MeteorNet: Deep Learning on Dynamic 3D Point Cloud Sequences Supplementary Materials

Xingyu Liu
Stanford University

Mengyuan Yan
Stanford University

Jeannette Bohg
Stanford University

A. Overview

In this document, we provide additional detail on MeteorNet as presented in the main paper. We present additional results on the accuracy of action recognition (Sec. B) and the outlier ratio in scene flow estimation (Sec. C). In Section D, we provide more details on the architectures used in various experiments. In Section E, we provide a runtime analysis for our model on the Synthia dataset. In Section F, we present the proof to our theorem. In Section G we provide qualitative example results for various experiments. Lastly, in Section H, we give a brief rationale for the name of our neural network.

B. MSRAction3D Per-class Accuracy

In the main paper, we showed that MeteorNet with multiple frames of point clouds as input outperforms various baselines. We obtained all possible clips of a certain length from a full-length point cloud sequence and computed the softmax classification scores on them individually. The final prediction is the average of softmax scores of all clips. We explored using extremely long sequence and its effect on final classification accuracy. The classification accuracy saturates at a sequence length of 24. Given the 15fps frame rate in MSRAction3D, a sequence length of 24, i.e. 1.6s, is close to the average length of a complete action.

In Figure 1, we illustrate the per-class accuracy gain of MeteorNet-cls with 24 frames as input compared to PointNet++ with 1 frame as input.

We can see that categories that may only be discriminated when observed over time show a significant gain in accuracy when using a sequence of point clouds as input. Categories that can be easily discriminated without temporal information show little or a negative gain in accuracy. For example, the categories "forward punch", "horizontal arm wave" and "draw x" show a large improvement in accuracy. These three categories are similar since they all involve stretching arms forward and thus requires temporal information to be correctly classified. Categories such as "pick up & throw" or "golf waving" have a very discriminative posture even in single frames and therefore show only the slightest or negative accuracy gain.

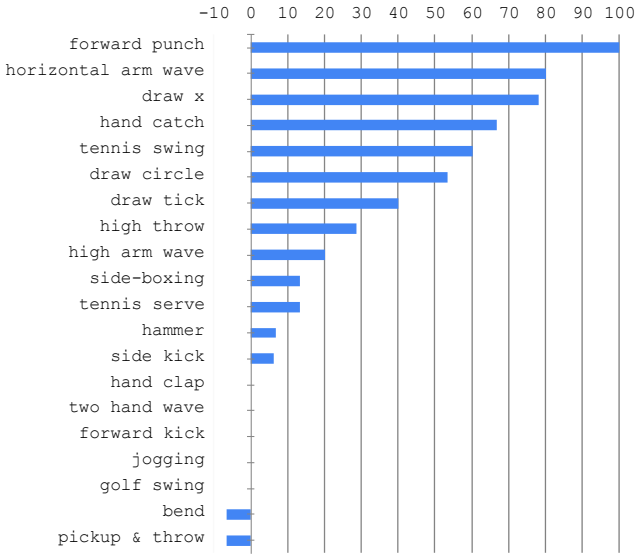

Figure 1: Per-class accuracy gain (%) of 24 frames MeteorNet-cls compared to PointNet++.

Method	Frames	Threshold for outlier (m)							
		0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
FlowNet3D [1]	2	6.88	4.28	2.19	1.31	1.02	0.77	0.59	0.54
MeteorNet-flow (direct)	3	8.43	3.71	1.87	1.12	**0.82**	**0.65**	0.55	**0.42**
	4	6.67	**3.32**	**1.44**	**1.09**	0.84	**0.65**	**0.53**	**0.42**
MeteorNet-flow (chain)	3	**6.35**	3.72	2.39	1.49	1.16	0.92	0.74	0.62
	4	7.88	3.50	1.95	1.27	0.85	0.72	0.63	0.58

Table 1: Scene flow EPE outlier ratio (%) given different threshold values.

The results support our intuition that the Meteor module effectively captures dynamic content of point cloud sequences.

C. Outlier Ratio of Scene flow

The ratio of outliers is an important metric that evaluates the robustness of scene flow estimation. We investigate scene flow outlier ratio on KITTI scene flow dataset [2]. We set different EPE threshold for determining outliers and list the outlier ratio in Table 1.

