ECCV'20 ONLINE

23-28 AUGUST 2020

16TH EUROPEAN CONFERENCE ON

COMPUTER VISION

WWW.ECCV2020.EU

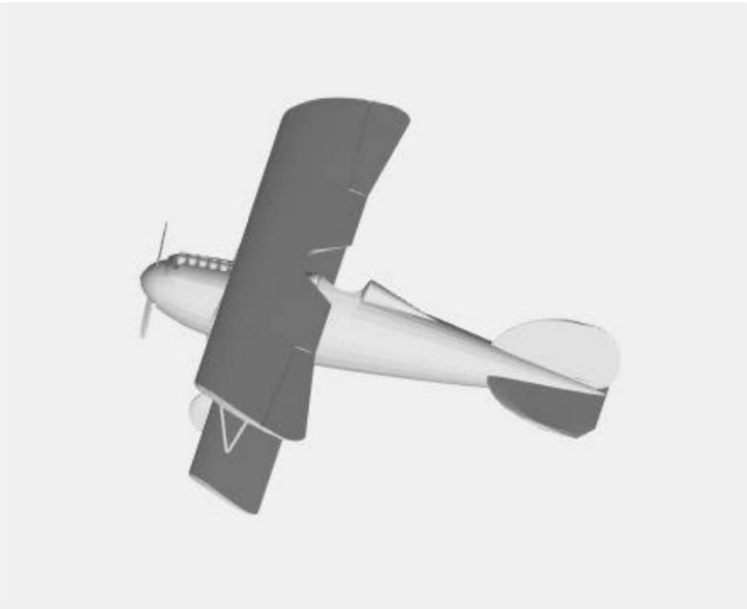
Paper ID: 3254

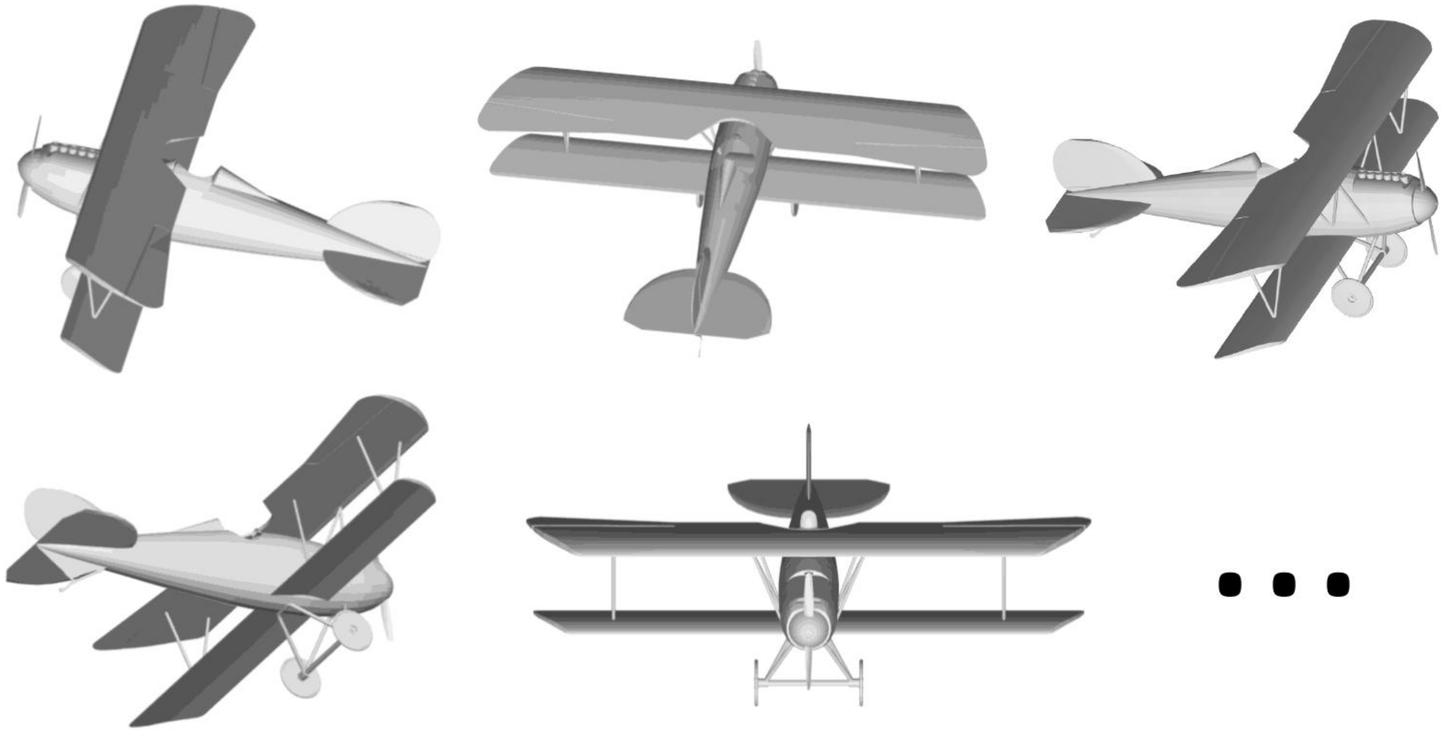
Unsupervised Multi-View CNN for Salient View Selection of 3D Objects and Scenes

Ran Song[1], Wei Zhang[1]
Yitian Zhao[2], and Yonghuai Liu[3]

