Marginalized Generalized IoU (MGIoU): A Unified Objective Function for Optimizing Convex Parametric Shapes

Duy-Tho Le[1], **Trung Pham**[2], **Jianfei Cai**[1], **Hamid Rezatofighi**[1]
[1]Monash University
[2]NVIDIA

Abstract

Optimizing the similarity between parametric shapes is crucial for numerous computer vision tasks, where Intersection over Union (IoU) stands as the canonical measure. However, existing optimization methods exhibit significant shortcomings: regression-based losses like L1/L2 lack correlation with IoU, IoU-based losses are unstable and limited to simple shapes, and task-specific methods are computationally intensive and not generalizable across domains. As a result, the current landscape of parametric shape objective functions has become scattered, with each domain proposing distinct IoU approximations. To address this, we unify the parametric shape optimization objective functions by introducing **Marginalized Generalized IoU (MGIoU)**, a novel loss function that overcomes these challenges by projecting structured convex shapes onto their unique shape Normals to compute one-dimensional normalized GIoU. MGIoU offers a simple, efficient, fully differentiable approximation strongly correlated with IoU. We extend MGIoU to **MGIoU**$^+$ that supports optimizing unstructured convex shapes. Together, MGIoU and MGIoU$^+$ unify parametric shape optimization across diverse applications. Experiments on standard benchmarks demonstrate that MGIoU and MGIoU$^+$ demonstrate higher performance while reducing loss computation latency up to 10-40x. Also, MGIoU and MGIoU$^+$ satisfy metric properties and scale-invariance, ensuring robustness as an objective function. We further propose **MGIoU**$^-$ for minimizing overlaps in tasks like collision-free trajectory prediction.

Code — https://ldtho.github.io/MGIoU/

Introduction

Parametric shape optimization is a core problem in computer vision, robotics, and graphics, with applications spanning 2D/3D object detection (Redmon et al. 2016; Yang et al. 2021c,b; Bi and Hu 2021), 6D pose estimation (Brazil et al. 2023; Li et al. 2024), shape registration, and trajectory prediction (Sun et al. 2020; Shi et al. 2022, 2023). This optimization seeks to maximize similarity between predicted and ground-truth shapes, commonly quantified using Intersection over Union (IoU). Despite IoU's popularity as a standardized and scale-invariant metric, directly optimizing IoU (or GIoU (Rezatofighi et al. 2019)) for arbitrary convex shapes presents significant challenges. Computing IoU/-GIoU analytically for complex shapes like 3D ellipsoids is non-trivial, and even for simpler shapes, using IoU/GIoU as an objective function remains computationally expensive with unstable gradients. Consequently, aside from specific cases like axis-aligned boxes (Rezatofighi et al. 2019), IoU/GIoU-based optimization has not been widely adopted for general shape optimization.

As a result, objective functions for parametric shape optimization remain fragmented across domains. Applications either: *(i)* use simple regression losses (e.g., L-norms), OKS (Maji et al. 2022), or Chamfer distance (Qi et al. 2017), or *(ii)* employ IoU/GIoU approximations, such as Gaussian-based models for rotated boxes (Yang et al. 2021a,b, 2022) and vertex-based methods for quadrilaterals (Bi and Hu 2021; He et al. 2018; Liu and Jin 2017; Liao, Shi, and Bai 2018). However, these lack direct correlation with true IoU/GIoU and may violate properties like scale invariance or metric consistency. Many methods also require extensive task-specific tuning (Yang et al. 2021b,c), ultimately yielding suboptimal or overfit solutions. Currently, no unified objective function works robustly across shape parameterizations while maintaining strong correlation with standard IoU-based metrics. This fragmentation particularly affects shape-alignment tasks that rely on IoU as their evaluation metric, spanning detection, localization, and visual grounding, serving as building blocks for higher-level applications.

We introduce **Marginalized Generalized IoU (MGIoU)**, a novel geometric loss function for flexible and stable optimization of arbitrary convex parametric shapes. MGIoU simplifies complex shape overlap calculations by projecting shapes onto their normals (e.g., edge normals in 2D, face normals in 3D) and marginalizing 1D GIoU operations. This enables differentiable overlap optimization even for non-overlapping shapes. MGIoU unifies shape optimization into a single coherent loss term applied directly to shape normals, enabling holistic adjustment of shape vertices and parameters (e.g., position, size, and orientation) without requiring balancing of multiple separate loss terms. Its simplicity allows integration into existing pipelines as a drop-in replacement across applications such as 2D oriented object detection, 3D shape estimation, polygonal shape fitting, and diffusion-based 6D grasp detection. The mathemat-

