Dexterous Grasp Transformer

Guo-Hao Xu*[1], Yi-Lin Wei*[1], Dian Zheng [1], Xiao-Ming Wu [1], Wei-Shi Zheng †[1,2]
[1] School of Computer Science and Engineering, Sun Yat-sen University, China
[2] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
{xugh23, weiylin5, zhengd35, wuxm65}@mail2.sysu.edu.cn wszheng@ieee.org

Abstract

In this work, we propose a novel discriminative framework for dexterous grasp generation, named ***D****exterous* ***G****rasp* ***TR****ansformer (****DGTR****), capable of predicting a diverse set of feasible grasp poses by processing the object point cloud with only* ***one forward pass****. We formulate dexterous grasp generation as a set prediction task and design a transformer-based grasping model for it. However, we identify that this set prediction paradigm encounters several optimization challenges in the field of dexterous grasping and results in restricted performance. To address these issues, we propose progressive strategies for both the training and testing phases. First, the dynamic-static matching training (DSMT) strategy is presented to enhance the optimization stability during the training phase. Second, we introduce the adversarial-balanced test-time adaptation (AB-TTA) with a pair of adversarial losses to improve grasping quality during the testing phase. Experimental results on the DexGraspNet dataset demonstrate the capability of DGTR to predict dexterous grasp poses with both high quality and diversity. Notably, while keeping high quality, the diversity of grasp poses predicted by DGTR significantly outperforms previous works in multiple metrics without any data pre-processing. Codes are available at https://github.com/iSEE-Laboratory/DGTR.*

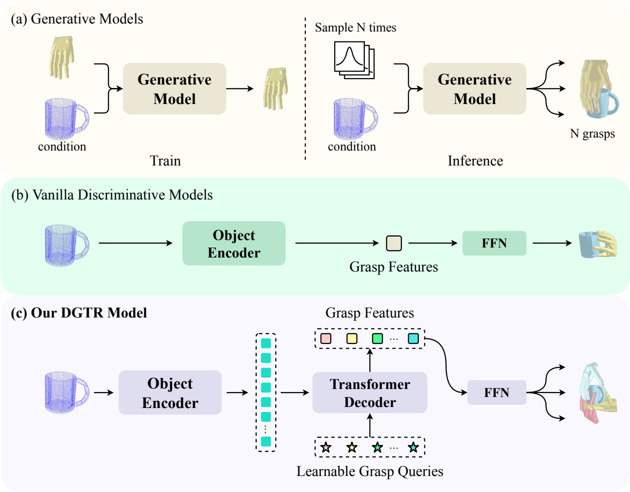

Figure 1. Comparison of DGTR and other dexterous grasping frameworks. The generative models (a) usually learn the distribution of the grasp poses conditioned on the object point cloud. At test time, they mainly infer multiple times to generate several grasps but produce nearly identical grasp poses with the same condition. The vanilla discriminative models (b) mainly learn to predict one grasp pose for the input point cloud. Our DGTR model (c) adopts a transformer decoder and learnable queries, and learns to predict a set of diverse grasps poses with one forward pass.

1. Introduction

Robotic dexterous grasping stands as a fundamental and critical task in the field of robotics and computer vision, offering a versatile and fine-grained approach with extensive applications in industrial production and daily scenarios.

With the development of deep learning and large-scale datasets for dexterous grasp generation, learning-based methods achieve considerable performance in grasping quality and generalizability [3, 11, 12]. Concurrently, acquiring grasping diversity (especially grasping from various directions) is also a crucial task [18, 41] as it provides the robot with robustness and task flexibility during the manipulation task. Previous learning-based approaches mostly utilize generative models to model the grasp distribution conditioned on the object point cloud as shown in Figure 1 (a). However, conditional generative models may consistently generate nearly identical outputs (given the same input) at inference time due to the powerful condition [31, 42], except for a diffusion-based model [11], which can generate diverse grasps but with low quality. Alternatively, vanilla discriminative models shown in Figure 1 (b) can only predict a single grasp pose for one input object [19]. Therefore, to obtain diversity, both of them have to rotate the

*Equal contribution.
†Corresponding author.

input point cloud and infer multiple times, which is time-consuming and quality-limiting.

In this work, we propose Dexterous Grasp Transformer (DGTR), a novel discriminative framework to tackle the task of predicting diverse and high-quality dexterous grasp poses given the complete object point cloud. We formulate dexterous grasp generation as a set prediction task and design a transformer-based grasping model inspired by the impressive success of Detection Transformers [1, 24]. As illustrated by Figure 1 (c), DGTR adopts a transformer decoder and utilizes learnable grasp queries representing different grasping patterns to predict a diverse set of feasible grasp poses by processing the object point cloud only once.

However, we observe that DGTR faces an optimization challenge in our task, which results in the dilemma between model collapse and unacceptable object penetration of the predicted grasps. As depicted in Figure 2 (a), applying a large weight on the object penetration loss causes the model to learn a trivial solution where all predictions are nearly identical. On the contrary, a zero weight for the penetration loss leads to severe object penetration of the grasps, as shown in Figure 2 (b). We identify the main cause of this challenge to be the instability of the Hungarian algorithm, which is exacerbated by the powerful object penetration loss. As the weight of the object penetration loss increases, the matching process becomes more unstable. Consequently, the unstable matching results misguide the optimization process of the model, ultimately causing the model collapse. We conduct abundant analysis and experiments for this in Section 3.3 and 4.5.1.