[1]Shandong University, China
[2]Cixi Institute of Biomedical Engineering, CAS, China
[3]Edge Hill University, UK

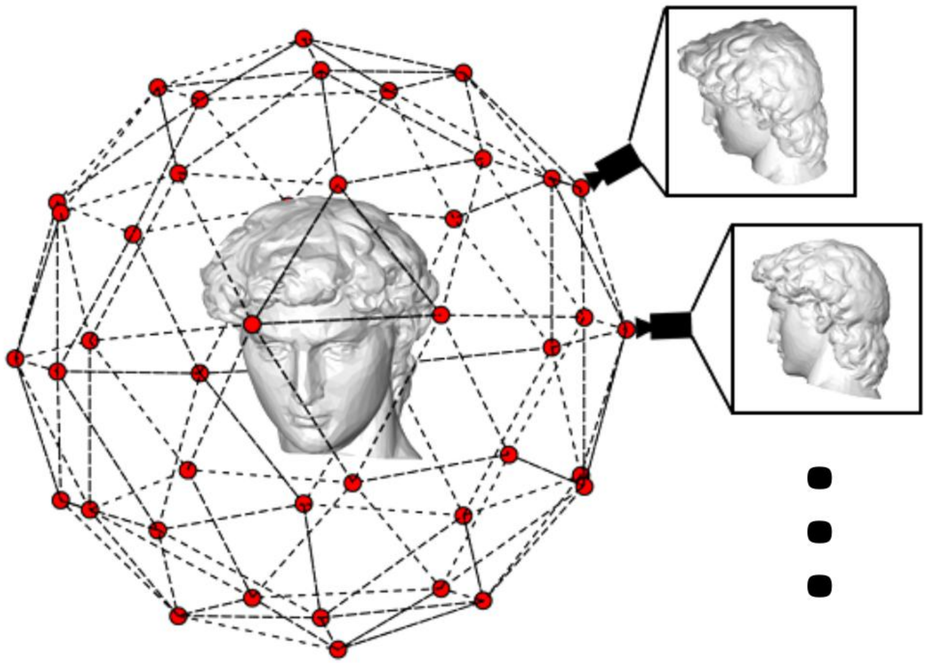
For a 3D object, which view is the best?





- Local features
- Global features
- Recognition
- Human aesthetics

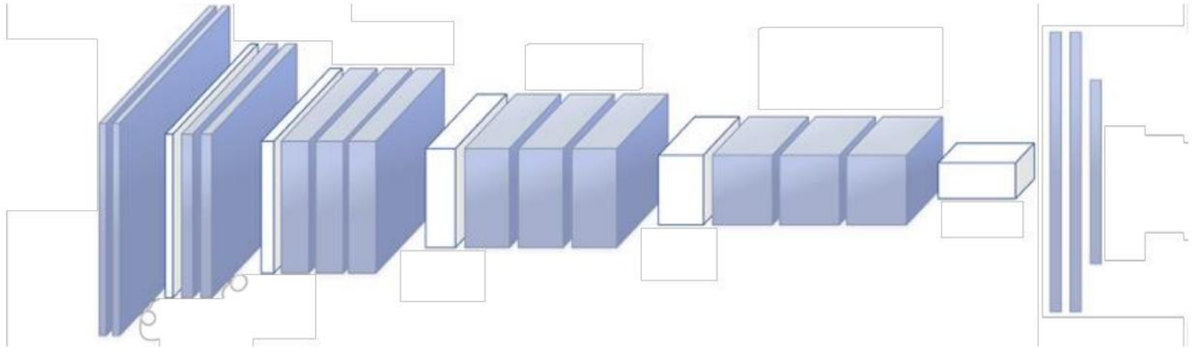
A data-driven method is naturally sound.



Data problem

- Direct data collection
  Dutagaci *et al.*: 68 objects
- Indirect data collection
  Secord *et al.*: 16 objects

Supervised deep neural network



Unsupervised 3D deep learning framework

View-object consistency

A 3D object and its projected 2D views always belong to the same object class no matter what taxonomy is applied to the classification.

MV-LogLoss:

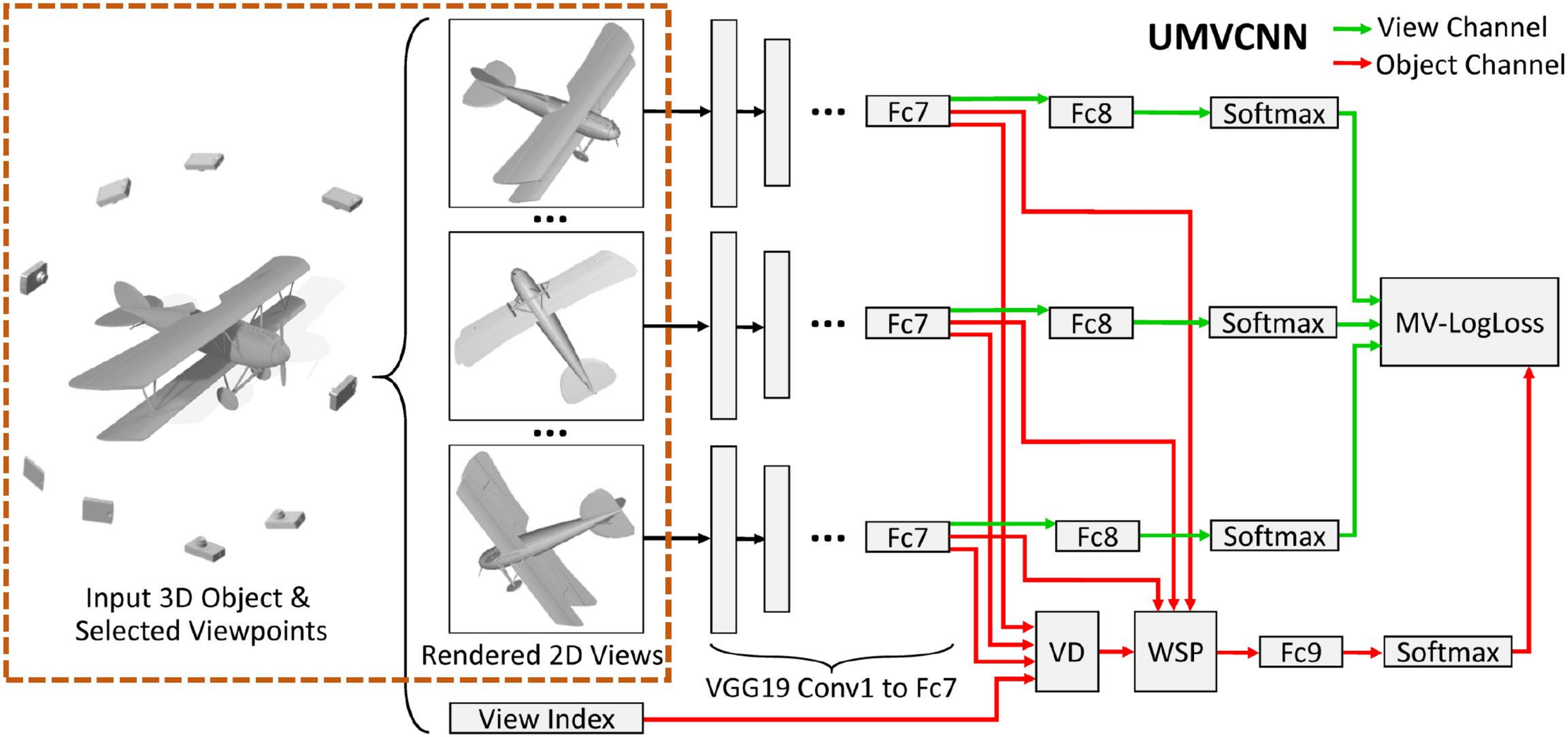
$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{c=1}^{C} \mathcal{O}_c \cdot \log\left(\mathcal{V}_c(i)\right)$$

