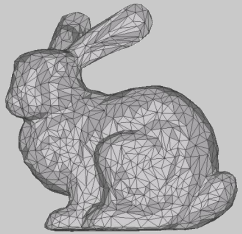
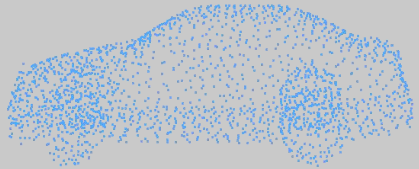
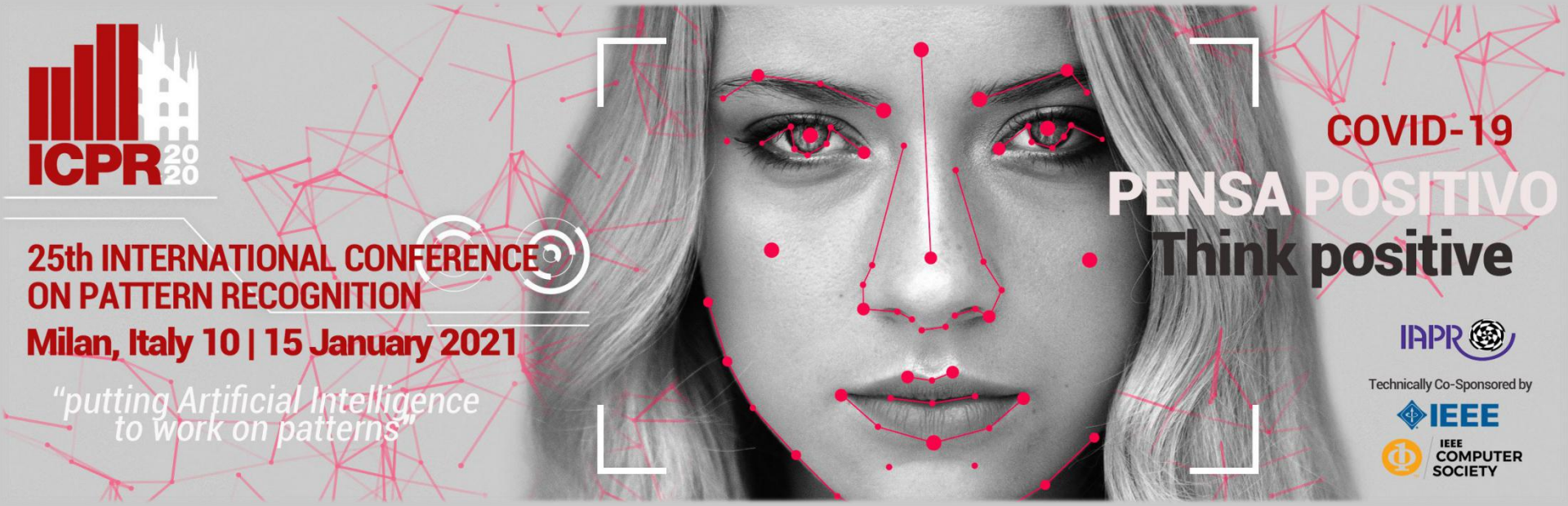
MANet: Multimodal Attention Network based Point- View fusion for 3D Shape Recognition





Yaxin Zhao, Jichao Jiao*, Ning Li, Zhongliang Deng
Beijing University of Posts and Telecommunications