To overcome this challenge, we propose progressive strategies for both the training and testing phases, which simultaneously enhance the diversity and quality of grasp poses as demonstrated in Figure 2 (c). Firstly, we present a dynamic-static matching training (DSMT) strategy, which is built on the insight of guiding the model to learn appropriate targets through dynamic matching training and subsequently optimize object penetration through static matching training. This strategy ensures effective optimization of the object penetration loss while directing the model optimization reasonably. Secondly, we present an adversarial-balanced test-time adaptation (AB-TTA) strategy to refine the predicted grasp poses directly in the parameter space of the dexterous hand. Specifically, we utilize a pair of adversarial losses: one repels the hand from the interior of the object, while the other attracts it towards the object's surface. The strategic interaction of the adversarial losses substantially enhances the quality of the grasp and mitigates the penetration. Notably, our AB-TTA neither relies on any 3D mesh information of the objects nor involves complex force analysis or auxiliary models.

Extensive experiments on DexGraspNet dataset show that our methods are capable of generating high-quality and high-diversity grasp poses on thousands of objects. To the best of our knowledge, this is the first work to predict a diverse set of dexterous grasp poses by processing the input object just once, without any need for data preprocessing.

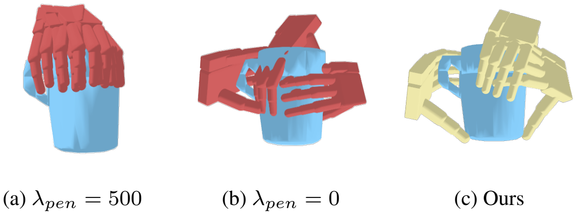
(a) $\lambda_{pen} = 500$ (b) $\lambda_{pen} = 0$ (c) Ours

Figure 2. Comparison of grasp quality and diversity under different penetration loss weights. We visualize 3 grasps for each circumstance. (a) large object penetration weight; (b) zero object penetration weight; (c) our progressive strategies.

2. Related Works

2.1. Dexterous Grasp Generation

Dexterous grasping is a promising task as it endows robots with the capability to manipulate objects like humans. Meanwhile, it also presents significant challenges due to the high degree-of-freedom design of dexterous hands. Early methods focus on analytical methods [8, 20, 28] and optimize the hand poses with kinematics and physical mechanisms to a force-closure state. Several works [18, 39] synthesize datasets for dexterous grasps with [20], but face challenges in the generating speed and success rate.

Recently data-driven methods [17, 22, 33, 35, 37, 45] have received increasing research attention with the development of deep neural networks. GraspTTA [12] utilizes a CVAE [34] to synthesize grasps with their hand-object consistency constraints. UnidexGrasp [41] proposes two variants of IPDF [25] and Glow [14] to predict object orientation, translation and articulation for the dexterous hand respectively. Some works [3, 11, 36] explore conditioned normalizing flow [14, 26], generative adversarial network [10] and conditioned diffusion models [30] to learn the probabilistic distribution of the dexterous grasps. In contrast, DDG [19] exploits a non-generative model and a differentiable Q_1 loss to learn one grasp pose for each instance.

However, these methods struggle to generate feasible and diverse grasps given the same input point cloud, either because the condition (*e.g.*, object point cloud) significantly restricts the generation direction of the model, or because of the limitation of the model architecture. To alleviate this problem, our work learns to predict a diverse set of grasps of an object at one time with a transformer-based framework specially designed for dexterous grasp generation.