Output of the object channel

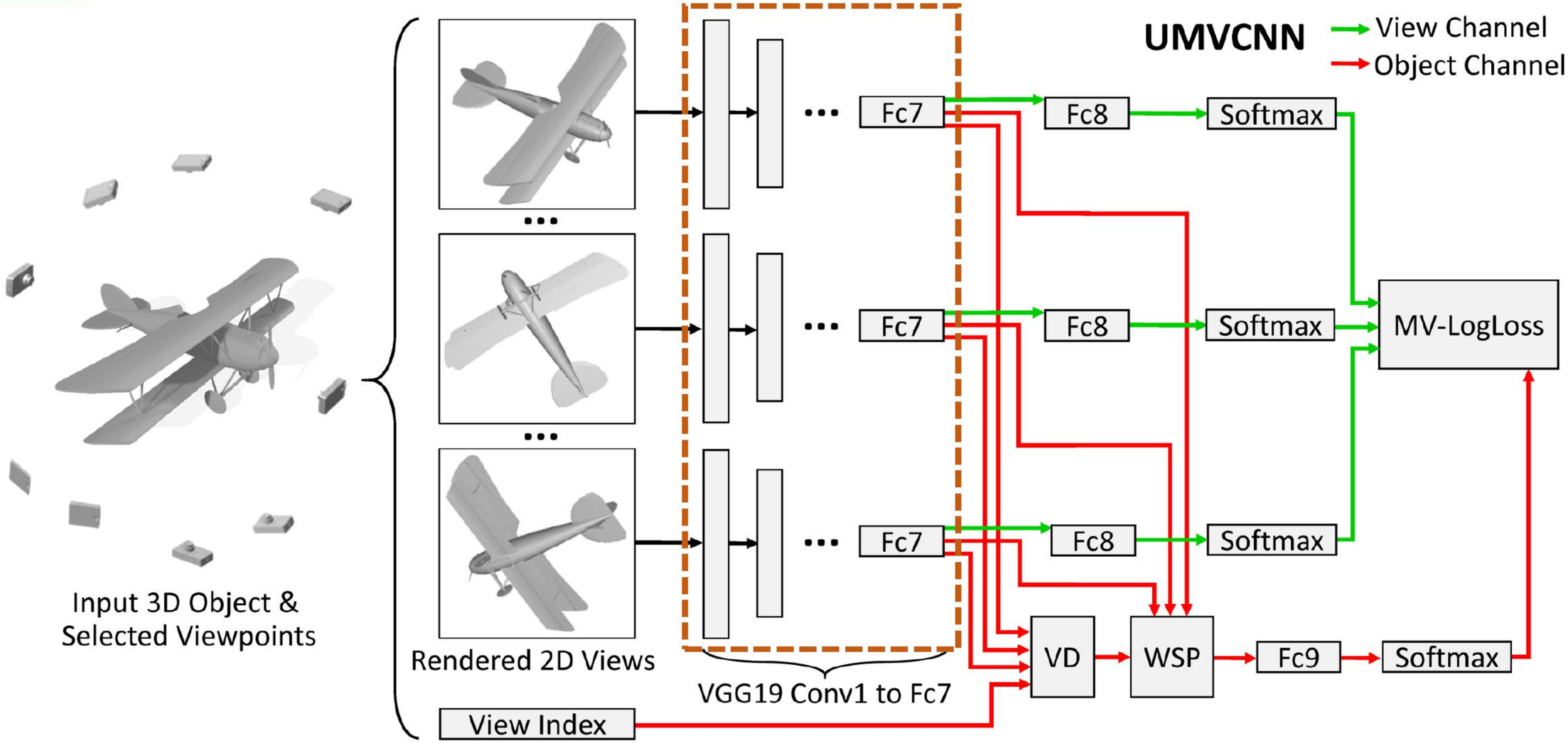
Output of the view channel

Classification loss

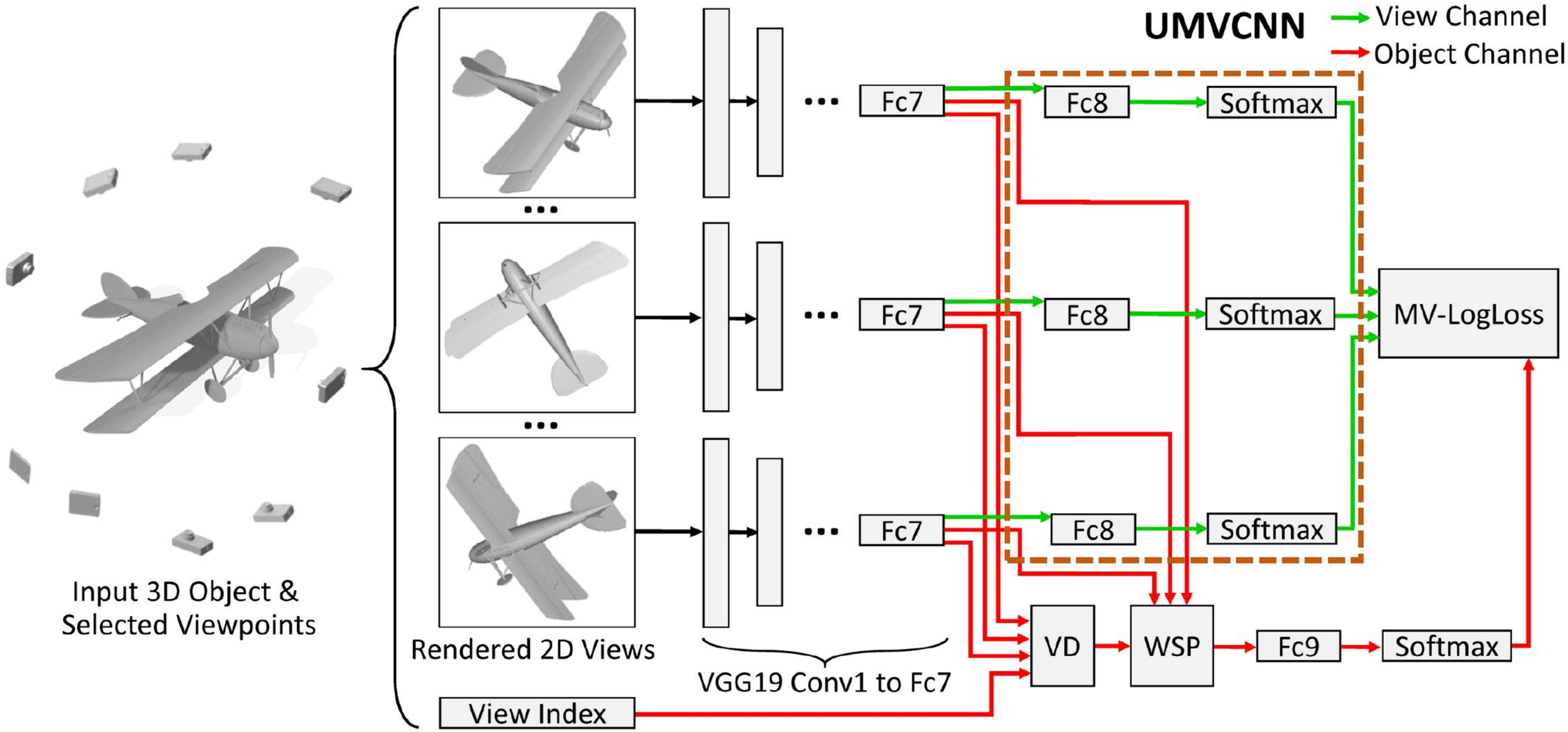
Unsupervised Multi-View CNN