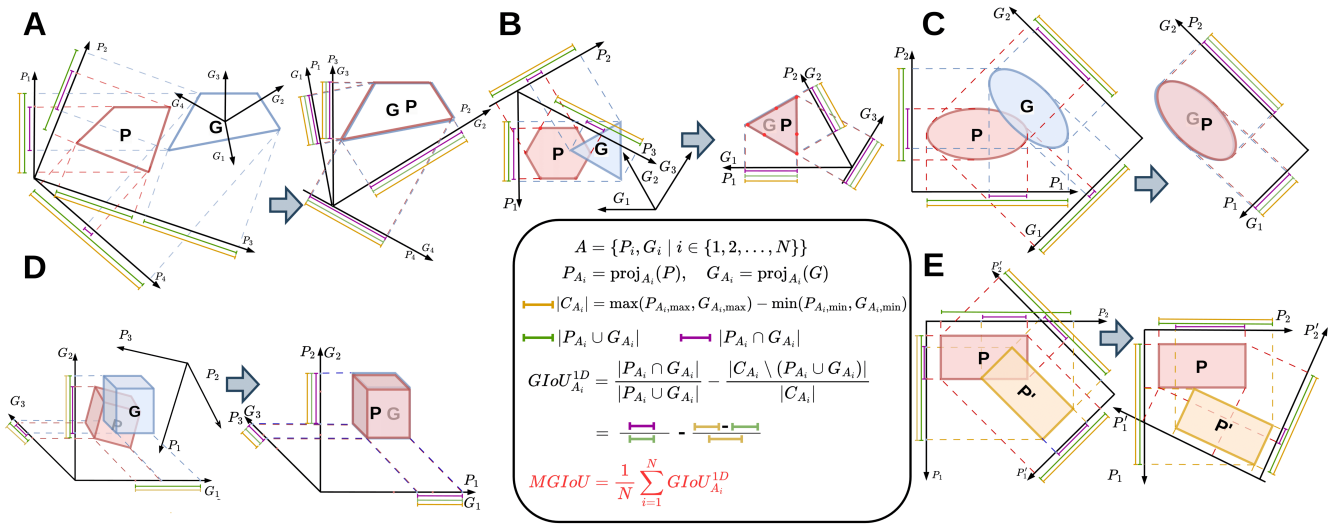

Figure 1: MGIoU and its variants computation. Predicted (P, red) and ground-truth (G, blue) shapes are projected onto their unique Normals (P_i, G_i) to calculate one-dimensional overlaps, measuring intersection (purple segments) relative to union (green segments). To reduce visual clutter, some panels depict projections onto unique Normals of only one shape. Examples: **MGIoU**$^+$[**A)** unstructured quadrilaterals, **B)** general polygons], **MGIoU**[**C)** ellipses, **D)** 3D cuboids], and **MGIoU**$^-$[**E)** rotated rectangles]. Best viewed zoomed in.

ical construction of MGIoU ensures compliance with scale-invariance and metric properties, reinforcing its theoretical soundness.

We propose three variations of this objective function, making MGIoU applicable across a wide range of tasks. The primary version, MGIoU, is designed for optimizing *structured* convex shapes, where both the source and target shapes share the same parametric domain, e.g. both being rectangles, ellipses, cuboids etc. (see fig. 1 (C, D)).

Second, an extended version, denoted as **MGIoU**$^+$, supports optimization between *unstructured* convex shapes, where the source and target may belong to different shape families or have differing numbers of vertices, such as convex polygons or polyhedra with varying and arbitrary vertex and edge counts (see fig. 1 (A, B) as two examples). To accommodate this flexibility, we use a vertex-based parametrization for both shapes, under the condition that the target shape always has more vertices than the source. In MGIoU$^+$, we introduce a convexity regularizer to ensure predicted shapes remain convex during optimization.

Third, we propose **MGIoU**$^-$, a complementary loss designed for minimizing overlap between arbitrary convex shapes. This is useful for trajectory prediction, where the goal is reducing collision risk. Unlike standard alignment objectives, MGIoU$^-$ directly minimizes IoU between predicted and reference shapes. Using the same projection mechanism to penalize overlaps, encouraging safer trajectory predictions. We extend MGIoU$^-$ to handle shape sequences, enabling *spatio-temporal* optimization for forecasting applications.

To validate our approach, we evaluate MGIoU and its variants across multiple applications on standard benchmarks. Our results demonstrate improved performance over traditional losses with reduced computational overhead.
Our key contributions can be summarized as follows:

- We propose **MGIoU**, a novel geometric loss function for optimizing arbitrary convex parametric shapes in any dimension. MGIoU simplifies complex shape overlap calculations into efficient 1D GIoU operations by projecting shapes onto their normals, unifying position, dimension, and orientation into a single differentiable objective.
- We extend MGIoU to **MGIoU**$^+$ for *unstructured* convex shapes with different parametric structures, and **MGIoU**$^-$ for minimizing shape overlap in applications such as collision-free trajectory prediction.
- We conduct experiments across diverse tasks, including 2D/3D shape alignment, oriented object detection, trajectory prediction, and 6D grasp detection, demonstrating MGIoU's effectiveness while reducing computational overhead.

Related Works

Loss functions have evolved across vision and robotics tasks, including object detection, 6-DoF pose estimation, and trajectory forecasting. This section reviews these developments, highlighting limitations that motivate MGIoU.

2D Axis-Aligned Object Detection. Early models (Redmon et al. 2016; Redmon and Farhadi 2017; Ren et al. 2015) used L1/L2 regression losses, which lack IoU correlation and lead to suboptimal overlap (Rezatofighi et al. 2019). IoU-based losses emerged with GIoU (Rezatofighi et al. 2019) introducing differentiable penalties for non-overlapping cases, followed by DIoU (Zheng et al. 2020) and CIoU (Zheng et al. 2021) incorporating distance and aspect ratio terms. However, these remain limited to axis-aligned geometries.

2D Oriented Object Detection. For aerial imagery and text recognition, specialized losses address rotation challenges. Methods include modulated rotation loss (Qian et al. 2021) for angle periodicity, and KFIoU (Yang et al. 2022), GWD (Yang et al. 2021b), and KLD (Yang et al. 2021c) modeling boxes as Gaussian distributions. SIoU (Yang et al. 2021a) targets skewed geometries. These approaches are computationally intensive and rotation-specific.

2D Quadrilateral Detection. Document analysis methods like Quadbox (Keserwani, Singh, and Shukla 2021) use vertex regression with L1 losses, Textboxes++ (Liao, Shi, and Bai 2018) employs rectangular distances for vertex ordering, and QRN (He et al. 2018) sorts vertices by polar angles. These solutions are hand-crafted and lack IoU correlation.

3D Object Recognition & 6-DoF Object Pose Estimation. 3D detection requires handling position, dimensions, and orientation simultaneously. While early approaches assumed 1-DoF rotation (yaw), recent work addresses 3-DoF rotations for VR (Brazil et al. 2023; Ahmadyan et al. 2021; Roberts et al. 2021; Baruch et al. 2021) and 6-DoF pose estimation (Chang et al. 2015; Dai et al. 2017). SO3 losses (Su et al. 2020) operate on rotation manifolds, while Chamfer distance (Qi et al. 2017) measures point set distances. However, both lack IoU correlation, limiting optimization of spatial-rotational alignment.

Collision-aware Penalties Losses. Trajectory prediction methods incorporate collision avoidance penalties through auxiliary losses like off-road loss (Niedoba et al. 2019), LaneLoss (Kim et al. 2022), and RouteLoss (Zhang et al. 2022). However, these primarily address static-ego interactions, neglecting multi-agent dynamics. TrafficSim (Suo et al. 2021) approximates objects as circles for collision avoidance but lacks precise boundary representation.

Diffusion-based 6D Grasp Detection. 6-DoF grasp detection involves predicting pose and grasp configurations simultaneously. Traditional methods use separate L1/L2 losses for translation and rotation, facing discontinuities and misalignment with grasp success metrics. Recent diffusion approaches (Nguyen et al. 2024a; Vuong et al. 2024; Nguyen et al. 2024b) introduce Gaussian noise to parameters and predict it using L2 loss, but they lack alignment with grasp coverage and geometry comprehension.