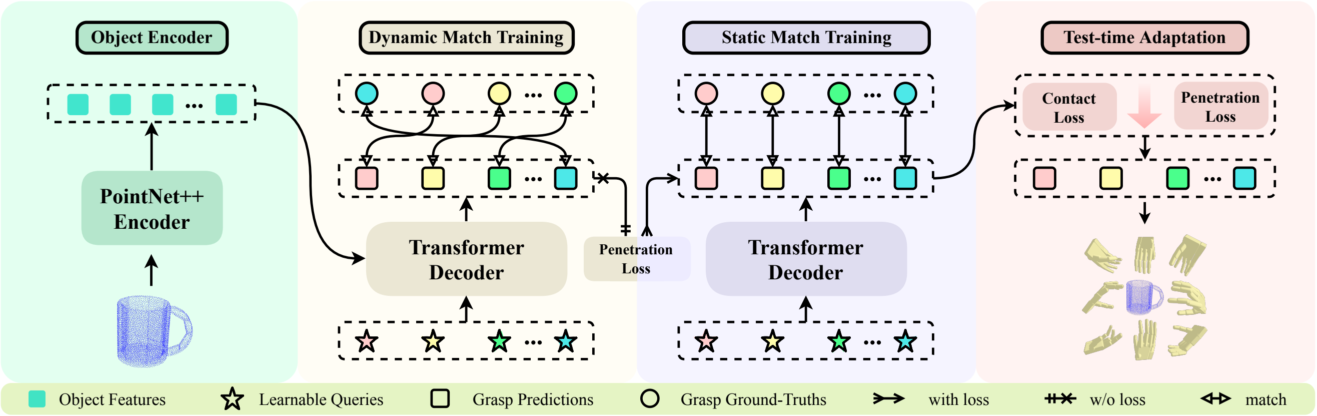

Figure 3. **Overview of our DGTR framework**. The input of DGTR is the complete point cloud $\mathcal{O}$ of an object. First, the PointNet++ [29] encoder downsamples the point cloud and extracts a set of object features. Next, the transformer decoder takes N learnable query embeddings as well as the object features as input and predicts N diverse grasp poses in parallel. In the dynamic matching training stage, our model is trained with the matching result produced by Hungarian Algorithm [15] and without object penetration loss. In the static matching training stage, we use static matching recorded in the DMT stage to train the model with object penetration loss. At test time, we adopt an adversarial-balanced loss to directly finetune the hand pose parameters.

2.2. Vision Transformer

Vision transformers [2, 4, 13, 21, 44, 46] have received an extensive amount of research attention in recent years, and several of them [1, 5, 40, 43] introduce novel paradigms for computer vision tasks. In our work, dexterous grasp generation from a complete point cloud is considered a set prediction task, which is one of the strengths of detection transformers [1, 24]. However, conventional detection transformers, which are specially designed for object detection, are unsuited for dexterous grasp generation, because of the absence of supervision for feasible grasps, as well as the optimization challenge arising from the grasp losses. To tackle this problem, we equip our model with a series of grasp losses for learning diverse and high-quality grasps, and progressive strategies for stable training and penetration optimization.

3. Dexterous Grasp Transformer

3.1. Problem Formulation

In this work, we focus on generating high-quality and diverse grasp poses from the complete object point cloud. Specifically, given an object point cloud $\mathcal{O} \in \mathbb{R}^{M\times3}$ of size M, our model learns to generate a set of N dexterous grasp poses $\{\mathbf{g}_i\}_{i=1}^{N} = \{(\mathbf{r}_i, \mathbf{t}_i, \mathbf{q}_i)\}_{i=1}^{N}$, where $\mathbf{r}_i \in \mathbf{SO(3)}$ and $\mathbf{t}_i \in \mathbb{R}^3$ are the global rotation and translation in the world coordinate, and $\mathbf{q}_i \in \mathbb{R}^J$ is the joint angles of the J-DoF dexterous hand ($J = 22$ for ShadowHand [32]).

3.2. DGTR Architecture

The model architecture of Dexterous Grasp Transformer (DGTR) contributes most to the diversity and efficiency (*i.e.* N various grasp poses in one forward pass) of our framework. As shown in Figure 3, it mainly consists of three components: 1) a point cloud encoder to extract the object feature, 2) a transformer decoder, and 3) feed-forward networks to predict the grasp poses.

Encoder. We adopt a three-layer PointNet++ [29] as the encoder to extract a set of object features. Given an object point cloud $\mathcal{O} \in \mathbb{R}^{M\times3}$, our encoder outputs the down-sampled point cloud $\mathcal{O}' \in \mathbb{R}^{M'\times3}$ and the corresponding features $\mathcal{F}' \in \mathbb{R}^{M'\times C'}$.

Decoder. Inspired by previous set-prediction frameworks [1, 24], we cascade Transformer blocks [38] as our decoder to predict an unordered set of grasp poses in parallel. This decoder takes as input the point features $\mathcal{F}'$ and a set of learnable grasping queries $\{q_i\}_{i=1}^{N}$ to produce grasp features $\{\mathcal{G}_i\}_{i=1}^{N}$. Since there is no explicit position information among the point features, we encode the raw points $\mathcal{O}' \in \mathbb{R}^{M'\times3}$ with an MLP module as the position embedding of encoder features $\mathcal{F}'$.

Prediction Heads. The grasp pose set $\{\mathbf{g}_i\}_{i=1}^{N} = \{(R_i, \mathbf{t}_i, \mathbf{q}_i)\}_{i=1}^{N}$ are predicted with the final decoder features $\{\mathcal{G}_i\}_{i=1}^{N}$ by three independent MLPs. Both the translation and the joint angle predictions are passed through a sigmoid activation to form a normalized value w.r.t. the limits of each dimension. And the rotation prediction is normalized to a unit quaternion with the $L2$ normalization.