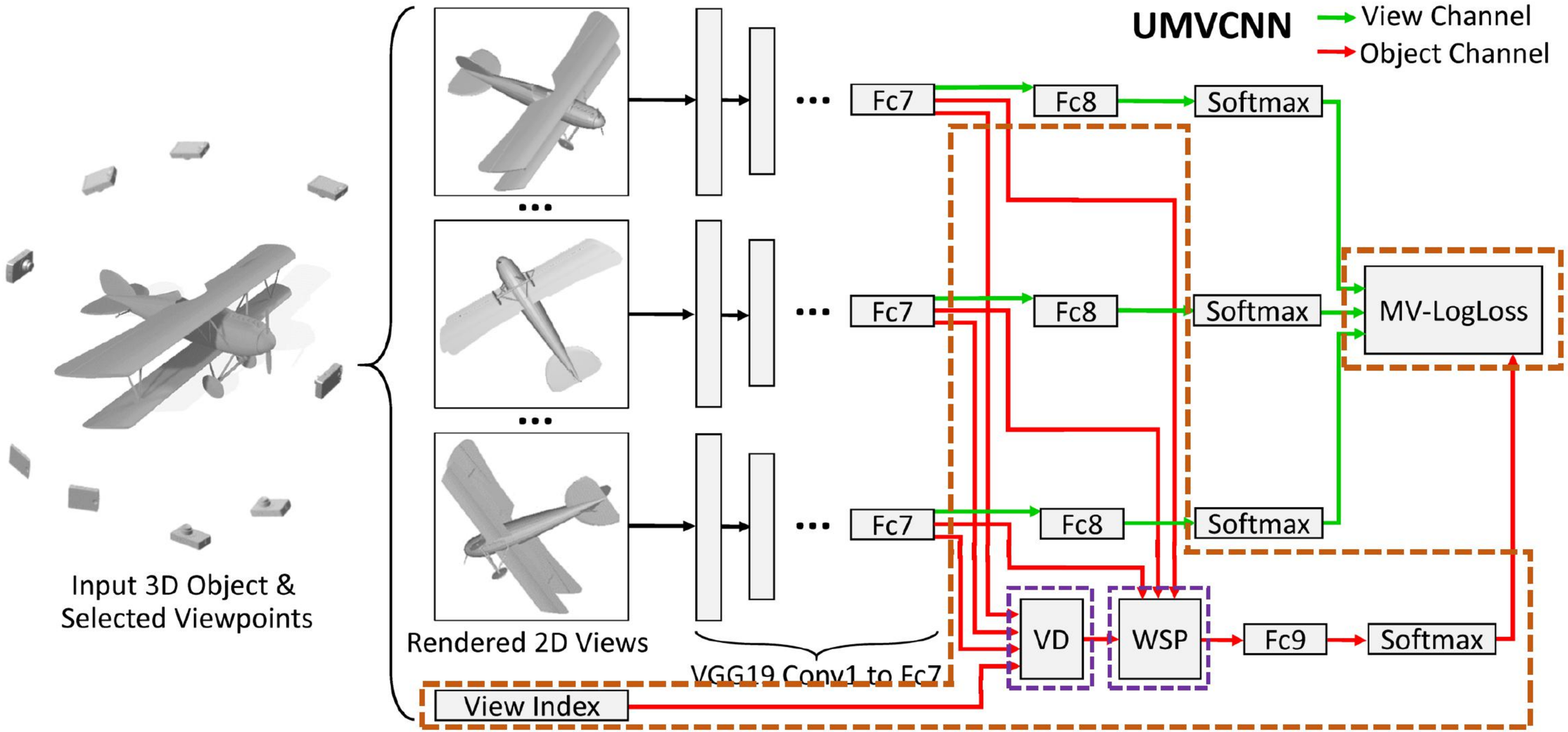
Unsupervised Multi-View CNN



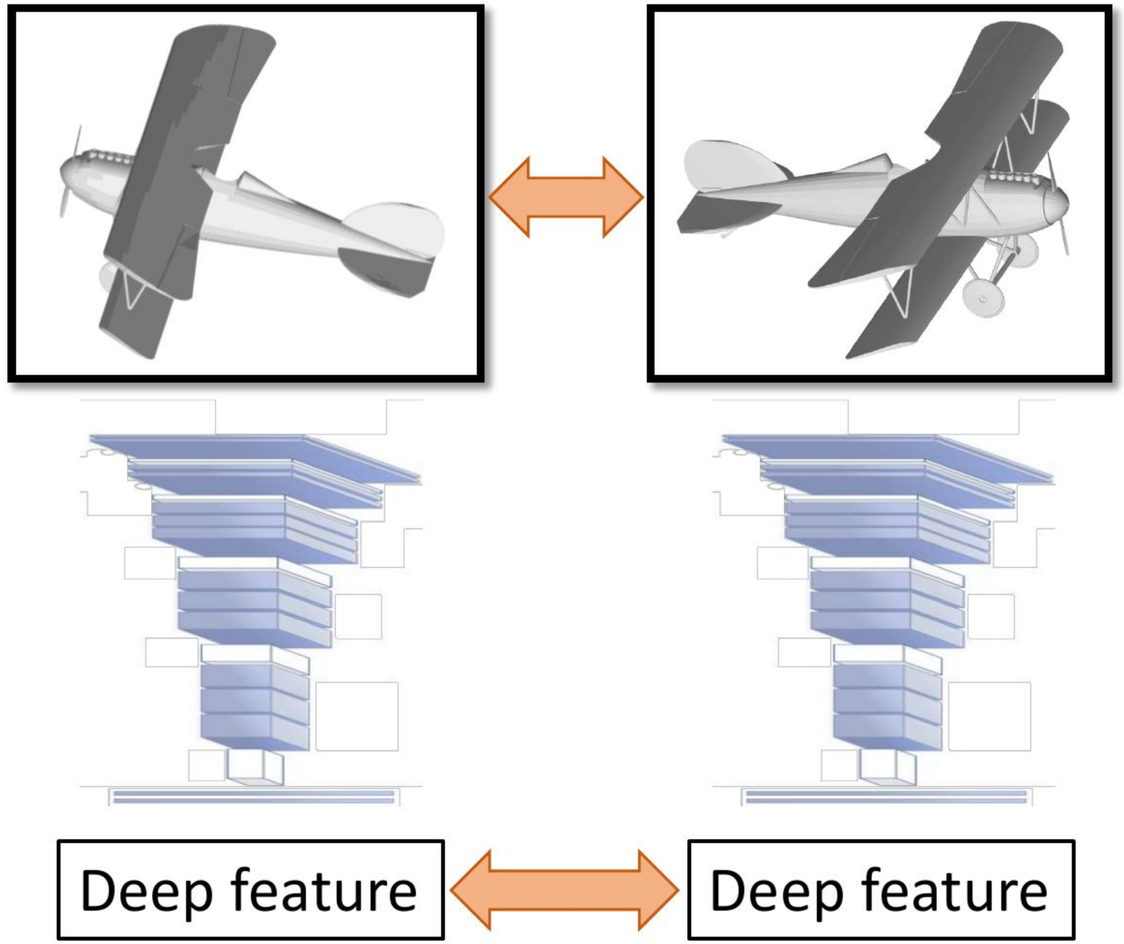
Unsupervised Multi-View CNN



Unsupervised Multi-View CNN



View Distinction Layer



