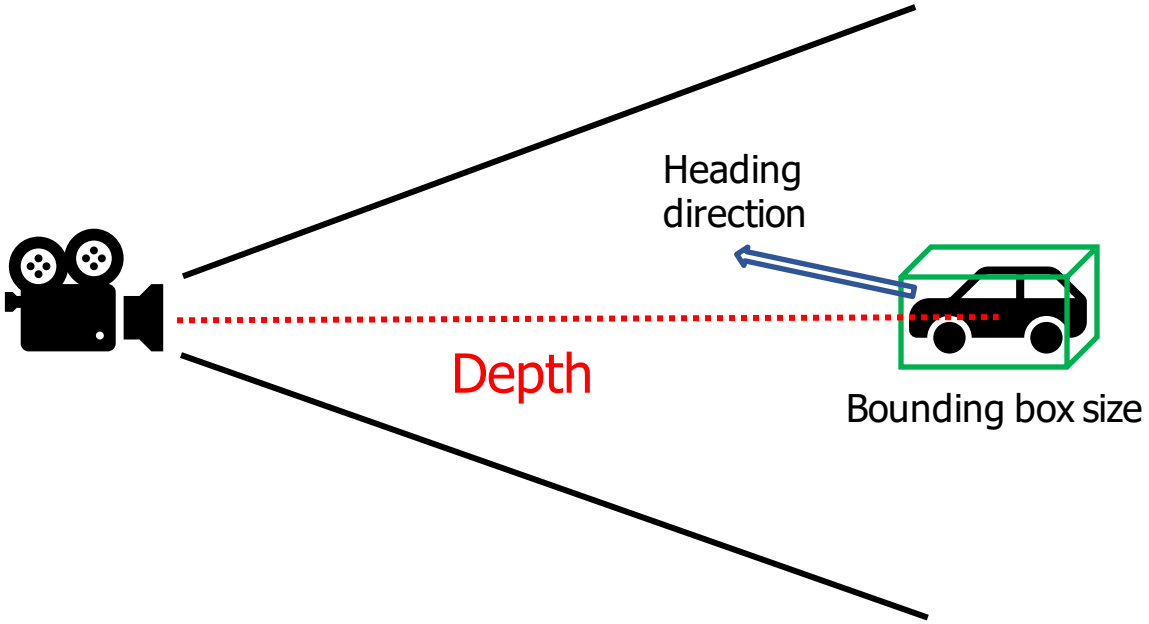
Depth-discriminative Metric Learning for Monocular 3D Object Detection