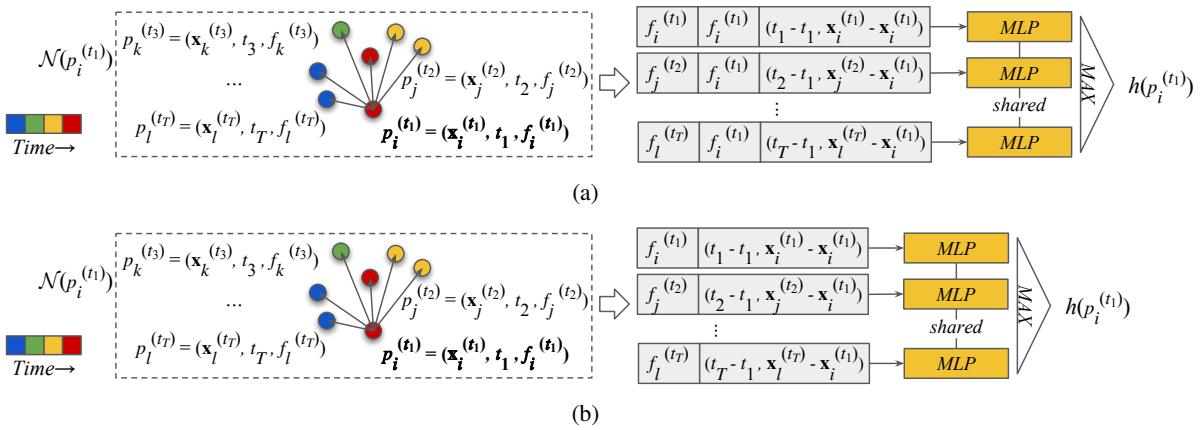

Figure 2: The architecture of (a) **Meteor-rel module** and (b) **Meteor-ind module**. The dashed box denotes the neighborhood $\mathcal{N}(p_i^{(t)})$ of $p_i^{(t)}$ (in bold) from which all arrows start. The neighborhood $\mathcal{N}$ can be determined by direct grouping or chained-flow grouping. In the figure, $\mathbf{x}$, t and f denotes the 3D spatial coordinate, time coordinate and feature vector of a point respectively; "MLP" denotes ζ in Equation (1) and (2), which is the multi-layer individually and independently perceptron applied.

As we can see, with more frames as input, MeteorNet-flow can reduce outlier ratio over FlowNet3D. Besides, MeteorNet-flow using chained-flow grouping with 3 frames as input has the best outlier ratio for a small threshold. However, when the threshold gets larger , MeteorNet-flow using direct grouping is advantageous.

D. Architecture Details

In this section, we provide details on the architectures used in the main paper. We used the same notation as the main paper and assume the input point cloud sequence is $(\{p_i^{(1)}\}, \ldots, \{p_i^{(T)}\}) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_T$ and the local spatiotemporal neighborhood of $p_i^{(t)}$ is $\mathcal{N}(p_i^{(t)})$.

D.1. Meteor Module Architecture

For every point $p_i^{(t)}$ in the point cloud sequence $\{p_i^{(t)}\}$, Meteor module calculates its updated feature vector $h(p_i^{(t)})$. In Section 3.2 of the main paper, we presented two instantiation of h.

The first instantiation is for applications where point correspondence is important, such as scene flow. For each $(p_j^{(t')}, p_i^{(t)})$ pair, we pass the feature vectors of two points and their 4D position difference into to an MLP with shared weights ζ, followed by an element-wise max pooling

$$h(p_i^{(t)}) = \underset{p_j^{(t')} \in \mathcal{N}(p_i^{(t)})}{MAX} \{\zeta(f_j^{(t')}, f_i^{(t)}, \mathbf{x}_j^{(t')} - \mathbf{x}_i^{(t)}, t' - t)\} \quad (1)$$

This instantiation is able to learn the relation between two frames of point clouds. We name the resulting Meteor module *Meteor-rel*. The architecture of Meteor-rel is illustrated in Figure 2(a).