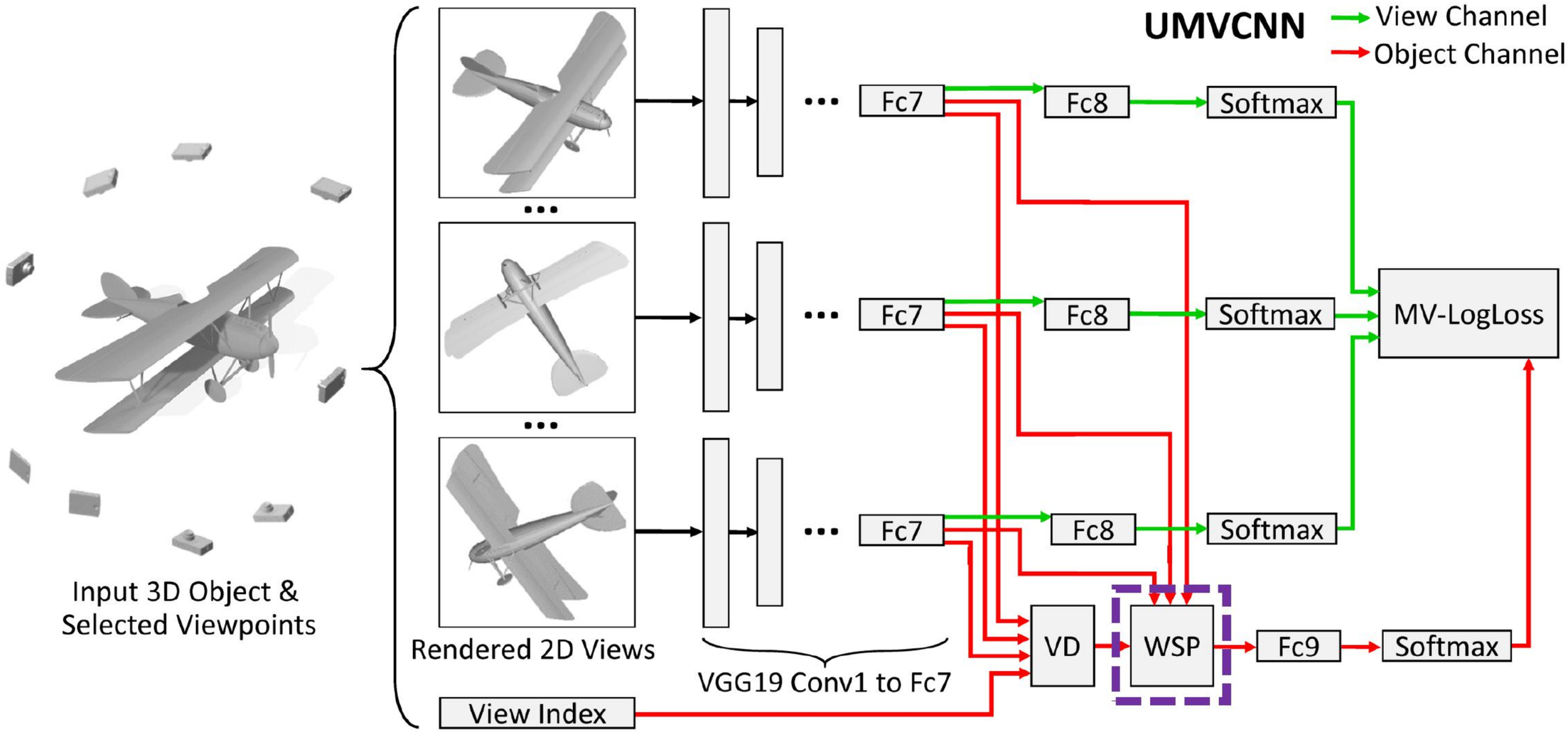
Human visual system tends to suppress the response to frequently occurring features, while at the same time it remains sensitive to features that deviate from the others.