Summary and Motivation for MGIoU. Existing loss functions exhibit trade-offs between accuracy, efficiency, and generalizability. Regression losses lack IoU alignment, while IoU-based losses struggle with non-axis-aligned geometries. Task-specific solutions are computationally expensive and narrowly focused. MGIoU addresses these limitations through a unified, efficient approach using 1D projections to optimize spatial relationships across diverse tasks.

Methodology

We introduce **Marginalized Generalized IoU (MGIoU)**, a novel loss function designed to optimize shape alignment for *structured* convex shapes in object detection tasks, along with its variants: **MGIoU$^+$** for *unstructured* convex shapes and **MGIoU$^-$** for minimizing overlaps. The key idea is to project vertices of convex shapes onto their normals, computing GIoU in 1D projection space. This enables MGIoU to generalize across 2D rotated, 3D 6-DoF detection and grasping tasks, unifying position, size, and orientation optimization in a single, differentiable loss term. MGIoU$^+$ extends this to unstructured shapes with a convexity regularizer, quantified on quadrilateral object detection task, while MGIoU$^-$ adapts it to penalize overlaps, enhancing collision avoidance in trajectory prediction.

Identifying Shape Normals

Given two arbitrary shapes $P, G \subseteq \mathbb{R}^D$ ($D = 2$ for 2D, $D = 3$ for 3D), with vertices $\mathbf{P} \in \mathbb{R}^{N_P \times D}$ and $\mathbf{G} \in \mathbb{R}^{N_G \times D}$, where N_P and N_G are the numbers of vertices for P and G respectively, MGIoU begins by constructing a set of unique directional Normals $\mathcal{A}$. In 2D polygons, these Normals originate from edges, computed as 90-degree rotations of consecutive vertex difference vectors. For ellipses or ellipsoids, the normals align with semi-axes directions. In 3D shapes, these Normals correspond to face normals derived from the surface geometry. Extremely regular shapes (e.g., rectangles in 2D, cuboids in 3D, see fig. 1) typically produce duplicate or parallel Normals; these redundancies are eliminated, leaving only a small number of unique directional Normals that capture essential geometric attributes. For example, a 2D rotated rectangle ($N = 4$) has just two unique Normals; a 3D cuboid ($N = 8$) only three. This reduces computational complexity.

MGIoU and MGIoU$^+$ for Structured and Unstructured Shapes

MGIoU$^-$ for Minimizing Shape Overlaps

For structured convex shapes (e.g., rectangles, cuboids), MGIoU maximizes overlap between the predicted shape P and ground truth G by projecting their vertices onto $\mathcal{A}$. The GIoU^{1D} is computed per Normal $a_i \in \mathcal{A}$ using a simplified version of GIoU for 1D (see Appendix), and then averaged across normals MGIoU $= \frac{1}{|\mathcal{A}|} \sum_{\mathbf{a} \in \mathcal{A}} \text{GIoU}^{1D}_{\mathbf{a}}$, finally $\mathcal{L}_{\text{MGIoU}} = \frac{1-\text{MGIoU}}{2}$.We have $\mathcal{L}_{\text{MGIoU}} = 0$ for perfect overlap (MGIoU $= 1$), $\mathcal{L}_{\text{MGIoU}} \to 1^-$ as shapes separate (MGIoU $\to -1^+$). As shown in algorithm 1, the MGIoU is positive when projections overlap and negative when they are disjoint, resulting in MGIoU $\in (-1, 1]$, and the final loss $\mathcal{L}_{\text{MGIoU}} \in [0, 1)$. Given its connection to the Jaccard index and the properties of GIoU, $\mathcal{L}_{\text{MGIoU}}$ satisfies key metric properties (See Appendix):

1. **Non-negativity**: $\mathcal{L}_{\text{MGIoU}}(P, G) \geq 0$
2. **Identity**: $\mathcal{L}_{\text{MGIoU}}(P, G) = 0$ if and only if $P = G$
3. **Symmetry**: $\mathcal{L}_{\text{MGIoU}}(P, G) = \mathcal{L}_{\text{MGIoU}}(G, P)$
4. **Triangle Inequality**: $\mathcal{L}_{\text{MGIoU}}(P, R) \leq \mathcal{L}_{\text{MGIoU}}(P, Q) + \mathcal{L}_{\text{MGIoU}}(Q, R)$
5. **Scale Invariance**: $\mathcal{L}_{\text{MGIoU}}(sP, sG) = \mathcal{L}_{\text{MGIoU}}(P, G)$ for any scalar $s > 0$

These properties enhance $\mathcal{L}_{\text{MGIoU}}$ correlation with IoU by measuring shape overlap in projected space, making it reliable and consistent.

For unstructured convex shapes (e.g., polygons with varying vertices), we apply GIoU^{1D} to all shape Normals and

Algorithm 1: MGIoU and MGIoU$^+$ - Maximize Shape Overlap

input : Predicted vertices $\mathbf{P} \in \mathbb{R}^{N_P \times D}$, ground truth vertices $\mathbf{G} \in \mathbb{R}^{N_G \times D}$, convexity weight λ, shape type (structured or unstructured)
output: $\mathcal{L}_{\text{total}}$