The unordered predictions are usually matched with their nearest ground truths using the Hungarian Algorithm [15] before the loss calculation. However, while the Hungarian Algorithm provides an effective solution to train the model regardless of the permutation of the predictions, it also brings ambiguity to the optimizing process of the model, which is a major factor of the dilemma of model collapse

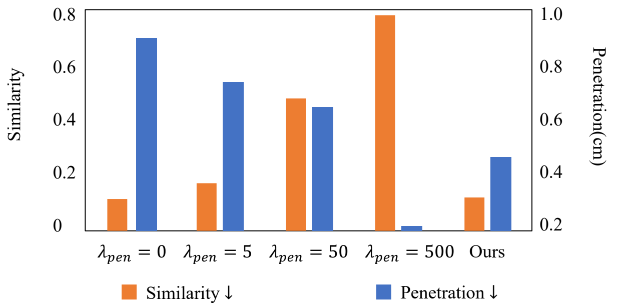

Figure 4. **Comparative analysis of grasp poses similarity and object penetration with various penetration loss weights.** *Similarity* is measured by the cosine similarity of N predicted grasp poses, which represents the *non-diversity*. *Penetration* is the object penetration from the object point cloud to the hand mesh. *Ours* denotes the model trained with our proposed DSMT strategy.

and unacceptable object penetration. We alleviate this problem with a dynamic-static matching training strategy (Section 3.3) and propose an adversarial-balanced loss to further enhance the practicality of the generated grasps at test time (Section 3.4).

3.3. Dynamic-Static Matching Training Strategy

Model Collapse *vs*. Object Penetration. We discover the optimization challenge when DGTR attempts to learn multiple grasping targets of one object simultaneously. As illustrated in Figure 4, DGTR encounters a dilemma between model collapse and the issue of unacceptable object penetration. On one hand, if we impose a heavy penalty on object penetration (*e.g.* $\lambda_{pen} = 500$), the model tends to be stuck in a trivial solution where it predicts nearly identical grasps for the object. On the other hand, if we reduce this penalty (*e.g.* $\lambda_{pen} = 5$) or even remove it ($\lambda_{pen} = 0$), the predicted grasps suffer from severe object penetration.

We analyze the reasons why the object penetration penalty could cause model collapse in the case of set prediction. Intuitively, there is a non-trivial gap between the optimizing difficulties of object penetration and hand pose reconstruction. The object penetration loss could be reduced easily by "pulling" the hand away from the object. While the latter involves a high-dimensional and non-convex optimization problem, which is inherently difficult to solve. Empirically, the object penetration loss increases the instability of Hungarian Algorithm matching results, which profoundly disturbs the optimizing process. As depicted in Figure 5, the instability of Hungarian matching increases as λ_{pen} becomes larger, which results in ambiguous optimization goals for each query [7, 16] and eventually causes the model to learn similar grasp poses for all queries.

DSMT. We serialize the optimizing process and propose a Dynamic-Static Matching Training (DSMT) strategy, aiming to alleviate the optimization challenge arising from the instability of the Hungarian Algorithm and the strong impact of object penetration loss. The key insight is to guide the model learning towards appropriate targets through dynamic training, and subsequently optimizing object penetration through static training.

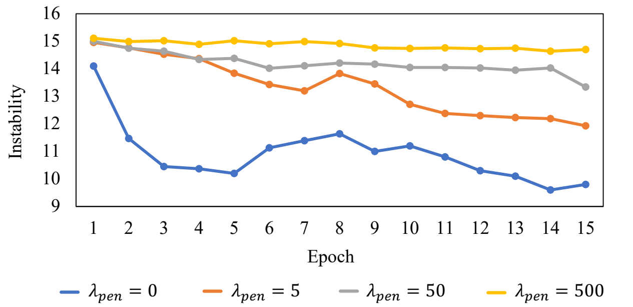

Figure 5. **Hungarian matching instability during training of different penetration loss weights.** The instability is measured by the *IS* metric introduced in [16], where a higher value indicates greater instability.

As illustrated in Algorithm 1, DGTR optimization begins with regular training with the hand regression loss and no object penetration loss for T_0 epochs (DMT). The matching results between the predictions and the targets are dynamically generated by the Hungarian Algorithm. The learnable queries are adequately trained to learn diverse grasping patterns in this stage.

In the Static Matching Warm-up (SMW) stage, we remove the Hungarian Matching process and utilize fixed and stable matching results recorded in the DMT stage. The objective of this stage is to finetune the model and make it adapt to the given static matching. Thus, we still exclude the object penetration loss in this stage.

In the Static Matching Penetration Training (SMPT) stage, the object penetration loss and the hand-object distance loss (Eq. (1)) are incorporated into the training process. The matching results used in the previous stage are preserved to maintain a stable optimization environment. In this way, the severe penetration issue arising from the lack of object penetration penalty in the previous training stages is significantly alleviated.

3.4. Adversarial-Balanced Test-Time Adaptation

Object Contact *vs*. Object Penetration. To further improve the practicality of the predicted grasps, we propose an adversarial-balanced test-time adaptation (AB-TTA) strategy to refine the predicted grasps during the test phase. It is worth noting that our AB-TTA eliminates the need for complex force analysis or auxiliary models. Specifically, this strategy mainly minimizes a pair of adversarial losses, the object penetration loss $\mathcal{L}_{pen}$ and hand-object distance loss $\mathcal{L}_{dist}$ in the parameter space of the dexterous hand. However, the comprehensive optimization of these two losses is challenging. The penetration loss can be easily reduced (*i.e.*, pulling the hand away from the object) in the parame-