Distinction of View i:

$$D_{ij} = \frac{\|F_i - F_j\|}{1 + \alpha \cdot \boxed{\mathrm{Geod}(\mathrm{VInd}_i, \mathrm{VInd}_j)}}$$

$$S_i = \sum_{j \neq i} D_{ij}$$

Unsupervised Multi-View CNN



Weighted Sum Pooling Layer

- Aggregate the learned knowledge from all 2D views to create a single descriptor for the 3D object.
- Cast the influence of view distinction into the aggregation process.

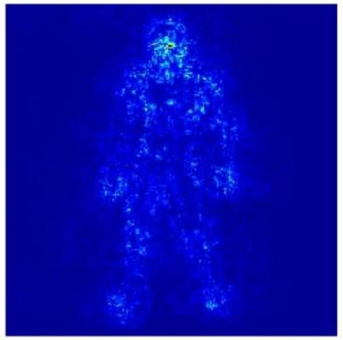
Weighted sum pooling (WSP): $$P = \sum_{i=1}^{N} F_i S_i$$

- WSP vs. max-pooling in terms of mean VSE (View Selection Error) proposed by Dutagaci et al.:

WSP: 0.367	Max-pooling: 0.384

Salient View Selection

Backpropagation



2D-to-3D saliency transfer

View-based 3D saliency map H_i

Aggregate multi-view 3D saliency maps into a single one:

$$H = \sum_{i=1}^{N} S_i H_i$$

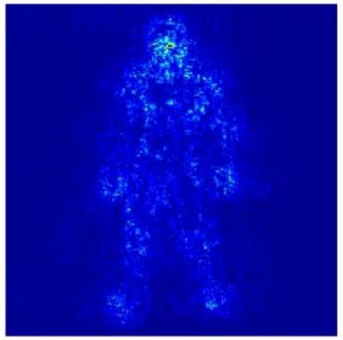
$$v_s = \arg\max_{v} \Big(\sum_{m \in B(v)} H(m) \Big)$$

The saliency of the vertex m

The set of vertices visible from the viewpoint v

Salient View Selection

Backpropagation



2D-to-3D saliency transfer

View-based 3D saliency map H_i

Aggregate multi-view 3D saliency maps into a single one:

$$H = \sum_{i=1}^{N} S_i H_i$$

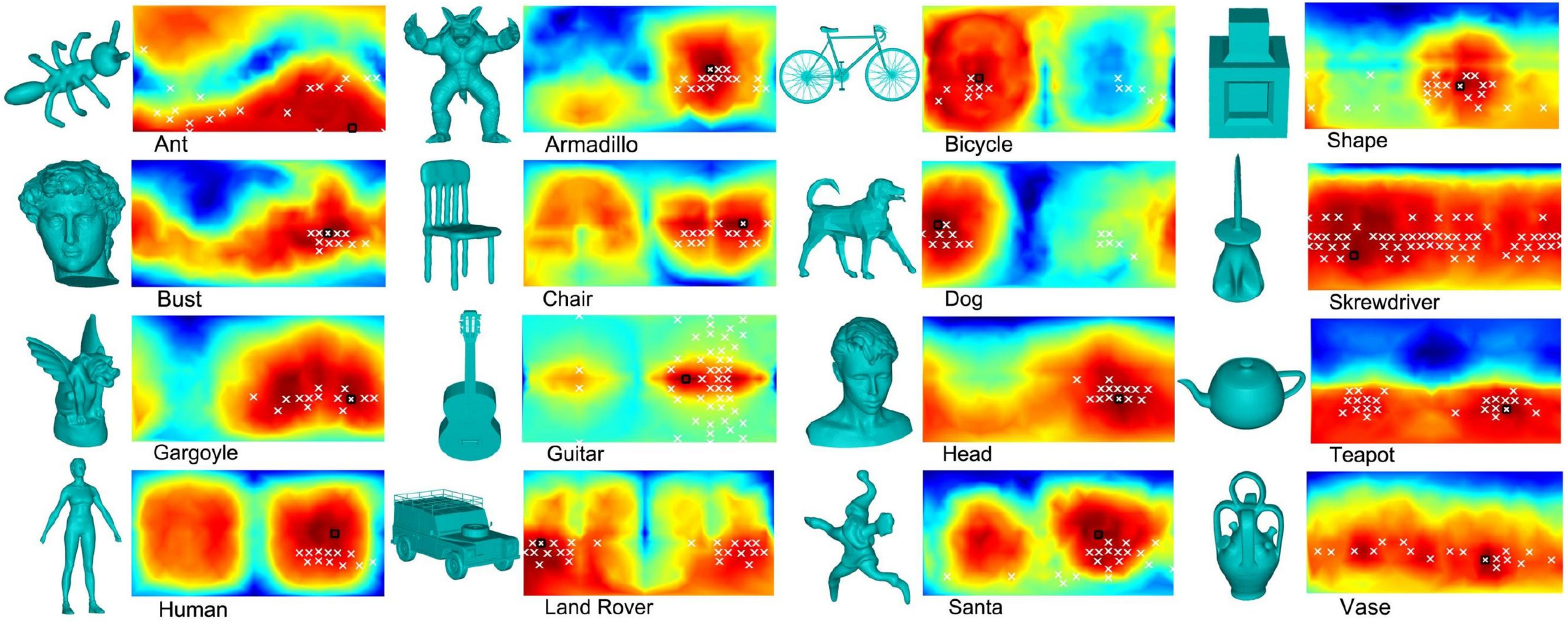
$$v_s = \arg\max_{v} \left(\sum_{m \in B(v)} H(m) \right)$$