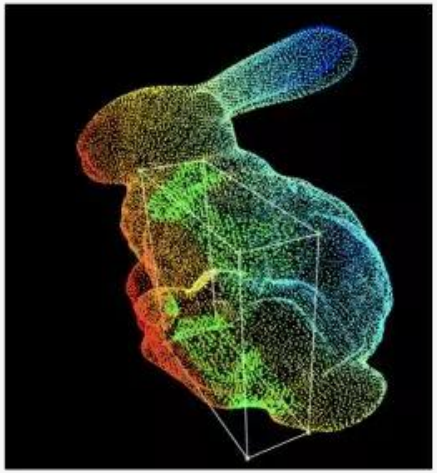
January 10-15, 2021

CONTENTS

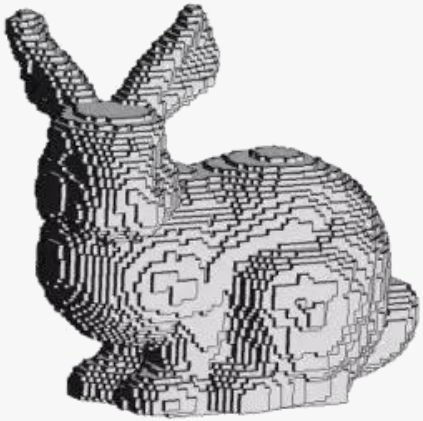
The Data Format

3D Data：

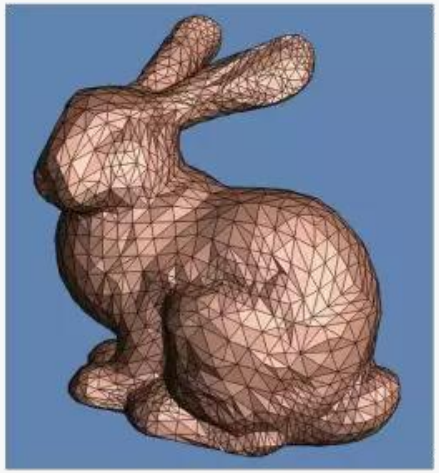
(a)



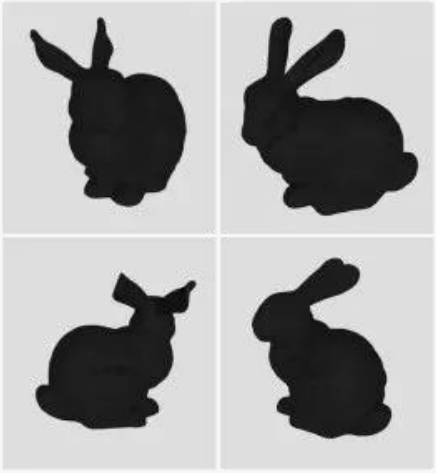
(b)



(c)



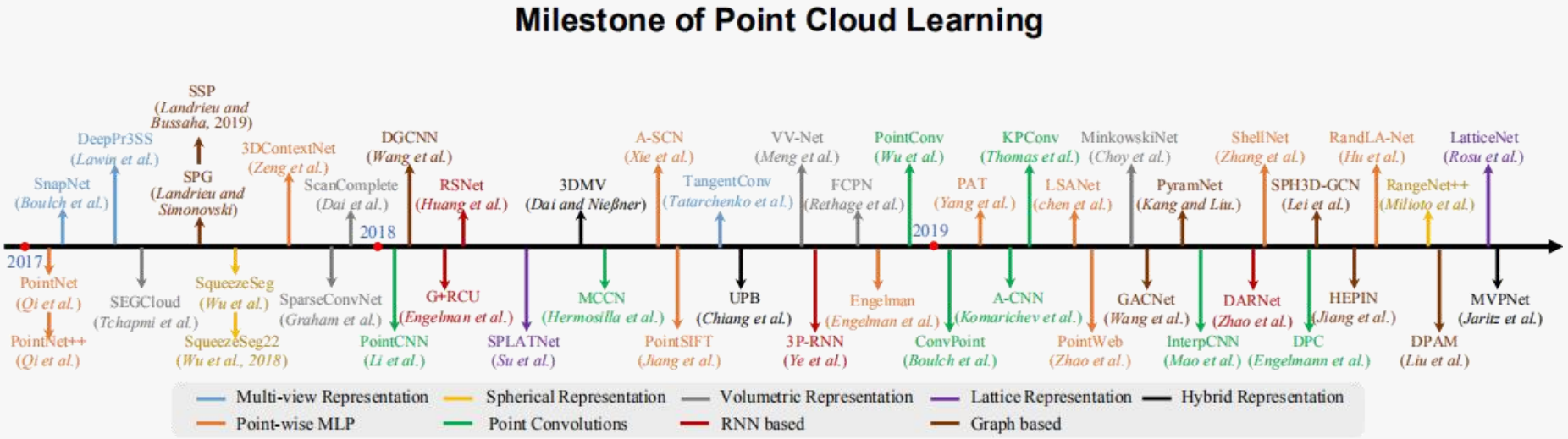
(d)



(a) Point-cloud (b) Voxel (c) Triangular mesh (d) Multi-view

CONTENTS

Development



Guo, Yulan , et al. "Deep Learning for 3D Point Clouds: A Survey." IEEE Transactions on Pattern Analysis and Machine Intelligence PP.99(2020):1-1.