The second instantiation is for applications where point correspondence is not important, such as semantic segmentation. We pass the feature vector of $p_j^{(t')}$ and 4D position difference between $p_j^{(t')}$ and $p_i^{(t)}$ to ζ followed by an element-wise max pooling

$$h(p_i^{(t)}) = \underset{p_j^{(t')} \in \mathcal{N}(p_i^{(t)})}{MAX} \{\zeta(f_j^{(t')}, \mathbf{x}_j^{(t')} - \mathbf{x}_i^{(t)}, t' - t)\} \quad (2)$$

We name the resulting Meteor module *Meteor-ind*. Its architecture is illustrated in Figure 2(b).

Similar to pooling in CNN, the output of both Meteor-ind and Meteor-rel modules can be downsampled by farthest-point-sampling.

D.2. MeteorNet-cls Architecture

MeteorNet-cls $\mathcal{C}$ takes a point cloud sequence $\{p_i^{(t)}\}$ as input and produces a classification score c for the whole sequence

$$c = \mathcal{C}(\{p_i^{(1)}\}, \{p_i^{(2)}\}, \ldots, \{p_i^{(T)}\})$$

MeteorNet-cls consists of four **Meteor-ind** modules and used **Early fusion** where the points from different frames are mixed at the first layer. The final Meteor-ind module will max-pool the point cloud to be only one point. The final fully-connected (FC) layer is 20 dimensional which corresponds to the number of classes in the MSRAction3D dataset. The final FC layer is deployed with a dropout layer with dropout rate of 0.5 for regularization. The architecture of MeteorNet-cls is illustrated in Figure 3.

D.3. MeteorNet-seg Architecture

MeteorNet-seg $\mathcal{S}$ takes a point cloud sequence $\{p_i^{(t)}\}$ as input and produces a classification score $c_i^{(t)}$ for every point in the sequence

$$(\{c_i^{(1)}\}, \ldots, \{c_i^{(T)}\}) = \mathcal{S}(\{p_i^{(1)}\}, \{p_i^{(2)}\}, \ldots, \{p_i^{(T)}\})$$

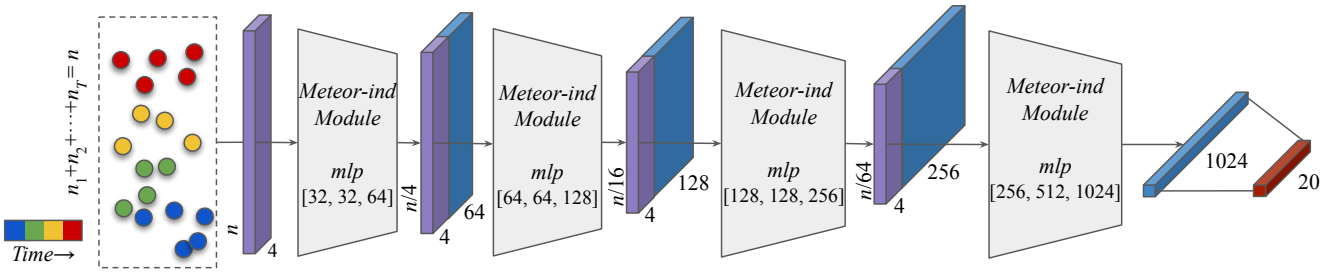

Figure 3: The architecture of **MeteorNet-cls**.

	MeteorNet-seg-s	MeteorNet-seg-m	MeteorNet-seg-l
mlp 1	[32,32,64]	[32,32,128]	[32,64,128]
mlp 2	[64,64,128]	[64,64,256]	[64,128,256]
mlp 3	[128,128,256]	[128,128,512]	[128,256,512]
mlp 4	[256,256,512]	[256,256,1024]	[256,512,1024]

Table 2: Architecture configuration for different versions of MeteorNet-seg. "mlp {1,2,3,4}" corresponds to MLPs of Meteor modules in Figure 4.