Wonhyeok Choi* Mingyu Shin* Sunghoon Im†

DGIST

Depth-discriminative Metric Learning for Monocular 3D Object Detection

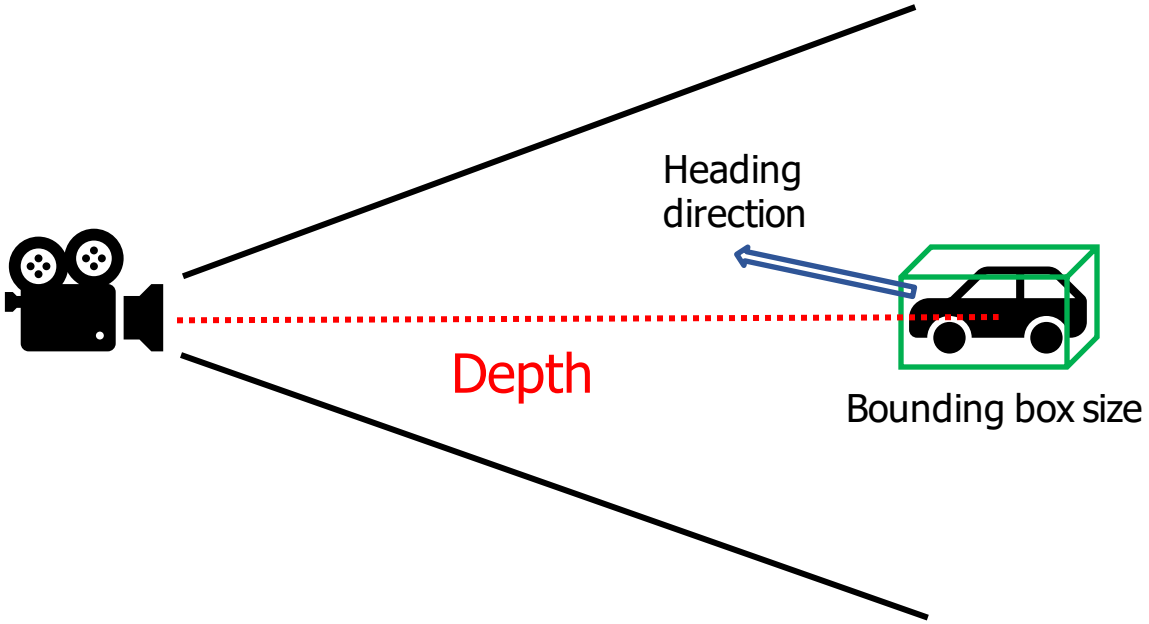


Metric: 3D AP on KITTI validation set (IoU: 0.7)

replace factors with gt values	original
Baseline (MonoDLE)	11.3
(a) with BB size	13.3
(b) with projected3D	12.0
(c) with yaw angle	11.8
(d) with 3D location	77.6
(e) with depth	67.0

[1]

Depth-discriminative Metric Learning for Monocular 3D Object Detection



Metric: 3D AP on KITTI validation set (IoU: 0.7)

replace factors with gt values	original
Baseline (MonoDLE)	11.3
(a) with BB size	13.3
(b) with projected3D	12.0
(c) with yaw angle	11.8
(d) with 3D location	77.6
(e) with depth	67.0

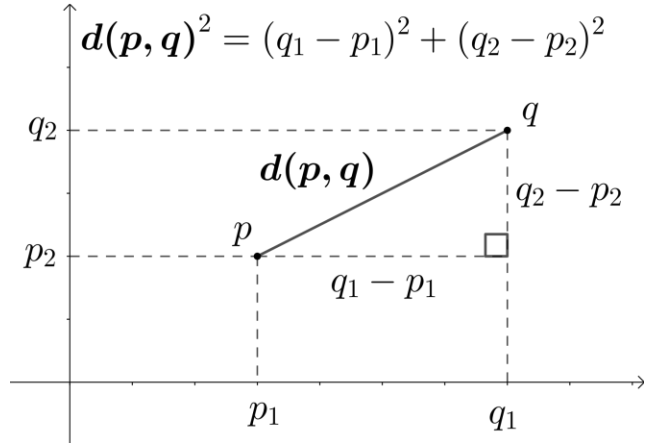
[1]

How to enhance the depth accuracy in monocular 3D object detection framework?
-> Use **Metric learning** that encourages the model to extract **Depth-discriminative** features.

Preliminary

Metric space. A metric space is a mathematical concept that characterizes a set of points and a function that measures the distance between any two points in the set. Formally, a metric space can be defined as a pair (M, d), where M represents a set and d is a distance function on M. The distance function d must satisfies the following axioms [13] for any three points $x, y, z \in M$:

1. Non-negativity: $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$.
2. Symmetry: $d(x, y) = d(y, x)$ for all $x, y \in M$.
3. Triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in M$.