CONTENTS

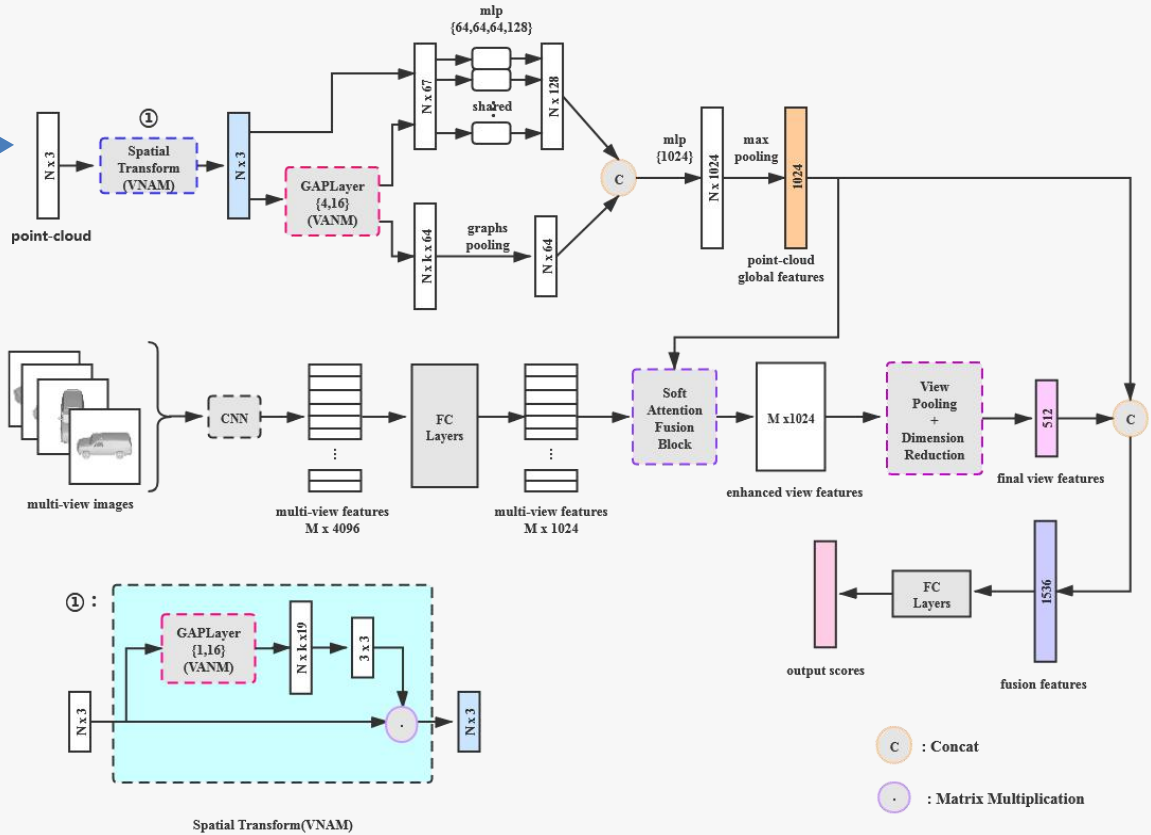
The main contributions of this paper are summarized as follows:

- MANet which uses multi-modal attention mechanism to well fuse multi-source data.
- In the point-cloud branch, a generalized processing method which named VNAM is introduced to explore the attention relationship between nodes and neighborhood points in the 3D point-cloud.
- A soft attention fusion scheme based on point-view data is proposed and the point-cloud global features are used to mine the contribution of each multi-view image to the whole shape recognition.

