Learning Relationships for Multi-View 3D Object Recognition

Ze Yang
yangze@pku.edu.cn

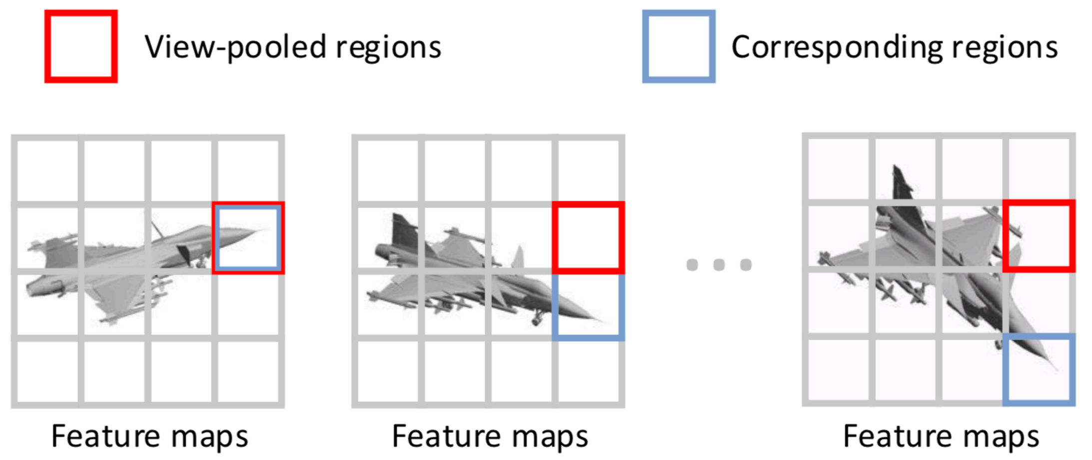
Liwei Wang
wanglw@cis.pku.edu.cn

Motivations

Given a set of view image of 3D object $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_N\}$ we want to predict the category of the object.

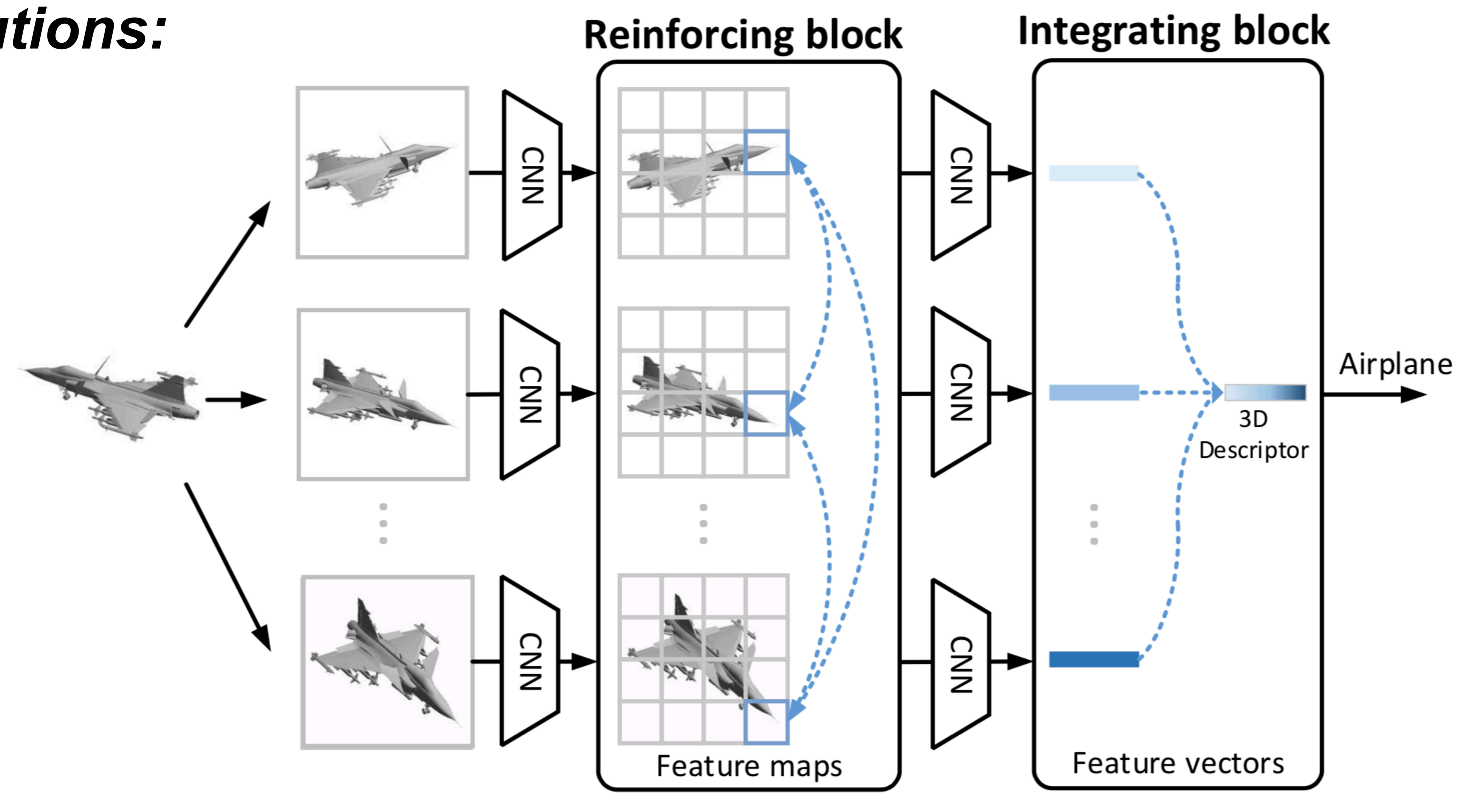
Problems:

The object usually have some regions that are occluded, reflective, incomplete or totally invisible from one particular viewpoint.
We need to **combine the information** from different viewpoints to understand the object.



A simple view pooling operation cannot aggregate the information from corresponding regions.

Solutions:



region-to-region relationships and *view-to-view* relationships

Reinforcing (region)

The region features of the *i-th* view:

$$\mathbf{R}_i = \{\mathbf{r}_{i1}, \mathbf{r}_{i2}, \cdots, \mathbf{r}_{iL^2}\},\ \mathbf{r}_{ij} \in \mathbb{R}^{D_r}$$

Matching score between $\mathbf{r}_{ij}$ and $\mathbf{r}_{mn}$:

$$M_{ij,mn} = \mathcal{M}(\mathbf{r}_{ij}, \mathbf{r}_{mn})$$

Normalize the ***matching score***:

- across view

$$\hat{M}_{ij,mn} = \frac{e^{\frac{M_{ij,mn}}{\sqrt{D_e}}}}{\sum_{m=1}^{N}\sum_{n=1}^{L^2} e^{\frac{M_{ij,mn}}{\sqrt{D_e}}}}$$

- inside view

$$\hat{M}_{ij,mn} = \frac{e^{\frac{M_{ij,mn}}{\sqrt{D_e}}}}{N \cdot \sum_{n=1}^{L^2} e^{\frac{M_{ij,mn}}{\sqrt{D_e}}}}$$

The reinforced region feature $\mathbf{r}^*_{ij}$ are calculated based on the matching score:

$$\mathbf{r}^*_{ij} = \mathbf{r}_{ij} + f\left(\sum_{m=1}^{N}\sum_{n=1}^{L^2} \hat{M}_{ij,mn} \cdot g(\mathbf{r}_{mn})\right)$$

Integrating (view)

Model the pair-wise relationships between views to determine the importance score of each view:

$$I_i = \sum_{j=1}^{N} \mathcal{R}(\mathbf{f}^*_i, \mathbf{f}^*_j)$$

Normalizing the importance score using ReLU (can be seen as first order approximation of SoftMax) to stabilize training.

$$\hat{I}_i = \frac{\mathrm{ReLU}(I_i)}{\sum_{j=1}^{N} \mathrm{ReLU}(I_j)}$$

Then the 3D object feature is calculated as the convex combination of the view feature.