MeteorNet-seg consists of four **Meteor-ind** modules and used **Early fusion** where the points from different frames are mixed at the first layer. The point cloud will first be downsampled and then upsampled to the original point cloud through feature propagation layers [4]. We added skip connections so that local features at early stages of the network can be used in the feature propagation. The output has 12 channels, same number as the number of classes in the Synthia dataset. The final FC layer is deployed with a dropout layer with dropout rate of 0.5 for regularization. The architecture of MeteorNet-seg is illustrated in Figure 4.

An ablation study in Section 5.2 of the main paper explored several architecture choices. We listed the architecture configurations in Table 2. Compared to MeteorNet-seg-s, MeteorNet-seg-m has a larger bottleneck dimension at each max pooling layer. Compared to MeteorNet-seg-m, MeteorNet-seg-l has the same max pooling dimensions but larger dimensions in non-bottleneck layers.

D.4. MeteorNet-flow Architecture

MeteorNet-flow $\mathcal{F}$ takes a point cloud sequence $\{p_i^{(t)}\}$ as input and estimates a flow vector $\mathfrak{f}_i^{(T)}$ for every point in frame T

$$\{\mathfrak{f}_i^{(T)}\} = \mathcal{F}(\{p_i^{(1)}\}, \{p_i^{(2)}\}, \ldots, \{p_i^{(T)}\})$$

MeteorNet-flow used **Late fusion**. It first employs per-frame set abstraction layers [4] to downsample the point clouds and learn local features for each frame individually. Then, one **Meteor-rel** module is used to aggregate information from all frames. Only the points in frame T are selected for subsequent part of the network. After further processing with feature propagation, MeteorNet-flow obtains the per-point flow vector for every point in frame T. We added skip connections so that local features at early stages of the network can be used in feature propagation. The architecture of MeteorNet-flow is illustrated in Figure 7.

E. Model Run Time Analysis

We use MeteorNet-seg on the Synthia semantic segmentation test set for runtime analysis. We tested with 8,192 points for the whole scene in each frame. We used a single GTX 1080 Ti GPU and Intel Core i7 CPU. The deep learning framework is Tensorflow 1.9.0. We performed a grid search over batch size and number of frames. The results for direct grouping and chained-flow grouping are illustrated in Figure 5 and Figure 6 respectively.

Interpolating and chaining flow introduces an additional computational overhead for chained-flow grouping. For 2 frames and a batch size of 1, MeteorNet-seg with direct grouping runs at 8.0 sequences per second (seq/s); MeteorNet-seg with chained-flow grouping runs at 4.1 seq/s. For 4 frames and batch size of 1, MeteorNet-seg with direct grouping runs at 4.3 seq/s; MeteorNet-seg with chained-flow grouping runs at 2.8 seq/s.

F. Proof of Theorem

Suppose $\forall t, \mathcal{X}_t = \{S_t \mid S_t \subseteq [0,1]^m, |S_t| = n, n \in \mathbb{Z}^+\}$ is the set of m-dimensional point clouds inside an m-dimensional unit cube at time $t \in \mathbb{Z}$. We define single-frame Hausdorff distance $d_H(S_i, S_j)$ for $S_i \in \mathcal{X}_i$ and $S_j \in \mathcal{X}_j$. $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_T$ is the set of point cloud sequences of length T. Suppose $f : \mathcal{X} \to \mathbb{R}$ is a continuous function on $\mathcal{X}$ w.r.t $d_{seq}(\cdot,\cdot)$, i.e. $\forall \epsilon > 0$, $\exists \delta > 0$, for any $S, S' \in \mathcal{X}$, if $d_{seq}(S, S') < \delta$, $|f(S) - f(S')| < \epsilon$. Here, we define the distance of point cloud sequences $d_{seq}(\cdot,\cdot)$ as the maximum per-frame Hausdorff distance among all respective frame pairs, i.e. $d_{seq}(S, S') = \max_t\{d_H(S_t, S'_t)\}$. Our theorem says that f can be approximated arbitrarily closely by a large-enough neural network and a max pooling layer with enough neurons.

We first have the following lemma from the supplementary material of [3], which ensures the universal approximation potential of PointNet.