Algorithm 1 Dynamic-Static Matching Training

```
Input: Object point clouds $\mathcal{O}$, target grasp poses $\hat{\mathbf{g}}$, training epochs $T_0$, $T_1$, $T_2$, and model parameters $\Theta$
  for $t = 1$ to $T_0$ do                                        ▷ DMT
      $\mathbf{g}_t = \Theta(\mathcal{O})$
      $\hat{\rho}_t = HungarianAlgorithm(\mathbf{g}_t, \hat{\mathbf{g}})$
      $\mathcal{L}(\mathbf{g}_t, \hat{\mathbf{g}}, \hat{\rho}_t) = \mathcal{L}_{regress}(\mathbf{g}_t, \hat{\mathbf{g}}, \hat{\rho}_t)$
      Update $\Theta$ with $\nabla_\Theta \mathcal{L}(\mathbf{g}_t, \hat{\mathbf{g}}, \hat{\rho}_t)$
  end for
  $\mathbf{g}_{T_0} = \Theta(\mathcal{O})$
  $\hat{\rho}_{T_0} = HungarianAlgorithm(\mathbf{g}_{T_0}, \hat{\mathbf{g}})$
  for $t = 1$ to $T_1$ do                                        ▷ SMW
      $\mathbf{g}_t = \Theta(\mathcal{O})$
      $\mathcal{L}(\mathbf{g}_t, \hat{\mathbf{g}}, \hat{\rho}_{T_0}) = \mathcal{L}_{regress}(\mathbf{g}_t, \hat{\mathbf{g}}, \hat{\rho}_{T_0})$
      Update $\Theta$ with $\nabla_\Theta \mathcal{L}(\mathbf{g}_t, \hat{\mathbf{g}}, \hat{\rho}_{T_0})$
  end for
  for $t = 1$ to $T_2$ do                                        ▷ SMPT
      $\mathbf{g}_t = \Theta(\mathcal{O})$
      $\mathcal{L}(\mathbf{g}_t, \hat{\mathbf{g}}, \hat{\rho}_{T_0}) = \mathcal{L}_{regress}(\mathbf{g}_t, \hat{\mathbf{g}}, \hat{\rho}_{T_0})$
                  $+ \mathcal{L}_{pen}(\mathbf{g}_t, \mathcal{O}) + \mathcal{L}_{van-dist}(\mathbf{g}_t, \mathcal{O})$
      Update $\Theta$ with $\nabla_\Theta \mathcal{L}(\mathbf{g}_t, \hat{\mathbf{g}}, \hat{\rho}_{T_0})$
  end for
Output: Optimized model parameters $\Theta$
```

ter space without appropriate constraints, causing the hand-object distance loss to lose efficacy. Hence, we incorporate two key designs to facilitate a balanced decrease of these adversarial losses, which brings considerable improvement in both hand-object contact and hand-object penetration.

AB-TTA. Our AB-TTA is based on the perception that the generated grasp poses are already or nearly valid, only requiring slight adjustments. Firstly, we propose to moderate the displacement of the global translation of the root link of the dexterous hand during the optimization process by downscaling its gradient with β_t. Moderating the global translation constrains the over-optimization of object penetration loss, which promotes the effectiveness and stability of the adaptation.

Secondly, we present a generalized tta-distance loss to address the ineffectiveness of vanilla distance loss used in [39]. The vanilla distance loss is defined as:

$$\mathcal{L}_{van-dist} = \sum_i \mathbb{I}(d(p_i) < \tau) * d(p_i), \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function, τ is a contact threshold to filter out the outliers, and $d(p_i)$ is the distance between the nearest point on the object point cloud and the i^{th} keypoint p_i on the predicted hand. We observe that the vanilla distance loss will be 0 if the hand is too far away from the object, where no point meets the conditions $(d(p_i) < \tau)$. As a result, the hand is unlikely to be "pushed" towards the object again since the distance loss has been 0. We improve the hand-object distance loss by defining a more general condition which constrains the hand keypoints that initially touched the object to remain in contact during optimization. The generalized tta-distance loss is defined as:

$$\mathcal{L}_{tta-dist} = \sum_i \mathbb{I}((d(p_i^c) < \tau) \vee (d(p_i^r) < \tau)) * d(p_i^r), \tag{2}$$

where p_i^c and p_i^r are the i^{th} keypoints of the initial coarse hand and the refined hand at the current iteration, respectively. As a result, a input hand which is nearly valid would not be pulled too far away from the object.

In addition, due to the high DoF of dexterous hands, we also add self-penetration loss $\mathcal{L}_{spen}$ in AB-TTA. Thus, the overall loss function for AB-TTA is

$$\mathcal{L}_{ab-tta} = \alpha_1 * \mathcal{L}_{pen} + \alpha_2 * \mathcal{L}_{tta-dist} + \alpha_3 * \mathcal{L}_{spen}. \tag{3}$$

The details of all losses are in Section 3.5 and *Appendix* A.

3.5. Grasp Losses

The optimization of DGTR involves the grasp losses and the bipartite matching between the predictions and the ground truths. We denote the i^{th} predicted item as $\mathbf{x}_i$ and the j^{th} ground-truth item as $\hat{\mathbf{x}}_j$ ($\mathbf{x} \in \{\mathbf{g}, \mathbf{t}, \mathbf{r}, \mathbf{q}\}$) in the following paragraphs of this section.