$$\mathbf{f} = \sum_{i=1}^{N} \hat{I}_i \cdot \mathbf{f}^*_i$$

Experiments

Quantitative results:

Methods	Number of Views		
	3 views	6 views	12 views
MVCNN [39]	91.3	92.0	91.5
RCPCNN [43]	92.1	92.2	92.2
MHBN [47]	**93.8**	**94.1**	93.4
Ours	93.5	**94.1**	**94.3**

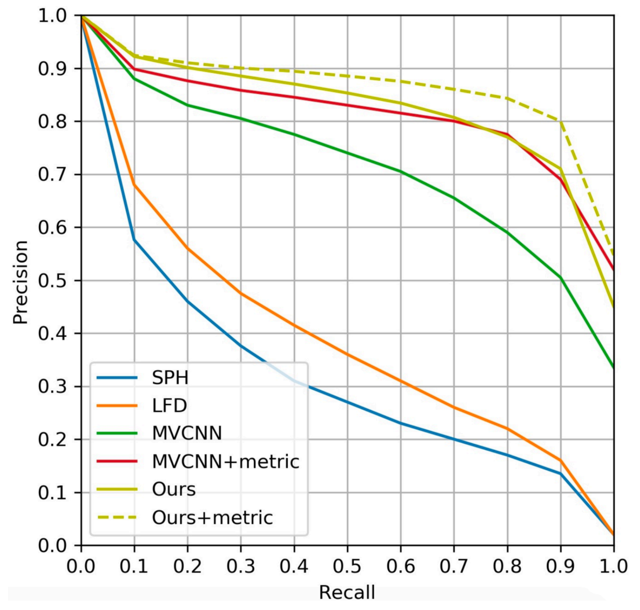
Ablation on view numbers

w/ Integrating block	Places of a single Reinforcing block	Average Instance Accuracy
✓	conv2	93.9
✓	conv3	93.7
✓	conv4	94.0
✓	conv5	**94.3**
×	conv5	93.8