Lemma 1. *Suppose $f : \mathcal{X} \to \mathbb{R}$ is a continuous set function w.r.t Hausdorff distance $d_H(\cdot,\cdot)$. $\forall \epsilon > 0$, $\exists$ continuous*

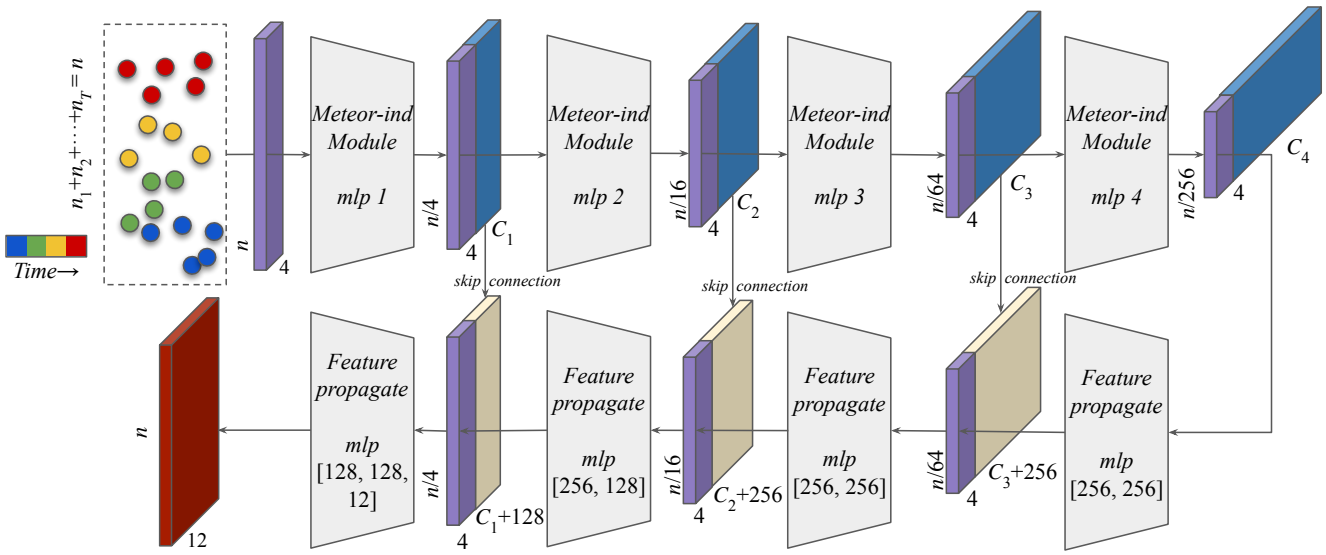

Figure 4: The architecture of **MeteorNet-seg**. The widths of "mlp {1,2,3,4}" for different configurations are listed in Table 2.

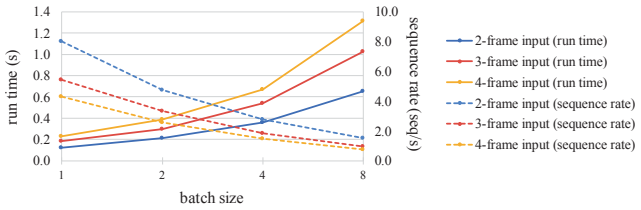

Figure 5: Run time and frame rate of Meteor-seg with direct grouping on Synthia test set.

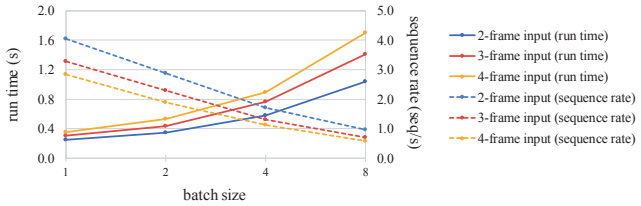

Figure 6: Run time and frame rate of Meteor-seg with chained-flow grouping on Synthia test set.

function η and γ such that for any $S \in \mathcal{X}$,

$$\left| f(S) - \gamma \circ \left(\underset{x \in S}{MAX} \{\eta(x)\} \right) \right| < \epsilon$$

where MAX is a vector max operator that takes a set of vectors as input and returns a new vector of the element-wise maximum.

Our theorem is proved based on Lemma 1. The core idea is that we can map the point cloud sequence indexed by t into the single point cloud space.