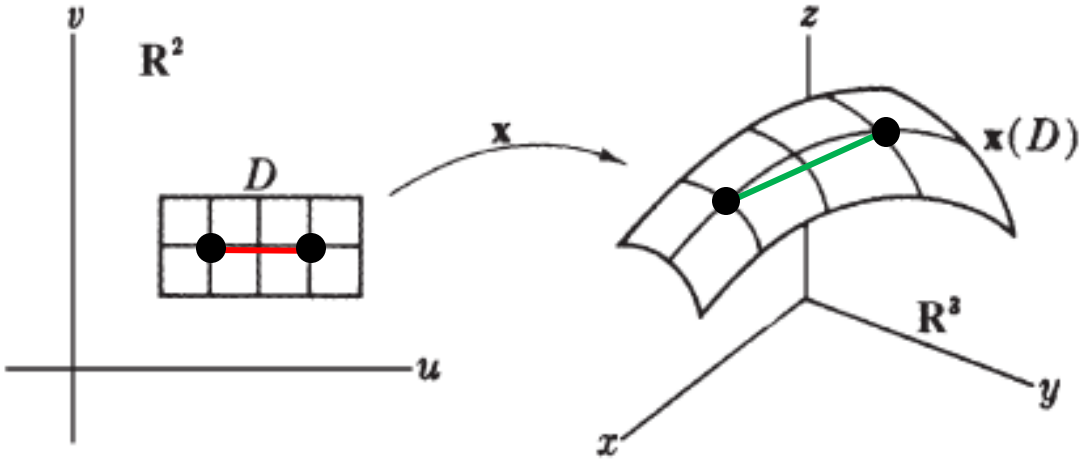
e.g.) Euclidean space (M, d)

- M: R^n space
- d(p,q) = ||p,q||

Preliminary

Quasi-isometry. A quasi-isometry is a function between two metric spaces that preserves distances up to a constant factor, even though it may locally distort angles and distances. Let $\mathcal{Q}$ be a function from one metric space (M_1, d_1) to another metric space (M_2, d_2). $\mathcal{Q}$ is considered a quasi-isometry from (M_1, d_1) to (M_2, d_2) if there exist constants $K \geq 1$, $B \geq 0$, and $\epsilon \geq 0$ such that both of the following properties hold:

1. $\forall x_1, x_2 \in M_1 \ : \ \frac{1}{K} \cdot d_1(x_1, x_2) - B \leq d_2(\mathcal{Q}(x_1), \mathcal{Q}(x_2)) \leq K \cdot d_1(x_1, x_2) + B.$
2. $\forall z \in M_2 \ : \ \exists x \in M_1 \ s.t. \ d_2(z, \mathcal{Q}(x)) \leq \epsilon.$



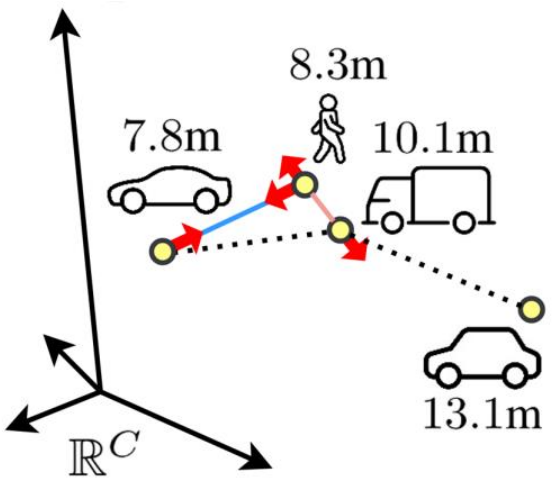
$\frac{1}{K} \cdot$ [—] $- B \leq$ [—] $\leq K \cdot$ [—]

Goal: Make Depth-discriminative feature space by using Quasi-isometry.

Key idea 1: Depth classifier can easily discriminate object features that roughly preserve the depth information

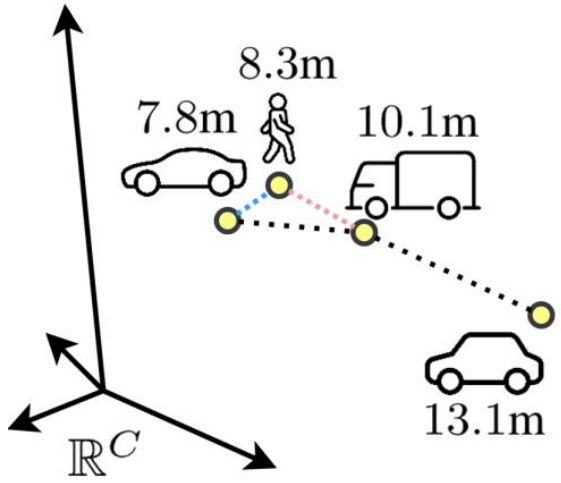
Key idea 2: We can learn the the features by contrasting the distance of each object feature with respect to the quasi-isometry condition.