Ablation on reinforcing block

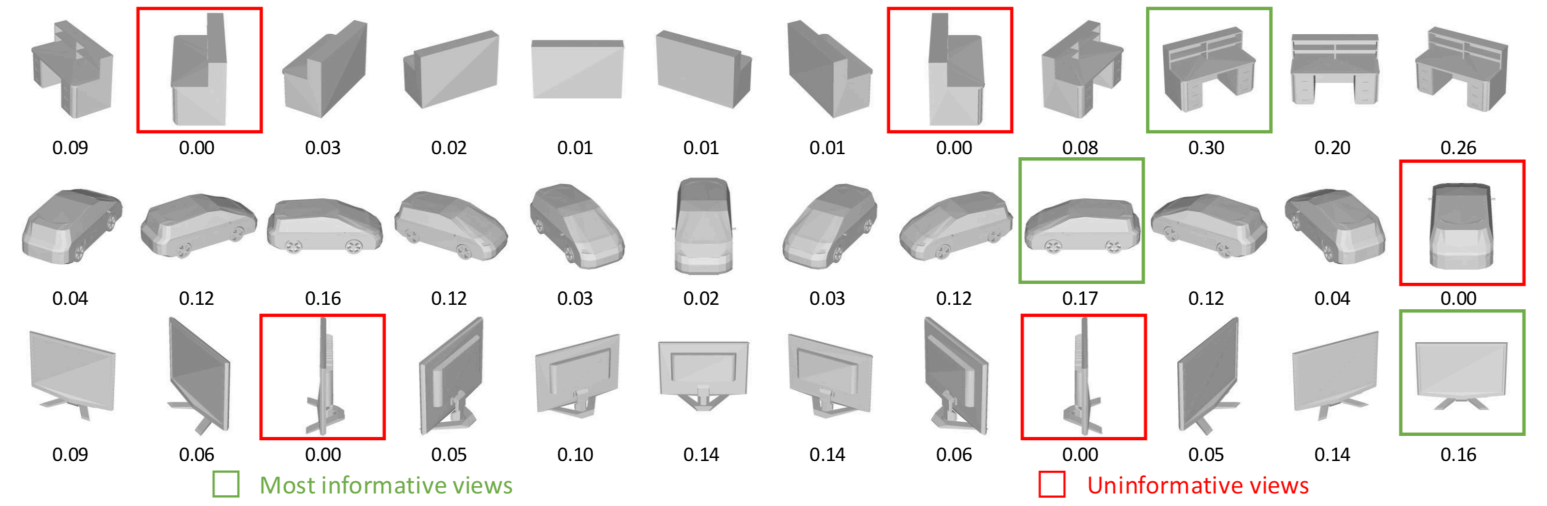
Methods	Input Modality	ModelNet40		ModelNet10		Retrieval on ModelNet40
		Inst Acc	Class Acc	Inst Acc	Class Acc	
SPH [19]	Handcraft	-	68.2	-	-	33.3
LFD [9]	Handcraft	-	75.5	-	-	40.9
3D ShapeNets [46]	Volume	-	77.3	-	83.5	49.2
VoxNet [27]	Volume	-	83.0	-	92.0	-
Subvolume Net [31]	Volume	89.2	86.0	-	-	-
Voxception-ResNet [4]	Volume	91.3	-	93.6	-	-
PointNet [30]	Points	89.2	86.2	-	-	-
PointNet++ [32]	Points w/ Normal	91.9	-	-	-	-
Kd-Networks [20]	Points	91.8	88.5	94.0	93.5	-
MVCNN [39]	12 Views	92.1	89.9	-	-	80.2
MVCNN-MultiRes [31]	Multi-resolution Views	93.8	91.4	-	-	-
Pairwise Network [17]	12 Views w/ Depth	-	91.1	-	93.2	-
RCPCNN [43]	12 Views w/ Depth, Normal	93.8	-	-	-	-
GVCNN [10]	8 Views	93.1	-	-	-	84.5
	12 Views	92.6	-	-	-	85.7
MHBN [47]	6 Views	94.1	92.2	94.9	94.9	-
	12 Views	93.4	-	-	-	
Ours	12 Views	**94.3**	**92.3**	**95.3**	**95.1**	**86.7**

Comparisons with state-of-the-art

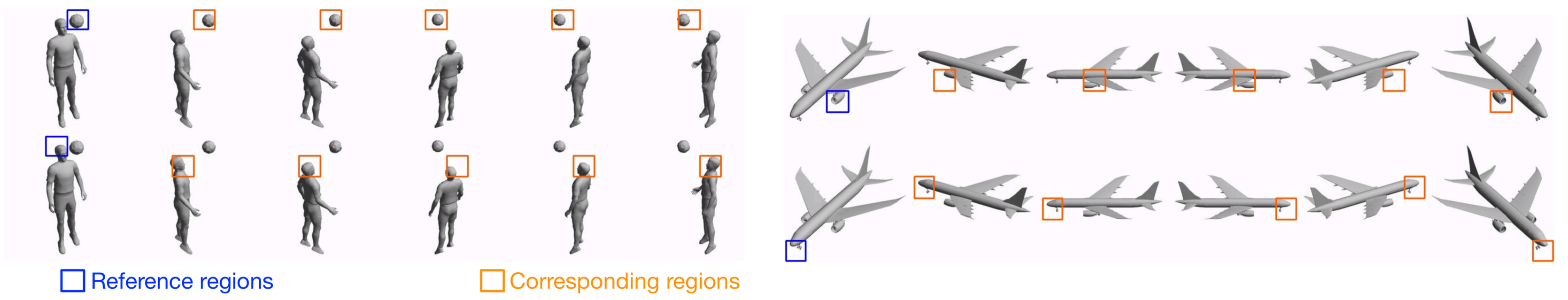


Retrieval results

Qualitative results:



Visualization of the learnt importance scores, The weight of each view is shown at the bottom of the image.



Visualization of the learnt correspondence regions, we frame the most relevant regions with orange rectangles.