Theorem 1. *Suppose $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_T \to \mathbb{R}$ is a continuous function w.r.t $d_{seq}(\cdot,\cdot)$. $\forall \epsilon > 0$, $\exists$ a continuous function $\zeta(\cdot,\cdot)$ and a continuous function γ, such that for any $S = (S_1, S_2, \ldots, S_T) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_T$,*

$$\left| f(S) - \gamma \circ \left(\underset{\mathbf{x}_i^{(t)} \in S_t, t \in \{1,2,\ldots,T\}}{MAX} \{\zeta(\mathbf{x}_i^{(t)}, t)\} \right) \right| < \epsilon$$

where MAX is a vector max operator that takes a set of vectors as input and returns a new vector of the element-wise maximum.

Proof. It suffices to prove for $m = 1$.

In the following proof, we use plain x_i instead of bold $\mathbf{x}_i$ to represent scalar value instead of a 3-D vector.

Let $\mathcal{T} = \{S \mid S \subseteq [0,1], |S| = n\}$. Define function $\psi : \mathcal{X} \to \mathcal{T}$ as

$$\psi(S_1, \ldots, S_T) = \{p_T(x_{i_t}, t) \mid x_{i_t} \in S_t, t \in \{1, \ldots, T\}\}$$

where $p_T(x,t) = \frac{x+t-1}{T}$ is a function that maps each of the T $[0,1]$ intervals into a unique place inside $[0,1]$ interval. Notice that $\mathcal{T} = \mathcal{X}_t$, so d_H can also be defined on $\mathcal{T}$.

For any $S \in \mathcal{X}$, $\forall \epsilon' > 0$, $\exists \delta = \epsilon' T$, such that $\forall S', d_{seq}(S, S') < \delta$, we have

$$\begin{aligned}
&d_H(\psi(S), \psi(S')) \\
&= d_H(\psi(S_1, \ldots, S_T), \psi(S'_1, \ldots, S'_T)) \\
&= \max_t \{ \sup_{x \in S_t} \inf_{y \in S'_t} d(p_T(x,t), p_T(y,t)), \\
&\sup_{y \in S'_t} \inf_{x \in S_t} d(p_T(y,t), p_T(x,t)) \} \\
&= \max_t \{ \sup_{x \in S_t} \inf_{y \in S'_t} \frac{1}{T} d(x,y), \sup_{y \in S'_t} \inf_{x \in S_t} \frac{1}{T} d(y,x) \} \\
&= \frac{1}{T} \max_t \{ \sup_{x \in S_t} \inf_{y \in S'_t} d(x,y), \sup_{y \in S'_t} \inf_{x \in S_t} d(y,x) \} \\
&= \frac{1}{T} \max_t d_H(S_t, S'_t) = \frac{1}{T} d_{seq}(S, S') < \frac{1}{T} \delta = \epsilon'
\end{aligned}$$