MANet

point-cloud branch

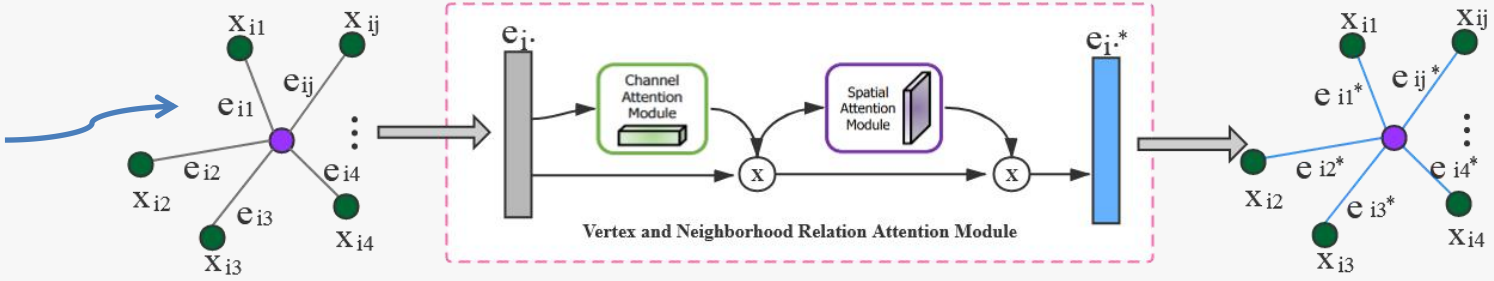
multi-view branch

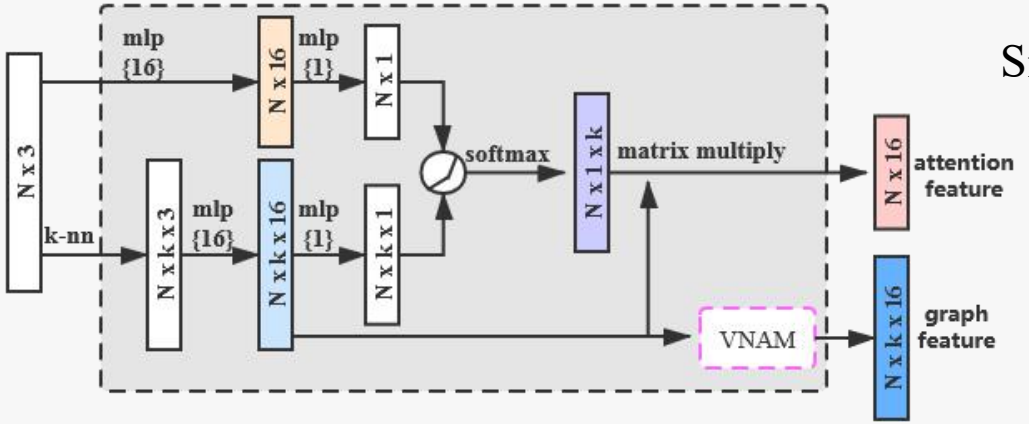


MANet

ICPR2020

VNAM





Single-head GAPLayer with VNAM

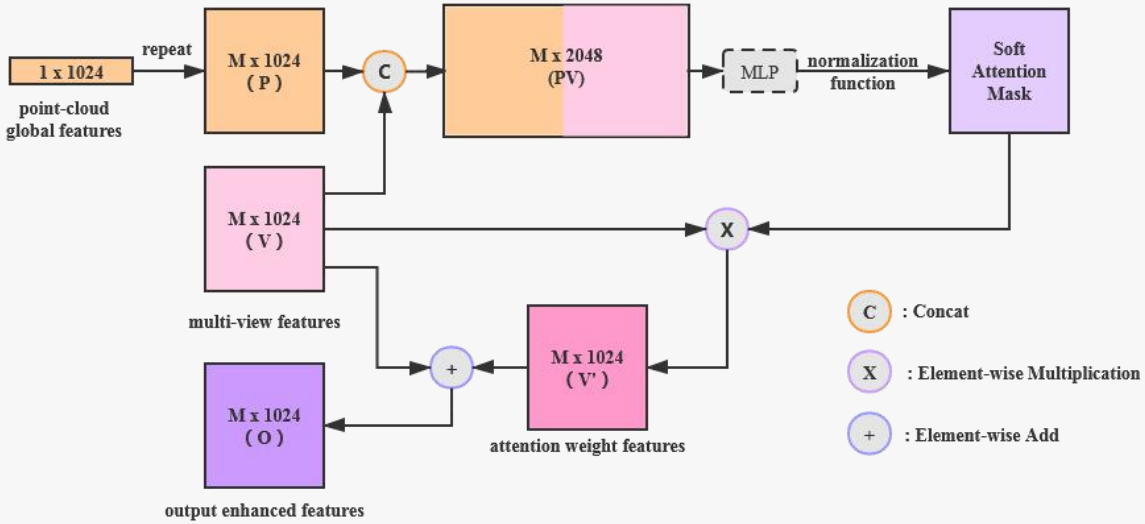
Basic Concepts

In the attention fusion block, in order to solve the problem that the two types of features are in different feature spaces, we map the global point-cloud features to the subspace of the multi-view features to obtain the features $P = \{P_1, P_2 \cdots, Pm\}$, and then fuse it with the multi-view features $V = \{V_1, V_2 \cdots, Vm\}$ to obtain the fused features $PV = \{I_1, I_2 \cdots, \mathrm{Im}\}$, where m is the number of multi-view images, which we set m=12 in the experiment. Weight coefficients $C(W) = \{W_1, W_2 \cdots, Wm\}$ are generated after normalizing the fused features, i.e.:

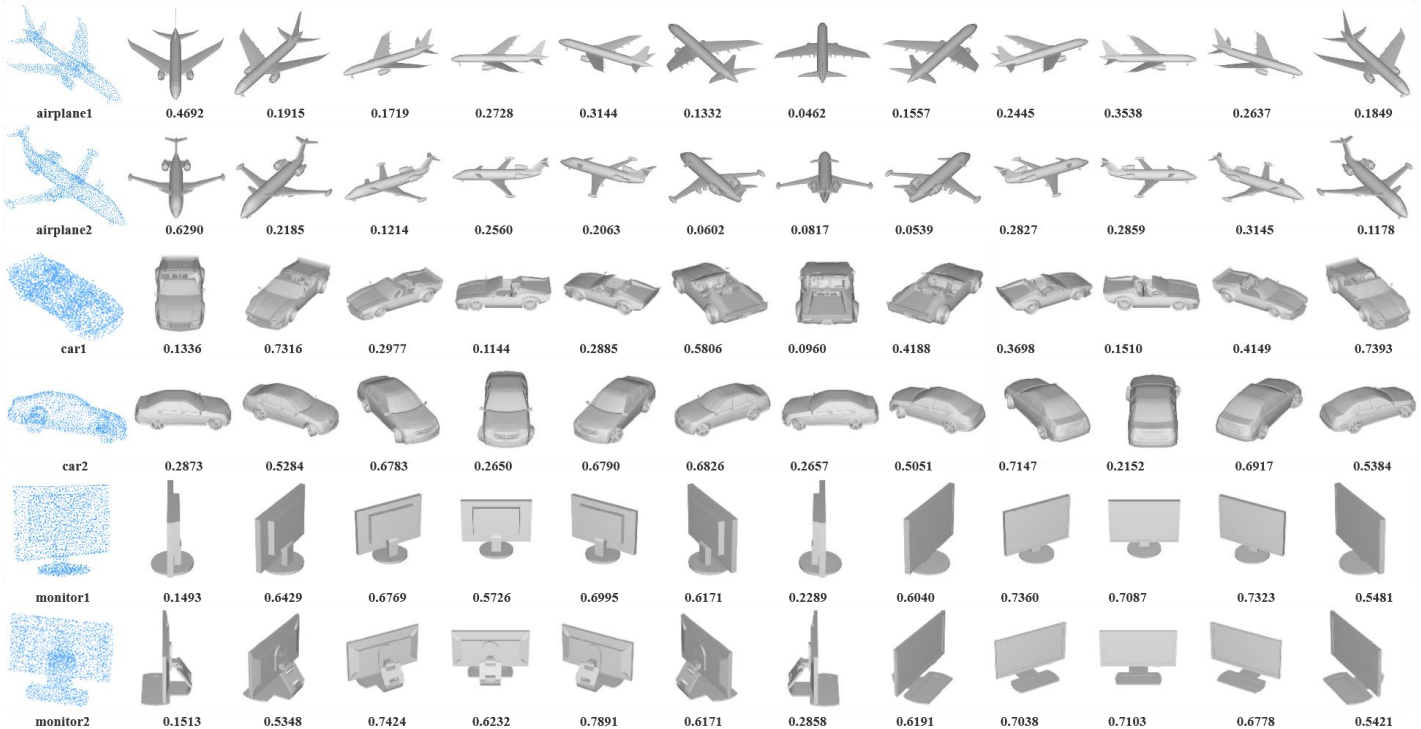
$$C(W) = F(MLP(P, V))$$

The normalization function $\mathbf{F}(\cdot)$ used in this article is the sigmoid function, so the calculation formula of $W_i \in C(W), 1 \le i \le m$ is:

$$W_i = \frac{exp(MLP(P_i, V_i))}{\sum_{k=1}^{m} exp(MLP(P_k, V_k))}, 1 \le i, k \le m$$



Soft attention fusion module. It takes point-cloud global features and multi-view features as input and outputs enhanced multi-view features.



Soft attention fusion mask.

Result

TABLE I. CLASSIFICATION AND RETRIEVAL RESULTS ON THE MODELNET40 DATASET

Method	Data	Classification (Overall Accuracy)	Retrieval (mAP)
3D ShapeNets	Volumetric	77.3%	49.2%
VoxNet	Volumetric	83.0%	-
MVCNN-MultiRes	Volumetric	91.4%	-
MVCNN(AlexNet)	12 views	89.9%	80.2%
MVCNN(GoogLeNet)	12 views	92.2%	83.0%
PointNet	Point-cloud	89.2%	-
PointNet++	Point-cloud	90.7%	-
KD-Network	Point-cloud	91.8%	-