Compute unique normals $\mathcal{A}$ of $\mathbf{P}$ and $\mathbf{G}$
Initialize overlap sum $S = 0$
for *each unique normal* $\mathbf{a}_i \in \mathcal{A}$ **do**
  $\text{proj}_{\mathbf{P},\mathbf{a}_i} = \mathbf{P} \cdot \mathbf{a}_i$, $\text{proj}_{\mathbf{G},\mathbf{a}_i} = \mathbf{G} \cdot \mathbf{a}_i$
  $\min_{\mathbf{P},\mathbf{a}_i} = \min(\text{proj}_{\mathbf{P},\mathbf{a}_i})$, $\max_{\mathbf{P},\mathbf{a}_i} = \max(\text{proj}_{\mathbf{P},\mathbf{a}_i})$
  $\min_{\mathbf{G},\mathbf{a}_i} = \min(\text{proj}_{\mathbf{G},\mathbf{a}_i})$, $\max_{\mathbf{G},\mathbf{a}_i} = \max(\text{proj}_{\mathbf{G},\mathbf{a}_i})$
  $|P_{A_i} \cap G_{A_i}| = \max(0, \min(\max_{\mathbf{P},\mathbf{a}_i}, \max_{\mathbf{G},\mathbf{a}_i}) - \max(\min_{\mathbf{P},\mathbf{a}_i}, \min_{\mathbf{G},\mathbf{a}_i}))$
  $|P| = \max_{\mathbf{P},\mathbf{a}_i} - \min_{\mathbf{P},\mathbf{a}_i}$
  $|G| = \max_{\mathbf{G},\mathbf{a}_i} - \min_{\mathbf{G},\mathbf{a}_i}$
  $|P_{a_i} \cup G_{a_i}| = |P| + |G| - |P_{a_i} \cap G_{a_i}|$
  $|C| = \max(\max_{\mathbf{P},\mathbf{a}_i}, \max_{\mathbf{G},\mathbf{a}_i}) - \min(\min_{\mathbf{P},\mathbf{a}_i}, \min_{\mathbf{G},\mathbf{a}_i})$
  $\text{IoU} = \frac{|P \cap G|}{|P \cup G|}$
  $\text{GIoU}_i^{1D} = \text{IoU} - \frac{|C| - |P_{a_i} \cup G_{a_i}|}{|C|}$
  $S = S + \text{GIoU}_i^{1D}$
$\text{MGIoU} = \frac{S}{|\mathcal{A}|}$
$\mathcal{L}_{\text{MGIoU}} = \frac{1 - \text{MGIoU}}{2}$
if *shape is unstructured* **then**
  Compute $\mathcal{L}_{\text{convexity}}$:
  for *each* $i = 0$ *to* $N_P - 1$ **do**
    $\mathbf{e}_i = \mathbf{p}_{(i+1) \mod N_P} - \mathbf{p}_i$
    $\mathbf{n}_i = (-e_{i,y}, e_{i,x})$
    $s_1 = \sum_{j=0}^{N_P-1} \max(0, -(\mathbf{p}_j - \mathbf{p}_i) \cdot \mathbf{n}_i)$
    $s_2 = \sum_{j=0}^{N_P-1} \max(0, (\mathbf{p}_j - \mathbf{p}_i) \cdot \mathbf{n}_i)$
    $\text{penalty}_i = \min(s_1, s_2)$
  $\mathcal{L}_{\text{convexity}} = \frac{1}{N_P} \sum_{i=0}^{N_P-1} \text{penalty}_i$
else
  return $\mathcal{L}_{MGIoU}$
$\mathcal{L}_{\text{MGIoU}^+} = \mathcal{L}_{\text{MGIoU}} + \lambda \mathcal{L}_{\text{convexity}}$
return $\mathcal{L}_{MGIoU^+}$

Algorithm 2: MGIoU$^-$ - Minimize Shape Overlap

input : Trajectory boxes $\mathbf{T} = \{\mathbf{T}_i^t \in \mathbb{R}^{4\times 2} \mid t = 1, \ldots, T; i = 1, \ldots, B\}$, masks $\mathbf{M} = \{m_i^t\}$, scores $\mathbf{S} = \{s_i\}$
output: $\mathcal{L}_{\text{MGIoU}^-}$

Initialize $\mathcal{L}_{\text{MGIoU}^-} = 0$
for *each* $t = 1$ *to* T **do**
  Initialize $\text{loss}_i^t = 0$ for $i = 1, \ldots, B$
  for *each pair* (i, j), $i \neq j$ **do**
    $\mathcal{A}_{ij}^t = \{\text{normals of } \mathbf{T}_i^t, \mathbf{T}_j^t\}$
    Initialize $\mathcal{O}_{ij}^t = []$
    for *each* $\mathbf{a}_k \in \mathcal{A}_{ij}^t$ **do**
      $\text{proj}_{\mathbf{T}_i^t,k} = \mathbf{T}_i^t \cdot \mathbf{a}_k$, $\text{proj}_{\mathbf{T}_j^t,k} = \mathbf{T}_j^t \cdot \mathbf{a}_k$
      Compute min and max of $\text{proj}_{\mathbf{T}_i^t,k}$, $\text{proj}_{\mathbf{T}_j^t,k}$
      $|P \cap G| = \max(0, \min(\max_{\mathbf{T}_i^t}, \max_{\mathbf{T}_j^t}) - \max(\min_{\mathbf{T}_i^t}, \min_{\mathbf{T}_j^t}))$
      $|P| = \max_{\mathbf{T}_i^t} - \min_{\mathbf{T}_i^t}$, $|G| = \max_{\mathbf{T}_j^t} - \min_{\mathbf{T}_j^t}$
      $|P \cup G| = |P| + |G| - |P \cap G|$, $|C| = \max(\max_{\mathbf{T}_i^t}, \max_{\mathbf{T}_j^t}) - \min(\min_{\mathbf{T}_i^t}, \min_{\mathbf{T}_j^t})$
      $\text{IoU} = \frac{|P \cap G|}{|P \cup G|}$
      $\text{GIoU}_k^{1D} = \text{IoU} - \frac{|C| - |P \cup G|}{|C|}$
      Append GIoU_k^{1D} to $\mathcal{O}_{ij}^t$
    Take the smallest $\text{GIoU}_{\text{min}}^{1D}$ from $\mathcal{O}_{ij}^t$
    $K_{ij}^t = \text{ReLU}\left(\text{GIoU}_{\text{min}}^{1D}\right)$
    $\text{loss}_i^t = \text{loss}_i^t + K_{ij}^t$
for *each* $i = 1$ *to* B **do**
  $L_i = \sum_{t=1}^{T} m_i^t \cdot \text{loss}_i^t$
  $\mathcal{L}_{\text{MGIoU}^-} = \mathcal{L}_{\text{MGIoU}^-} + s_i \cdot L_i$
return $\mathcal{L}_{MGIoU^-}$

introduce a convexity regularizer ($\mathcal{L}_{\text{convexity}}$) to ensure geometric consistency. The convexity loss, $\mathcal{L}_{\text{convexity}}$, is added for unstructured shapes like polygons to prevent unrealistic concavities, which the MGIoU loss alone might allow. It penalizes vertices that fall on the wrong side of any edge's outward normal, using signed distances to enforce a convex shape, as outlined in algorithm 1. Specifically, for each edge i, we compute the outward normal $\mathbf{n}_i$ and then calculate the signed distances of all other vertices to this edge. If a vertex lies inside the half-plane defined by the edge (i.e., on the wrong side), it contributes to the penalty. The penalty for each edge is the minimum of the sums of positive and negative signed distances, ensuring that the loss is zero only when all vertices are on the correct side of every edge. This convexity loss is then averaged over all edges to obtain $\mathcal{L}_{\text{convexity}}$. The signed distance provides a continuous measure of deviation from convexity, guiding the optimization toward a valid shape, especially critical for polygons with variable vertex counts. Combined with MGIoU loss as $\mathcal{L}_{\text{MGIoU}^+} = \mathcal{L}_{\text{MGIoU}} + \lambda \mathcal{L}_{\text{convexity}}$, it ensures both accurate overlap and geometric consistency.