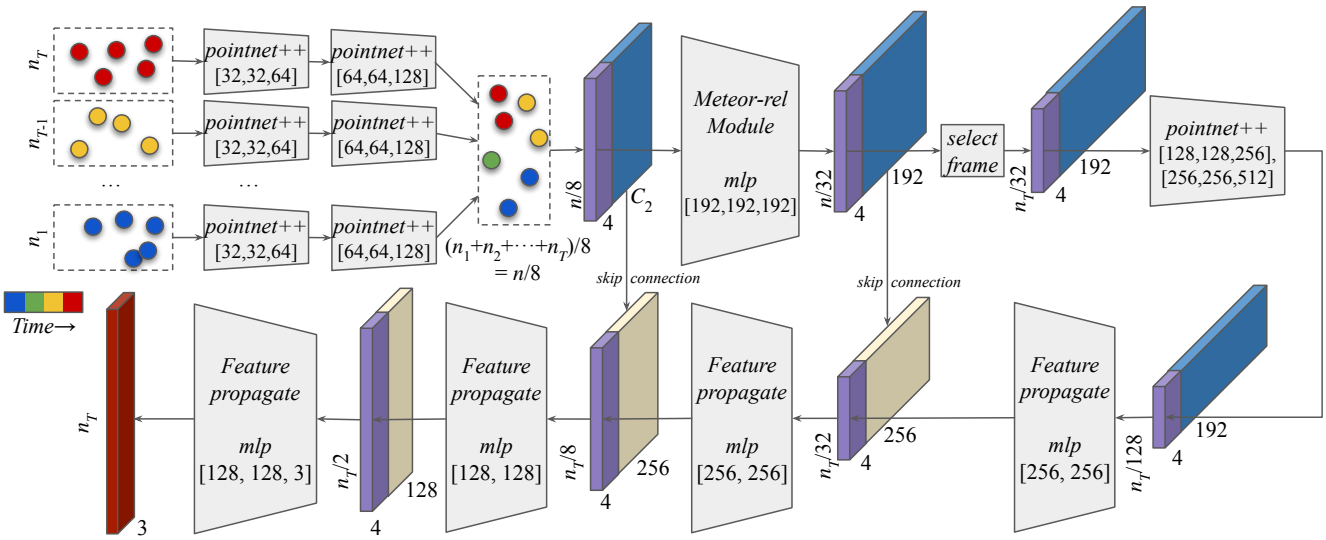

Figure 7: The architecture of **MeteorNet-flow**.

So ψ is a continuous function w.r.t. $d_H : \mathcal{X} \to \mathbb{R}$ and $d_{seq} : \mathcal{X}' \to \mathbb{R}$. It's easy to show that the inverse of ψ is also a continuous function.

According to Lemma 1, $\forall \epsilon > 0$, $\exists$ continuous function η and γ such that for any $\psi(S) \in \mathcal{T}$,

$$\begin{aligned}
&\left| f(S) - \gamma \circ \left(\underset{x \in \psi(S)}{MAX} \{\eta(x)\} \right) \right| \\
&= \left| f(S) - \gamma \circ \left(\underset{x_{i_t} \in S_t, t \in \{1,\dots,T\}}{MAX} \{\eta(p_T(x_{i_t}, t))\} \right) \right| \\
&= \left| f(S) - \gamma \circ \left(\underset{x_{i_t} \in S_t, t \in \{1,\dots,T\}}{MAX} \{\zeta(x_{i_t}, t)\} \right) \right| < \epsilon
\end{aligned}$$

where ζ is defined as $\zeta(\cdot, t) = \eta(p_T(\cdot, t))$.

This concludes the proof. □

G. More Visualization

G.1. Synthia

We provide additional qualitative results for segmentation results on the Synthia test set in Figure 9. Again, MeteorNet-seg can accurately segment most objects.

G.2. KITTI scene flow

We provide additional qualitative results for scene flow estimation results on KITTI scene flow dataset in Figure 8. Again, MeteorNet-flow can accurately estimate flow for moving objects.

H. Name Metaphor

The universe is all of space and time. When we look deep into the universe, stars are the visible points in the sky. Meteor shower is a group of stars that move together as a "dynamic point cloud sequence". It brings fortune and good luck to anyone who sees it.

We hope our MeteorNet can also bring fortune and good luck to our readers and benefit related research domains.

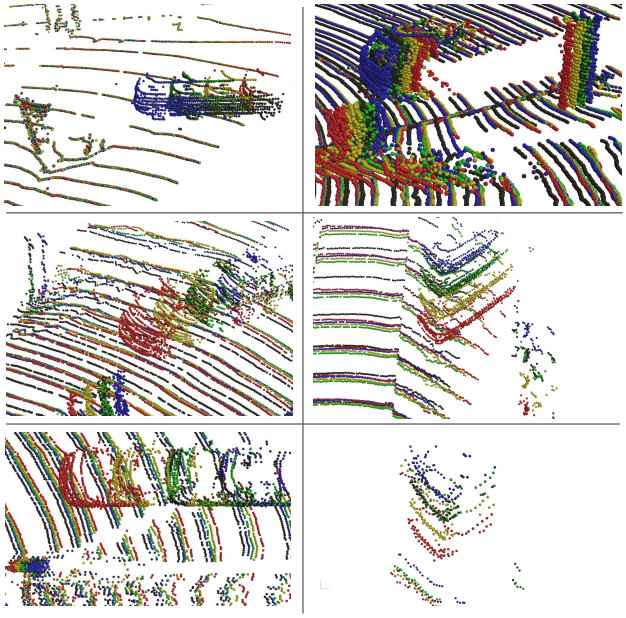

Figure 8: Additional visualization of MeteorNet example results on the KITTI scene flow dataset. Point are colored to indicate which frames they belong to: frames $t-3$, frame $t-2$, frame $t-1$, frame t. **Translated points** (frame $t-3$ + estimated scene flow) is in black. Green and black shapes are supposed to overlap for perfect estimation.