SO-Net	Point-cloud	90.9%	-
DGCNN	Point-cloud	92.2%	81.6%
GAPNet	Point-cloud	92.4%	-
FusionNet	Volumetric and 20/60 views	90.8%	-
PVNet	Point-cloud and 12 views	93.2%	89.5%
Ours	Point-cloud and 12 views	93.4%	90.1%
PVRNet	Point-cloud and 12 views	93.6%	90.5%

[a] Symbol '-' Means Results Are Unavailable.

TABLE II. COMPARISON BETWEEN MANET AND PVRNET

Method	Classification (Overall Accuracy)	Model Size
MANet	93.4%	507.1 MB
PVRNet	93.6%	579.2 MB

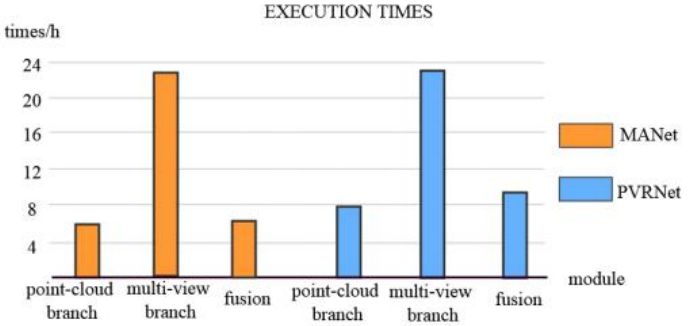


Fig. 7. Comparison of execution time between MVNet and PVRNet.

TABLE III. ABLATION EXPERIMENTS OF VNAM

Method	Classification (Overall Accuracy)	Model Size
GAPNet	91.21% (92.40%)	22.9 MB
GAPNet(VNAM)	91.45%	21.9 MB

MANet: Multimodal Attention Network based Point- View fusion for 3D Shape Recognition

Yaxin Zhao, Jichao Jiao*, Ning Li, Zhongliang Deng

Beijing University of Posts and Telecommunications

THANKS!