Hand Parameters Regression Loss. We utilize the smooth $L1$ loss [9] as $\mathcal{L}_{trans}$ and $\mathcal{L}_{joints}$ to regress the translations and joint angles. For the rotation, we maximize the similarity of the predicted and ground-truth quaternions with $\mathcal{L}_{rotation}(\mathbf{r}_i, \hat{\mathbf{r}}_j) = 1.0 - |\mathbf{r}_i \cdot \hat{\mathbf{r}}_j|$, where $(\cdot)$ is the inner product operation. The overall regression loss for hand parameters is a weighted sum of the above losses:

$$\begin{aligned}\mathcal{L}_{param}(\mathbf{g}_i, \hat{\mathbf{g}}_j) &= \lambda_1 * \mathcal{L}_{trans}(\mathbf{t}_i, \hat{\mathbf{t}}_j) \\ &+ \lambda_2 * \mathcal{L}_{joints}(\mathbf{q}_i, \hat{\mathbf{q}}_j) + \lambda_3 * \mathcal{L}_{rotation}(\mathbf{r}_i, \hat{\mathbf{r}}_j).\end{aligned} \tag{4}$$

Hand Chamfer Loss. We incorporate a hand chamfer loss $\mathcal{L}_{chamfer}(\mathbf{g}_i, \hat{\mathbf{g}}_j)$ to explicitly minimize the discrepancies between the actual shapes of the predicted and ground-truth hands. Specifically, we apply $\mathbf{g}_i$ and $\hat{\mathbf{g}}_j$ to the dexterous hand and obtain the hand meshes $\mathcal{H}(\mathbf{g}_i)$ and $\mathcal{H}(\hat{\mathbf{g}}_j)$ by forward kinematics. Then we sample the hand point clouds $\Phi(\mathbf{g}_i)$ and $\Phi(\hat{\mathbf{g}}_j)$ from the corresponding meshes and calculate the Chamfer distance [6] between them.

Penetration Loss. We employ two penetration losses: 1) $\mathcal{L}_{pen}(\mathbf{g}_i, \mathcal{O})$ [41]: object penetration calculated by the signed squared distance function from object point cloud to the hand mesh, and 2) $\mathcal{L}_{spen}(\mathbf{g}_i)$ [39]: self penetration depth from the keypoints of the hand to themselves.

Cost Function for Bipartite Matching. To obtain a bipartite matching between predictions and ground truths, the cost function for each pair of $(\mathbf{g}_i, \hat{\mathbf{g}}_j)$ is defined as:

$$\begin{aligned}\mathcal{C}(\mathbf{g}_i, \hat{\mathbf{g}}_j) &= \omega_1 * \mathcal{L}_{trans}(\mathbf{t}_i, \hat{\mathbf{t}}_j) + \omega_2 * \mathcal{L}_{joints}(\mathbf{r}_i, \hat{\mathbf{r}}_j) \\ &+ \omega_3 * \mathcal{L}_{rotation}(\mathbf{q}_i, \hat{\mathbf{q}}_j).\end{aligned} \tag{5}$$

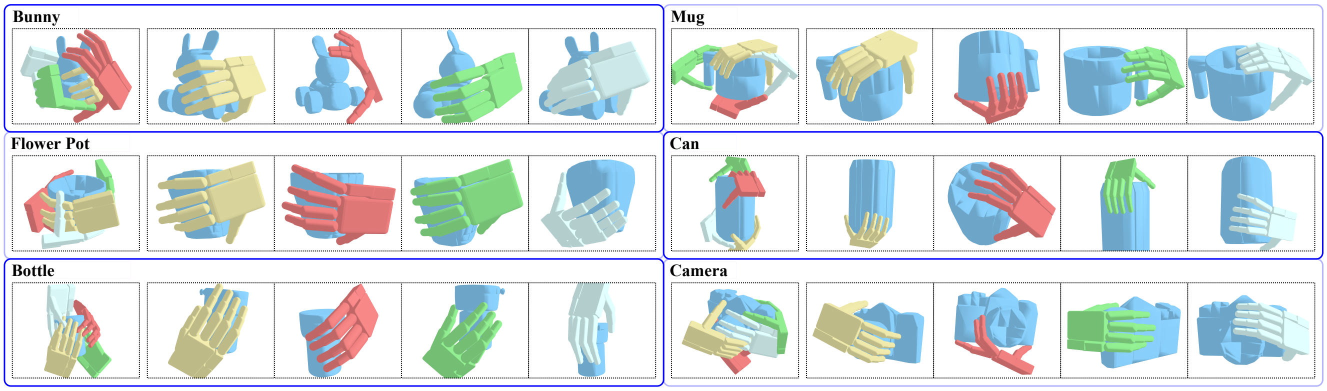

Figure 6. Visualization of predicted dexterous hand poses. We visualize four grasp poses in five images for each object. The first image visualizes all grasps together to demonstrate their global positions. The following four images mainly visualize the details of the grasp pose. These visualization results qualitatively indicate that the proposed DGTR framework is capable of generating diverse and feasible grasps with the same input and only **in one forward pass**. More visualization results can be found in *Appendix* C.