References

[1] Xingyu Liu, Charles. R. Qi, and Leonidas J. Guibas. Flownet3d: Learning scene flow in 3d point clouds. In *CVPR*, 2019. 1

[2] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 1

[3] Charles R. Qi, H. Su, Kaichun Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 3

[4] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*. 2017. 3

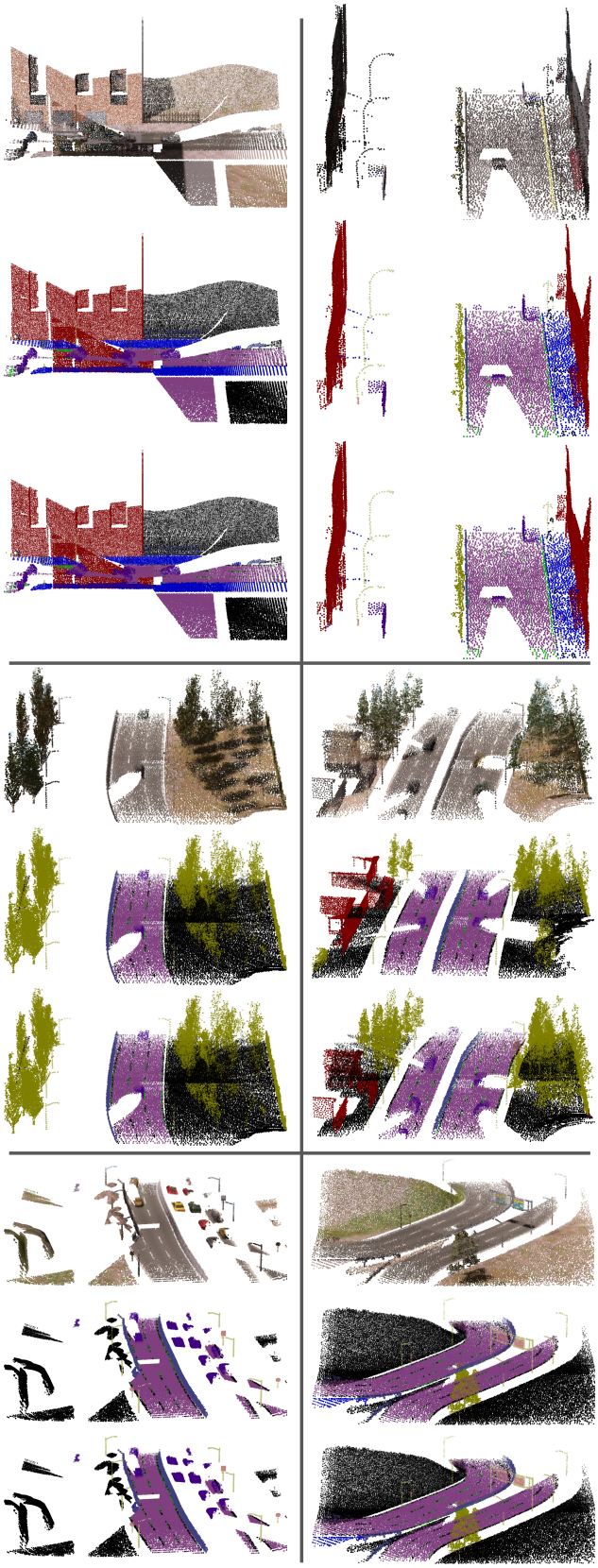

Figure 9: Visualization of two example results from the Synthia dataset. Each cell from the top: RGB input, ground truth, predictions.