Feature space

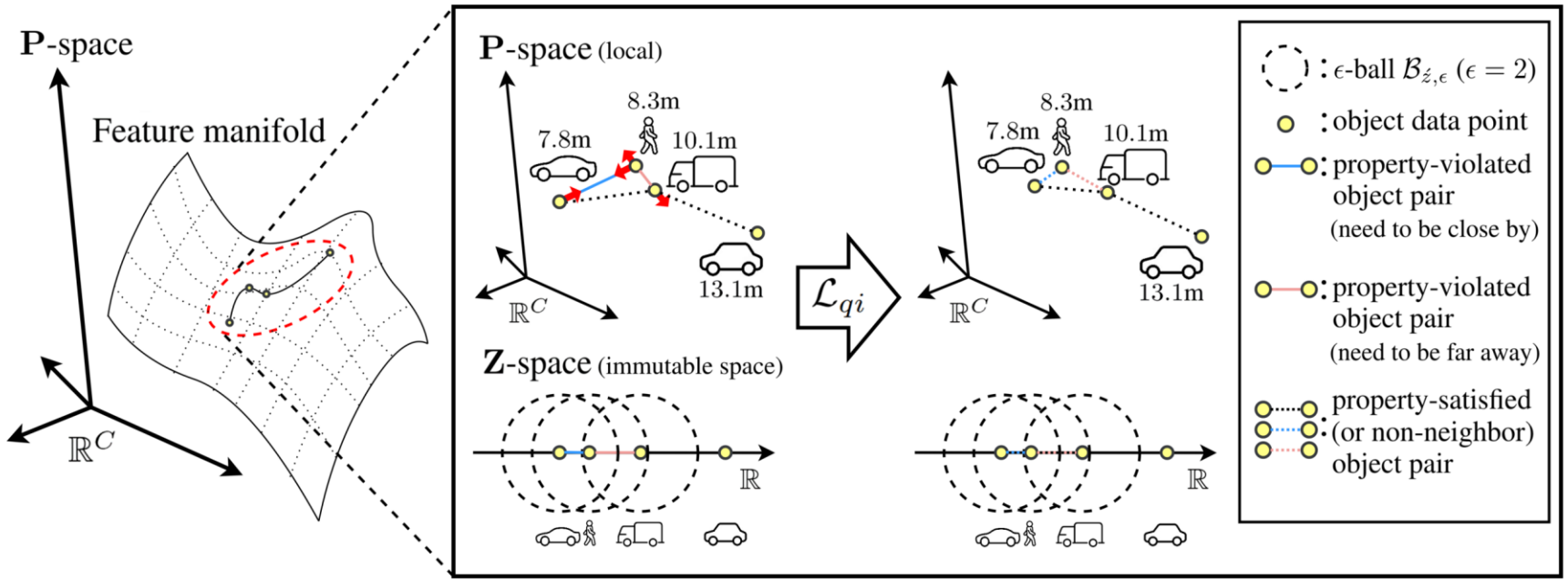


Contrast

Depth-discriminative feature space



Proposed Method 1. Quasi-isometric loss



Summary of high-level process of quasi-isometric loss:

1. Extract the feature descriptor from each object
2. Find the feature pair that not satisfied the quasi-isometric property
3. Train with the proposed quasi-isometric loss.

$$\frac{1}{K} \cdot \left[\text{———}\right] - B \leq \left[\text{———}\right] \leq K \cdot \left[\text{———}\right]$$

$$\mathcal{L}_{qi} = -\frac{1}{|\mathbf{P}_b^+|} \sum_{(\rho_i,\rho_j)\in\mathbf{P}_b^+} \log \frac{\mathrm{S}^+(\rho_i,\rho_j)}{\mathrm{S}^+(\rho_i,\rho_j) + \sum_{(\rho_k,\rho_l)\in\mathbf{P}_b^-} \mathrm{S}^-(\rho_k,\rho_l)},$$

Proposed Method 2. Object-wise depth map loss

$$\mathcal{L}_{obj} = \frac{1}{|\mathfrak{D}|} \sum_{z^p \in \mathfrak{D}} \frac{\sqrt{2}}{\hat{\sigma}^p} |z^p - \hat{z}^p| + \log(\hat{\sigma}^p),$$

Add Auxiliary Task: projected object center depth supervision with entire bounding box region of the object.