Instead of maximizing parametric shape overlaps like MGIoU and MGIoU$^+$, we propose **MGIoU$^-$** variant minimize them. We chose trajectory prediction task to showcase our loss capability. Here, the goal shifts from maximizing overlap to minimizing it, ensuring predicted trajectories remain separated. Inspired by Separating Axis Theorem, MGIoU$^-$ adapts the MGIoU framework by computing pairwise overlaps between trajectory boxes at each timestep t. MGIoU$^-$ naturally extends to temporal scenarios by computing overlap penalties between predicted trajectory boxes at each timestep t for all pairs of trajectories i and j. For each timestep t, the algorithm evaluates the smallest 1D GIoU ($\text{GIoU}_{\text{min}}^{1D}$) across the normals of trajectory pairs, penalizing overlaps using a ReLU function. These penalties are summed across all timesteps $t = 1$ to T,

weighted by ground truth masks m_i^t, ensuring the loss accounts for the entire prediction horizon. For each trajectory i, the loss L_i integrates temporal penalties over T timesteps, modulated by masks to handle invalid or variable-length sequences. The final MGIoU$^-$ loss is computed by summing each trajectory's loss L_i, weighted by prediction scores s_i, prioritizing confident predictions across all timesteps. This temporal formulation ensures that overlaps are minimized throughout the sequence, promoting collision-free trajectories over T timesteps. The final $\mathcal{L}_{\text{MGIoU}^-}$ is normalized by the classification head's confidence scores, encouraging the model to either select safer trajectories or adjust their location/orientation regressions to reduce overlaps. Efficiency-wise, MGIoU$^-$ also benefits from parallelizable projections, with a cost of $O(B^2 \cdot T \cdot |\mathcal{A}|)$ per batch, where B is the number of trajectories, T is the number of timesteps, and $|\mathcal{A}|$ is typically 4 in 2D, as outlined in algorithm 2.

In summary, MGIoU and its variants form a general framework for shape optimization tasks, offering a solution for convex shape optimization.

Experimental Settings

This section describes datasets, baselines, and training setups to evaluate MGIoU, MGIoU$^+$, and MGIoU$^-$ across tasks: 2D oriented object detection, monocular 3D 6-DoF object regconition, quadrangle object detection, collision avoidance in trajectory prediction, and 6D Grasp Detection.

2D Oriented Object Detection: We used the DOTAv1.5 (Xia et al. 2018) dataset (single-scaled split). The baseline model was RetinaNet (Lin et al. 2017) (similar baseline to (Yang et al. 2021c,b, 2022)), trained for 12 epochs with SGD, a step LR scheduler (epochs 8 and 11), and a 500-iteration linear warmup.

Monocular 3D 6-DoF Object Recognition: We select Omni3D (Brazil et al. 2023) large-scale dataset, which includes SUNRGBD (Song, Lichtenberg, and Xiao 2015), Hypersim (Roberts et al. 2021), ARKitScenes (Baruch et al. 2021), Objectron (Ahmadyan et al. 2021), KITTI (Geiger et al. 2013), and nuScenes (Caesar et al. 2020) datasets. This ensures robust evaluation across indoor and outdoor environments. The baseline model is CubeRCNN (Brazil et al. 2023) proposed by dataset authors, trained for 5,568,000 iterations on a 4090 GPU. Training followed Omni3D (Brazil et al. 2023), using SGD with a step learning rate scheduler.

Quadrangle Object Detection: The quadrangle object detection task, we choose ICDAR2017 (Gomez et al. 2017) dataset. The ICDAR2017 competition includes the MLT (Multilingual Text Detection) task, covering text in nine languages and various orientations. This dataset suits quadrangle detection due to irregular text region shapes. The baseline model selected was YOLO-NAS (Aharon et al. 2021) due to performance and ease of implementation, we train for 40 epochs with AdamW and a cosine LR scheduler.

Collision Avoidance in Trajectory Prediction: The Waymo (Sun et al. 2020) motion prediction dataset was utilized together with MTR (Shi et al. 2022) baseline. We trained the model on a subset of 20% of the training set and report results on validation set, values in table 4 is for the most confident trajectory per object (not the average of top 6 trajectories). The original architecture was retained to isolate and evaluate MGIoU$^-$ effectiveness in reducing collisions.

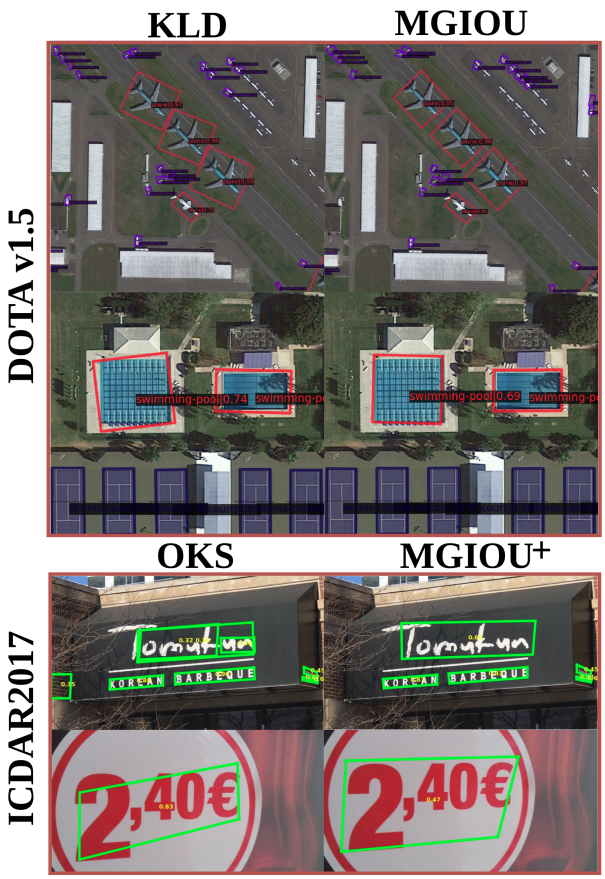

Figure 2: Qualitative visualisation (Test set images) comparing MGIoU vs KLD losses on DOTA dataset, and MGIoU$^+$ vs OKS distance on ICDAR2017 Dataset. MGIoU and MGIoU$^+$ can capture the orientation (2D oriented detection) and vertices better (quadrilaterals detection)

.

