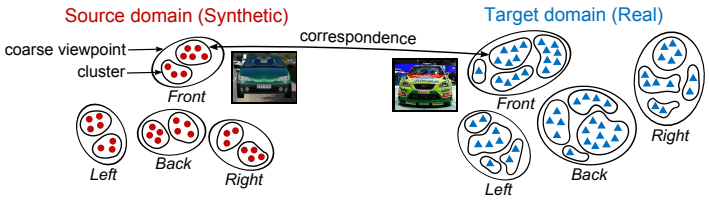

Figure 2: Synthetic data (source) is domain-adapted towards the real data (target) with a transformation estimated by minimizing the distance of correspondences between both domains. Each cluster in the target domain is assigned to a source cluster that belongs to the same coarse viewpoint. In this example, for an 8-view refinement: $V_i = 2$ and $K_i = 4$.

		8 views	16 views	32 views
	gt	*80.06*	*73.57*	*60.59*
w/o DA	syn	65.98	60.92	46.55
	real	76.04	65.46	49.90
	joint	72.52	63.81	50.04
with DA	syn	74.62	67.01	51.06
	real	**78.37**	69.04	**55.22**
	joint	75.73	**71.93**	53.00

Table 1: Pose estimation accuracy on test data using synthetic data, viewpoint refined real data or both training sets. In EPFL Dataset [4].

The quality of learning-based pose estimation still heavily relies on manual training data annotations. However, the manual labeling of large datasets is costly [6] and frequently limited to a few coarse viewpoint annotations of varying accuracy [5]. In this work, we propose to refine such coarse pose annotations with a domain adaptation approach, where the source domain consists of fine-grained pose annotations generated from synthetic computer graphics models, and the target domain of coarse manual pose annotations of a real dataset (see Figure 1), *i.e.*, we ask humans to annotate only four coarse views, namely front, back, left and right view.

Our domain adaptation technique starts by clustering the source and target domains and establishes correspondences between the clusters for each coarse viewpoint (see Figure 2). To cluster the synthetic data, we use the known fine-grained poses where each pose can be associated with one of the four coarse viewpoints $i = \{\text{front}, \text{back}, \text{left}, \text{right}\}$, *i.e.*, $V = \sum_i V_i$ where V is the total number of fine-grained poses. For the target domain, we only have the coarse viewpoints and thus we cluster the N_i training samples of one viewpoint further by K-Means where the number of clusters for each coarse viewpoint is given by K_i, *i.e.*, $K = \sum_i K_i$ and $V_i \leq K_i \leq N_i$. We represent each image by a HOG feature vector and append the aspect ratio of the bounding box surrounding the object. To this end, we represent each cluster by its centroid. The sets of centroids are denoted by $\hat{S}^i = \{\hat{s}^i_1, ..., \hat{s}^i_{V_i}\}$ and $\hat{T}^i = \{\hat{t}^i_1, ..., \hat{t}^i_{K_i}\}$. The correspondences are then established by solving a bipartite matching problem:

$$\begin{aligned} &\underset{e_{vk}}{\operatorname{argmin}} \sum_{v=1}^{V_i} \sum_{k=1}^{K_i} e_{vk} \left\| \hat{s}^i_v - \hat{t}^i_k \right\|_2^2 \\ &\text{subject to } \sum_v e_{vk} = 1 \;\; \forall k \,, \;\; \sum_k e_{vk} = a_v \;\; \forall v \;\text{ and }\; e_{vk} \in \{0,1\} \;\; \forall v,k \,. \end{aligned} \tag{1}$$

It assigns to each cluster in the target domain a unique cluster in the source domain. Since there can be more clusters in the target domain than in the source domain, each source is associated to $a_v = \frac{K_i}{V_i}$ target clusters. We use the Hungarian algorithm [3] to solve the problem.

The correspondences are then used to learn a mapping from the source domain, $\mathcal{S} \in \mathbb{R}^D$, to the target domain, $\mathcal{T} \in \mathbb{R}^D$, where D denotes the dimensionality of the features. We consider a linear transformation, which is represented by a matrix $W \in \mathbb{R}^{D \times D}$, *i.e.*, $t = Ws$. Let $S = \{s_1, ..., s_M\}$ and $T = \{t_1, ..., t_N\}$ denote the training samples of the source and target domain, respectively. M and N are the total amount of samples of each domain and we can assume that $M \geq N$ since we can always generate more synthetic data than annotated real images. Given the correspondences $C = \{c_1, ..., c_K\}$ with (s_{c_k}, t_k) and $K \leq N$, W can be learned by minimizing the objective

$$f(W) = \frac{1}{2} \sum_{k=1}^{K} ||W s_{c_k} - t_k||_2^2 \tag{2}$$

Lastly, the viewpoint refinement of the real training images is seen as a classification problem, where we train on the transformed synthetic samples a linear SVM for each of the fine viewpoints $v = \{1, ..., V\}$. Then, we apply the linear SVMs corresponding to the coarse viewpoint i of the real image and assign the fine pose with the highest scoring function:

$$f(x, i) = \underset{v=\{1,...,V_i\}}{\operatorname{argmax}} \; w_v^T x + b_v, \tag{3}$$

where w_v and b_v are the weights and bias of the linear SVM for the fine viewpoint v. The performance of our viewpoint refinement is evaluated in the paper on well-known car datasets with annotated poses, outperforming popular domain adaptation techniques (*e.g.* Geodesic flow kernel [1], Maximum margin domain transform [2]).

For further pose estimation on real test images, we also use linear SVMs in the same one-vs-all classification procedure. For each fine viewpoint, we evaluate trained SVMs using the real training images with refined pose labels, the synthetic training images, which have been transformed by domain adaptation, and both combined. As observed in Table 1, the accuracy increases in all setups when using domain adaptation.

[1] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

[2] J. Hoffman, E. Rodner, J. Donahue, B. Kulis, and K. Saenko. Asymmetric and category invariant feature transformations for domain adaptation. *IJCV*, 2014.

[3] H. W. Kuhn. The hungarian method for the assignment problem. *NRLQ*, 1955.

[4] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *CVPR*, 2009.

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

[6] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond Pascal: A benchmark for 3D object detection in the wild. In *WCACV*, 2014.