Let $\rho \in \mathcal{P}_N$ be a permutation of N elements, and assume that $K = M = N$. We utilize Hungarian Matching Algorithm [15] to compute the optimal assignment $\hat{\rho}$:

$$\hat{\rho} = \underset{\rho \in \mathcal{P}_N}{argmin} \sum_{i}^{K} \mathcal{C}(\mathbf{g}_i, \hat{\mathbf{g}}_{\rho_i}). \tag{6}$$

The process of computing $\hat{\rho}$ when $M \neq N$ is similar, except that we leave the redundant predictions or ground truths unmatched. As a result, there are $K = min\{M, N\}$ matched pairs accounting for the overall loss.

Overall Loss Function. The overall grasp loss for DGTR training is a weighted sum of the aforementioned losses, which is formulated as:

$$\begin{aligned} \mathcal{L}_{grasp}(\mathbf{g}_i, \hat{\mathbf{g}}_{\rho_i}, \mathcal{O}) = {} & \mathcal{L}_{param}(\mathbf{g}_i, \hat{\mathbf{g}}_{\hat{\rho}_i}) \\ & + \lambda_4 * \mathcal{L}_{chamfer}(\mathbf{g}_i, \hat{\mathbf{g}}_{\hat{\rho}_i}) \\ & + \lambda_5 * \mathcal{L}_{spen}(\mathbf{g}_i) + \lambda_6 * \mathcal{L}_{pen}(\mathbf{g}_i, \mathcal{O}). \end{aligned} \tag{7}$$

This loss is averaged among all matched pairs $\{\mathbf{g}_i, \hat{\mathbf{g}}_{\hat{\rho}_i}\}_{i=1}^{K}$.

4. Experiments

4.1. Dataset and Evaluation Metrics

We evaluate the proposed DGTR framework in the challenging dexterous grasping benchmark DexGraspNet [39], which contains 1.32 million grasps of ShadowHand [32] for 5355 objects from more than 133 object categories. The official training-validation split is used in our experiments.

We use five metrics to conduct comprehensive evaluations of the generating quality of DGTR. That is, 1) **Mean** Q_1 [8] reflects grasp stability. We follow [39] to set the contact threshold to 1cm and set the penetration threshold to 5mm. 2) **Maximal penetration depth (cm)** ($Pen.$), which is the maximal penetration depth from the object point cloud to hand meshes. 3) **Non-penetration ratio** η_{np} **(%)**, which is the proportion of the predicted hands with a maximal penetration depth of less than 5mm. 4) **Torque balance ratio** η_{tb} **(%)**, denoting the percentage of torque-balanced grasps (*i.e.* $Q_1 > 0$). 5) **Grasping success rate** $\eta_{success}$ **(%) in Isaac Gym** [23]. Following [39], we consider a grasp pose valid if the grasp can hold the object steadily under any one of the six gravity directions.

For diversity, we introduce the new metrics, 6) **occupancy proportion** of translations δ_t, rotations δ_r and joint angles δ_q (%), to quantitatively measure the ability of a model to grasp objects from a diverse range of directions, orientations, and joint angles. Generally, we discretize the continuous parameter space into $\xi = 16$ uniform bins and calculate the proportion of occupied spaces for different grasps of each object. For δ_t, we uniformly sample ξ points as the bins on a unit sphere with Fibonacci sampling, and then assign each grasp to a bin based on the cosine similarity between its global translation and the corresponding direction of the point. For δ_r and δ_q, we discretize the range of Euler angle into ξ bins. Intuitively, higher values of δ_t indicate that the predicted grasps can move to more areas of the object and grasp it from more directions, while higher δ_r and δ_q suggest more various hand orientations and gestures. All details of metrics can be found in *Appendix A*.

4.2. Implementation Details

Our DGTR is implemented with PyTorch [27] and trained on a single RTX 4090 GPU. The number of queries N is set to 16. The training epochs for each stage in DSMT are $T_0 = 15$, $T_1 = 5$, $T_2 = 5$. We set $\omega_1 = 2.0$, $\omega_2 = 1.0$, and $\omega_3 = 2.0$ for the Hungarian Algorithm cost function. During DMT and SMW, the loss weight are $\lambda_1 = 10.0$, $\lambda_2 = 10.0$, $\lambda_3 = 10.0$, $\lambda_4 = 1.0$, $\lambda_5 = 10.0$, $\lambda_6 = 0.0$. In the SMPT stage, λ_6 is set to 50.0, and distance loss weight is 10.0. For AB-TTA, we set $\beta_t = 0$, $\alpha_1 = 5$, $\alpha_2 = 3$, and $\alpha_3 = 5$. More details can be found in *Appendix* A.