Diffusion-based 6D Grasp Detection: We evaluated MGIoU on the Grasp-Anything-6D dataset (Vuong et al. 2024), containing over 3 million objects with 6-DoF grasp poses. We trained the diffusion model on 20% of the data for 20 epochs using AdamW with Step LR scheduler, replacing the standard L2 denoising loss with MGIoU demonstrating improved grasp metrics.

Comparisons with Losses and Discussions

We evaluate MGIoU on five tasks accross domains, demonstrating consistent improvements over existing losses. **2D Oriented Object Detection.** On DOTAv1.5 with RetinaNet(Lin et al. 2017), MGIoU achieves the highest mAP (0.554) with minimal computational overhead (0.45ms latency). table 3 shows MGIoU significantly outperforms KFIoU (51.1x slower), GWD (16.9x slower), and on par with KLD while 17.3x faster.

Monocular 3D 6-DoF Object Recognition. For 3D 6-DoF

Loss	SUNRGBD	Hypersim	ARKitScenes	Objectron	KITTI	nuScenes	$\text{Omni3D}_{\text{Out}}$	$\text{Omni3D}_{\text{In}}$	Omni3D
L1 + SO3	14.19	6.44	41.27	52.87	30.86	26.42	29.22	19.34	22.21
Distangled-L1 + Chamfer	14.64	6.89	40.59	54.94	28.37	27.07	29.93	20.02	22.82
MGIoU	**16.82**	**8.22**	**43.76**	**56.84**	**31.95**	**29.66**	**32.08**	**21.87**	**24.86**
Abs Imp.	2.18↑	1.33↑	3.17↑	1.90↑	3.58↑	2.59↑	2.15↑	1.85↑	2.04↑
Rel Imp.	18.34%	19.30%	7.81%	3.46%	12.62%	9.57%	7.18%	9.24%	8.94%

Table 1: mAP3D Comparison for 3D 6-DoF Object Detection on Omni3D. Baseline model is CubeRCNN (Brazil et al. 2023). Improvement values are relative to the Distangled-L1+Chamfer loss (Brazil et al. 2023). MGIoU consistently outperform the baselines in both indoor and outdoor settings

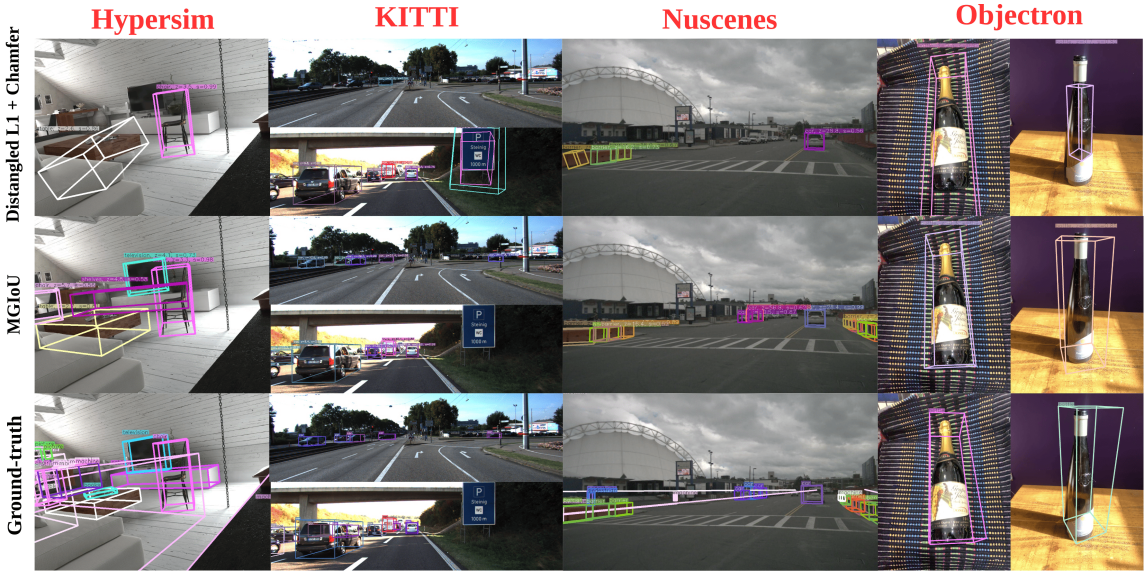

Figure 3: Qualitative visualisation on Omni3D dataset. Best viewed zoomed in.

Method	CR ↑	EMD ↓	CFR ↑
Baseline	0.2505	2.5459	0.865
MGIoU (Ours)	**0.2763**	**2.5394**	**0.875**

Table 2: Comparison of grasp quality metrics on Grasp-Anything-6D dataset.

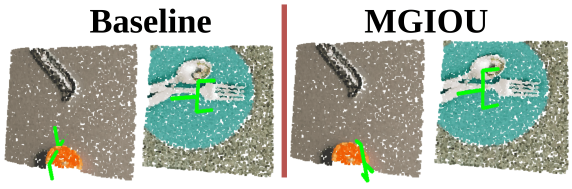

Figure 4: Qualitative results for Diffusion-based 6D Grasp Detection on Grasp-Anything-6D dataset. MGIoU shows better geometric understanding and compliance compared to the baseline L2 loss, generating more accurate and feasible grasp poses that better align with object geometry. Left - Get the juicy orange, Right - Grasp the shiny metal fork.