The saliency of the vertex m

The saliency map of the viewpoints

The set of vertices visible from the viewpoint v

Results: Salient View + Viewpoint Saliency Map

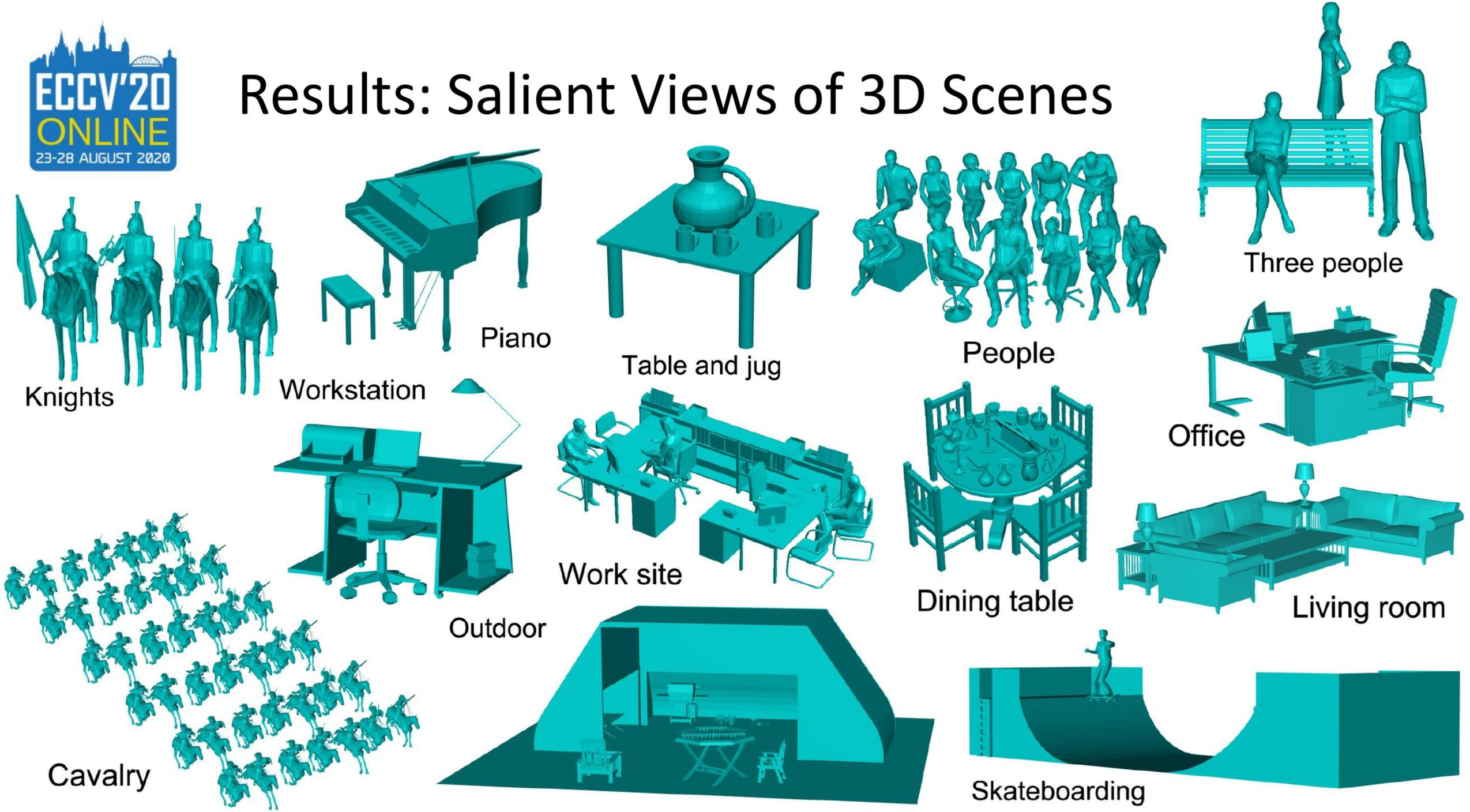


Results: View Selection Error (VSE)

VSE: the geodesic distance between the viewpoint found by a method and the ground truth supplied by a human subject on a unit view sphere and is averaged over the choices of all subjects with the consideration of object-specific symmetry.

View Selection Method	mean VSE	median VSE	SD of VSE	IQR of VSE	n
View area [7]	0.517	0.539	0.186	0.306	6
Ratio of visible area [22]	0.473	0.473	0.196	0.338	1
Surface area entropy [29]	0.396	0.386	0.144	0.195	8
Silhouette length [22]	0.446	0.445	0.172	0.275	7
Silhouette entropy [21]	0.484	0.469	0.153	0.241	5
Curvature entropy [21]	0.474	0.466	0.139	0.239	8
Mesh saliency [16]	0.430	0.395	0.165	0.233	2
Deep mesh distinction [26]	0.380	0.346	0.173	0.314	11
UMVCNN-30	0.367	0.336	0.165	0.236	20

Results: Salient Views of 3D Scenes



Conclusions

- The view-object consistency enables the proposed UMVCNN to effectively learn the salient view of a 3D object or scene.
- The UMVCNN benefits from the heuristic of view distinction for the specific task of salient view selection.
- Future work will investigate if the unsupervised learning framework based on the view-object consistency is amenable to a range of 3D shape understanding tasks.
- Data and codes will be made available at https://github.com/rsong/UMVCNN.

Thanks for your time and attention!