-> To reduce the localization error.

Results

KITTI

Extra data	Method	Car, $AP_{3D\|R40\|Mod.}$	Ped, $AP_{3D\|R40\|Mod.}$	Cyc, $AP_{3D\|R40\|Mod.}$
LiDAR	DID-M3D	16.29	-	-
	DID-M3D + Ours	16.42 (+0.8%)	9.05 (-)	3.11 (-)
None	MonoDLE	12.26	6.55	2.66
	MonoDLE + Ours	15.30 (+24.8%)	7.80 (+19.1%)	4.12 (+54.9%)
	GUPNet	14.20	9.53	2.56
	GUPNet + Ours	15.78 (+11.1%)	9.03 (-5.2%)	3.61 (+41.0%)
	MonoCon	16.46	8.41	1.92
	MonoCon + Ours	16.36 (-0.6%)	10.28 (+22.2%)	2.89 (+50.5%)

+23.51%

Waymo Open Dataset

Difficulty	Method	Vehicle, $AP_{3D\|R40\|overall}$	Vehicle, APH_{3D}
Level_1 (IoU = 0.7)	MonoCon	2.30	2.29
	MonoCon + Ours	2.50 (+8.7%)	2.48 (+8.3%)
Level_2 (IoU = 0.7)	MonoCon	2.16	2.15
	MonoCon + Ours	2.34 (+8.3%)	2.33 (+8.4%)

+5.78%

Results

Extensive method: Anchor-based / Bird-eye-view paradigm

Method	Car, $AP_{3D\|R40\|Mod.}$
MonoDTR	18.45
MonoDTR + Ours	19.07 (+3.4%)
ImVoxelNet	17.80
ImVoxelNet + Ours	18.20 (+7.2%)

Visualized feature space

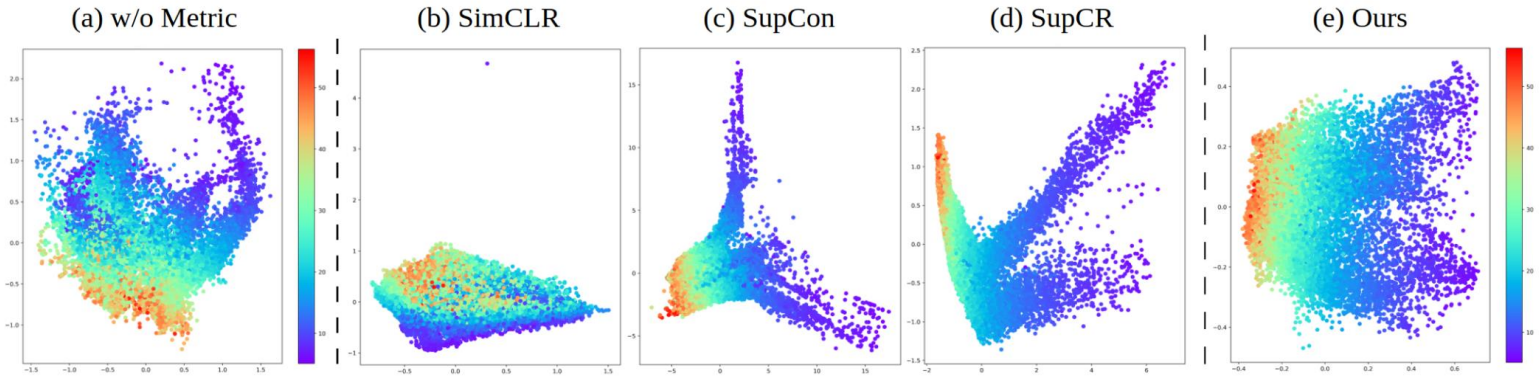


Figure 1: Visualization results (PCA) of the learned feature space with various contrastive methods.

Thank you!