Loss	mAP	Latency (ms)	Rel. Latency
L1	0.522	**0.03**	×0.10
KFIoU	0.546	23	×51.1
GWD	0.547	7.6	×16.9
KLD	0.55	7.8	×17.3
MGIoU (ours)	**0.554**	0.45	×1.00

Table 3: Comparison of Losses on DOTAv1.5 dataset

object recognition on the Omni3D dataset, we evaluated MGIoU using CubeRCNN (Brazil et al. 2023) as the baseline model. Performance was assessed with AP3D across indoor and outdoor scenes. table 1 demonstrates that MGIoU consistently improved AP3D across all datasets compared to Distangled-L1 + Chamfer, with gains from 1.15% (Objectron) to 18.34% (SUNRGBD). The overall Omni3D improvement was 2.04 (9% relative improvement), showing MGIoU's robustness in both indoor ($\text{Omni3D}_{\text{In}}$) and outdoor ($\text{Omni3D}_{\text{Out}}$) settings. Other losses typically involve multiple components—such as separate terms for location, dimension, and rotation—requiring hyperparameter tuning. For example, in the baseline Distangled-L1 + Chamfer loss,

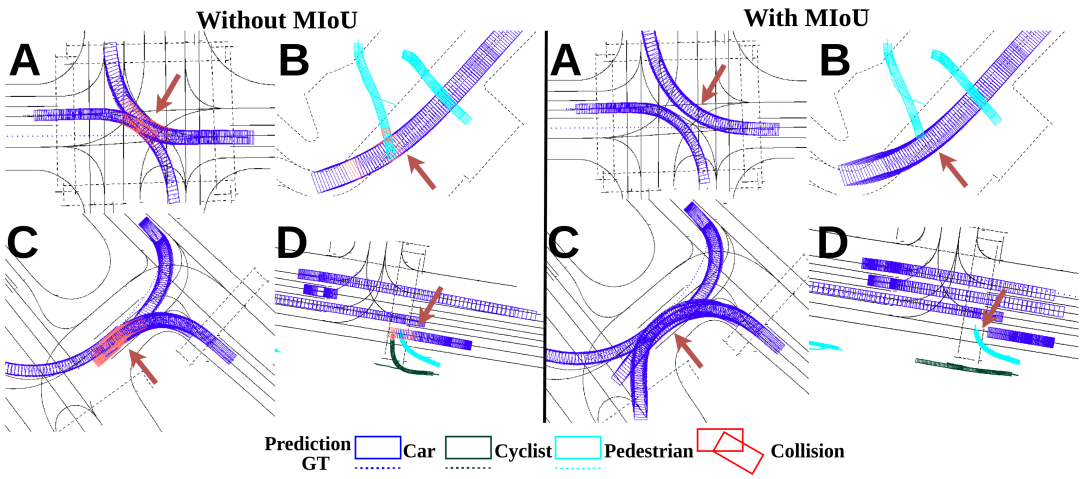

Figure 5: Qualitative visualisation on Waymo dataset, we visualize the predicted future bounding boxes of road agents in the next 8 seconds (80 timesteps), with and without MGIoU$^-$ during training. With MGIoU$^-$ incorporated in the training stage, model now have a better understanding of the physical world and can make safer interactions between road agents. **A)** 2 cars avoid collision at an intersection, **B)** Pedestrian stops and wait, **C)** Smooth interaction among three vehicles without collisions, **D)** Car appropriately yields to a pedestrian

.

Metric	Without MGIoU$^-$	With MGIoU$^-$
mAP↑	0.2823	**0.2961**
minADE↓	1.3668	**1.3449**
minFDE↓	3.3398	**3.3046**
MissRate↓	0.4466	**0.4407**
# Collisions↓	7493	**6443 (↓ 14%)**

Table 4: Trajectory Prediction task on Waymo Dataset (out of 192,172 predicted trajectories). Detailed in Appendix.

Loss	AP	AR
QRN	48.60	42.67
OKS distance	49.10	43.39
MGIoU$^+$ (Ours)	**49.84**	**44.91**

Table 5: Comparison of Losses on ICDAR2017 dataset.

the L1 loss handles location and dimension, while Chamfer distance addresses rotation discrepancies. This multi-component setup necessitates careful tuning to optimize the trade-off between these aspects, as illustrated in fig. 3. In contrast, MGIoU unifies location, dimension, and rotation into a single loss function, optimizing them holistically without additional hyperparameters. There are cases where the baseline predicts a nearly perfect location, but slight errors in dimension or rotation lead to significant drops in 3D IoU, and vice versa. As shown in fig. 3, the baseline's predicted boxes often show misalignments, whereas MGIoU's boxes better match the ground-truth boxes, mitigating these issues.

Quadrangle Object Detection. On ICDAR2017 with YOLO-NAS (Aharon et al. 2021), MGIoU$^+$ achieves superior AP (49.84) and AR (44.91) compared to QRN and OKS distance losses (table 5), demonstrating effectiveness for irregular quadrilateral shapes.

Collision Avoidance in Trajectory Prediction. On Waymo with MTR baseline, MGIoU$^-$ improves mAP (0.2823→0.2961) and reduces collisions by 14% (7,493→6,443) (table 4). As shown in fig. 5, MGIoU$^-$ enables safer agent interactions by penalizing trajectory overlaps, encouraging collision-free path generation while maintaining prediction accuracy.

Diffusion-based 6D Grasp Detection. table 2 shows MGIoU outperforms L2 loss across all metrics: CR improved 10.3% (0.2505→0.2763), EMD decreased (2.5459→2.5394), and CFR increased 1.2% (0.865→0.875). These improvements validate that MGIoU's shape-aware optimization better captures geometric constraints for successful grasping (fig. 4).

Conclusion

We introduce Marginalized Generalized IoU (MGIoU), a novel loss function that unifies parametric shape optimization across computer vision applications. Experiments across five tasks demonstrate that MGIoU and its variants consistently outperform existing objective functions without additional inference overhead, providing a robust solution for convex shape optimization.

Acknowledgments

This work was partially supported by the DARPA Assured Neuro-Symbolic Learning and Reasoning (ANSR) program (FA8750-23-2-1016), the ONR Global X-Challenge Grant (N62909-25-1-2067), and the Australian Research Council Discovery Project ARC DP2020102427.

References

Aharon, S.; Louis-Dupont; Ofri Masad; Yurkova, K.; Lotem Fridman; Lkdci; Khvedchenya, E.; Rubin, R.; Bagrov, N.; Tymchenko, B.; Keren, T.; Zhilko, A.; and Eran-Deci. 2021. Super-Gradients.

Ahmadyan, A.; Zhang, L.; Ablavatski, A.; Wei, J.; and Grundmann, M. 2021. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7822–7831.

Baruch, G.; Chen, Z.; Dehghan, A.; Dimry, T.; Feigin, Y.; Fu, P.; Gebauer, T.; Joffe, B.; Kurz, D.; Schwartz, A.; et al. 2021. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*.

Bi, Y.; and Hu, Z. 2021. Disentangled contour learning for quadrilateral text detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 909–918.

Brazil, G.; Kumar, A.; Straub, J.; Ravi, N.; Johnson, J.; and Gkioxari, G. 2023. Omni3D: A Large Benchmark and Model for 3D Object Detection in the Wild. In *CVPR*. Vancouver, Canada: IEEE.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828–5839.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11): 1231–1237.

Gomez, R.; Shi, B.; Gomez, L.; Numann, L.; Veit, A.; Matas, J.; Belongie, S.; and Karatzas, D. 2017. Icdar2017 robust reading challenge on coco-text. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, 1435–1443. IEEE.

He, Z.; Zhou, Y.; Wang, Y.; Wang, S.; Lu, X.; Tang, Z.; and Cai, L. 2018. An end-to-end quadrilateral regression network for comic panel extraction. In *Proceedings of the 26th ACM international conference on Multimedia*, 887–895.

Keserwani, P.; Singh, R.; and Shukla, R. 2021. Quadbox: A new approach for arbitrary quadrilateral detection. *Neurocomputing*, 448: 188–200.

Kim, S.; Jeon, H.; Choi, J. W.; and Kum, D. 2022. Diverse multiple trajectory prediction using a two-stage prediction network trained with lane loss. *IEEE Robotics and Automation Letters*, 8(4): 2038–2045.

Li, Z.; Xu, X.; Lim, S.; and Zhao, H. 2024. UniMODE: Unified monocular 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16561–16570.

Liao, M.; Shi, B.; and Bai, X. 2018. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, 27(8): 3676–3690.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, Y.; and Jin, L. 2017. Deep matching prior network: Toward tighter multi-oriented text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1962–1969.

Maji, D.; Nagori, S.; Mathew, M.; and Poddar, D. 2022. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2637–2646.

Nguyen, N.; Vu, M. N.; Huang, B.; Vuong, A.; Le, N.; Vo, T.; and Nguyen, A. 2024a. Lightweight Language-driven Grasp Detection using Conditional Consistency Model. In *IROS*.

Nguyen, T.; Vu, M. N.; Huang, B.; Vuong, A.; Vuong, Q.; Le, N.; Vo, T.; and Nguyen, A. 2024b. Language-driven 6-dof grasp detection using negative prompt guidance. In *ECCV*.

Niedoba, M.; Cui, H.; Luo, K.; Hegde, D.; Chou, F.-C.; and Djuric, N. 2019. Improving movement prediction of traffic actors using off-road loss and bias mitigation. In *Workshop on'Machine Learning for Autonomous Driving'at Conference on Neural Information Processing Systems*, volume 1, 2.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *arXiv preprint arXiv:1612.00593*.

Qian, W.; Yang, X.; Peng, S.; Yan, J.; and Guo, Y. 2021. Learning modulated loss for rotated object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 2458–2466.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Redmon, J.; and Farhadi, A. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263–7271.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Roberts, M.; Ramapuram, J.; Ranjan, A.; Kumar, A.; Bautista, M. A.; Paczan, N.; Webb, R.; and Susskind, J. M. 2021. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10912–10922.

Shi, S.; Jiang, L.; Dai, D.; and Schiele, B. 2022. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*.

Shi, S.; Jiang, L.; Dai, D.; and Schiele, B. 2023. MTR++: Multi-Agent Motion Prediction with Symmetric Scene Modeling and Guided Intention Querying. *arXiv preprint arXiv:2306.17770*.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 567–576.

Su, H.; Jampani, V.; Sun, D.; Maji, S.; Koltun, V.; and Torr, P. 2020. Deep Learning on the Rotation Manifold for Object Detection in 3D Point Clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2446–2454.

Suo, S.; Regalado, S.; Casas, S.; and Urtasun, R. 2021. Trafficsim: Learning to simulate realistic multi-agent behaviors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10400–10409.

Vuong, A. D.; Vu, M. N.; Huang, B.; Nguyen, N.; Le, H.; Vo, T.; and Nguyen, A. 2024. Language-driven Grasp Detection. In *CVPR*, 17902–17912.

Xia, G.-S.; Bai, X.; Ding, J.; Shen, Z.; Ma, Z.; Wang, H.; Gao, J.; Zhang, D.; Li, M.; He, B.; et al. 2018. Dota: A large-scale dataset for object detection in aerial images. *arXiv preprint arXiv:1803.06165*.

Yang, X.; Liu, Q.; Yan, J.; and Li, A. 2021a. R3Det: Refined Joint Representation for Oriented Object Detection. *arXiv preprint arXiv:2105.02445*.

Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; and Tian, Q. 2021b. Rethinking rotated object detection with gaussian wasserstein distance loss. In *International conference on machine learning*, 11830–11841. PMLR.

Yang, X.; Yang, X.; Yang, J.; Ming, Q.; Wang, W.; Tian, Q.; and Yan, J. 2021c. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *Advances in Neural Information Processing Systems*, 34: 18381–18394.

Yang, X.; Zhou, Y.; Zhang, G.; Yang, J.; Wang, W.; Yan, J.; Zhang, X.; and Tian, Q. 2022. The KFIoU loss for rotated object detection. *arXiv preprint arXiv:2201.12558*.

Zhang, Q.; Gao, Y.; Zhang, Y.; Guo, Y.; Ding, D.; Wang, Y.; Sun, P.; and Zhao, D. 2022. Trajgen: Generating realistic and diverse trajectories with reactive and feasible agent behaviors for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(12): 24474–24487.

Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; and Ren, D. 2020. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12993–13000.

Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; and Zuo, W. 2021. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. *IEEE Transactions on cybernetics*, 52(8): 8574–8586.