Method	Quality					Diversity		
	$Q_1 \uparrow$	$\eta_{np} \uparrow$	$\eta_{tb} \uparrow$	$\eta_{success} \uparrow$	$Pen. \downarrow$	$\delta_t \uparrow$	$\delta_r \uparrow$	$\delta_q \uparrow$
GraspTTA [12]	0.0271	18.95	15.90	24.5	0.678	8.09	7.53	7.90
UniDexGrasp [41]	0.0462	**97.29**	50.94	37.1	0.121	9.64	7.49	29.29
SceneDiffuser [11]	0.0129	96.21	22.88	25.5	**0.107**	**54.84**	**52.27**	**39.75**
DGTR (ours)	**0.0515**	75.78	**69.62**	**41.0**	0.421	47.77	51.66	27.81
DDG [19]	0.0582	84.53	56.63	**67.5**	**0.173**	6.25	6.25	6.25
DGTR* (ours)	**0.0921**	**99.51**	**81.28**	66.6	0.313	**19.66**	**20.68**	**15.11**

Table 1. Results on DexGraspNet[39] compared with the state-of-the-art **in one forward pass** condition. DGTR* is a practical variant of DGTR concentrating on grasp quality. Note that DDG [19] is not in the same setting as ours and serves as a quality reference here.

Method	n_{pass}	n_{grasp}	T_{inf} (ms) $\downarrow$	$\delta_t \uparrow$	$\delta_r \uparrow$	$\delta_q \uparrow$
Uni. [41]	1	16	58.3 ± 4.1	9.64	7.49	**29.29**
Uni. [41]	4	4	153.7 ± 8.8	18.37	22.20	36.36
Uni. [41]	16	1	530.6 ± 12.2	25.04	44.31	38.65
Ours	1	16	$\mathbf{20.4 \pm 3.3}$	**47.77**	**51.66**	27.81

Table 2. Comparison with multiple pass methods. T_{inf} is the total time to generate all grasp poses. n_{pass} refers to the times of object's point cloud being rotated and passed to the grasping model. n_{grasp} is the number of grasp poses generated per pass.

Method	$Q_1 \uparrow$	$Pen. \downarrow$	$\eta_{np} \uparrow$	$\eta_{tb} \uparrow$
DGTR	0.0515	0.421	75.78	69.62
w/o AB-TTA	0.0278	0.466	52.36	65.10
w/o DSMT	0.0115	0.869	7.69	96.84

Table 3. Effectiveness of each component of DGTR.

4.3. Dexterous Grasp Generation Performance

4.3.1 Comparison with SOTA in one forward pass

We first compare SOTA dexterous grasp generation methods with DGTR in our setting, where each method is allowed to infer once. DDG [19] takes multi-view images as input and only predicts one grasp pose for each object, which serves as a quality reference. SceneDiffuser [11], GraspTTA [12] and UniDexGrasp [41] samples 16 times in a batch, with the same object point cloud as condition.

The evaluation results are shown in Table 1. For grasp quality, DGTR surpasses the SOTA generative models in several important metrics. Note that UniDexGrasp has remarkable performance in η_{np} and $Pen.$ but with a low η_{tb}, which suggests low contact with the object, while DGTR has a more balanced performance and higher success rate. Moreover, owing to the capability of generating diverse grasps, DGTR can efficiently select top-4 results (DGTR*) by the number of contact points and object penetration during inference without extra inputs. In this scenario, DGTR* has comparable results with DDG [19].

For diversity, DGTR surpasses UniDexGrasp [41] and GraspTTA [12] by a large gap in terms of δ_t and δ_r, which indicates that DGTR is able to grasp the object from a variety of directions. SceneDiffuser [11] has higher diversity but with much lower quality. More comparisons with SceneDiffuser are in *Appendix B*. The results demonstrate that DGTR achieves overall SOTA performance and excels in generating high-quality and diverse grasps.

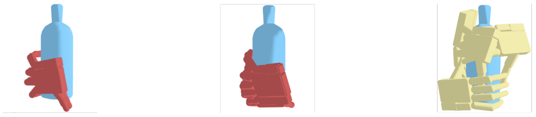
(a) GraspTTA [12] (b) UniDexGrasp [41] (c) Ours

Figure 7. Comparison of grasp diversity in one forward pass with 4 outputs. The diversity of our DGTR significantly surpasses [12] and [41] in one forward pass.

We visualize the predicted grasp poses of several objects in Figure 6 to provide a qualitative result of DGTR. DGTR is capable of generating high-quality grasps of an object from various directions with different poses in one forward pass. Furthermore, Figure 7 highlights the diversity of DGTR in comparison to two other generative methods.

4.3.2 Comparison with SOTA in multiple forward pass

Table 2 presents a comparison of grasping diversity and inference time between UniDexGrasp in multiple forward passes and DGTR in one forward pass. UniDexGrasp first utilizes a probabilistic model to sample rotations and then rotates object point clouds to generate grasps in multiple passes. As shown in Table 2, DGTR exhibits significantly lower time consumption compared to the multi-pass UniDexGrasp. More importantly, DGTR outperforms UniDexGrasp with 16 forward passes in δ_t and δ_r. This indicates that DGTR offers more diverse grasping hand positions and enables grasping from a wider range of directions.

4.4. Ablation Study

4.4.1 Dynamic-Static Matching Training Strategy

As demonstrated in Table 3, our DSMT significantly enhances Q_1 by 3.5 times, while reducing $Pen.$ by nearly 50%. Table 4 provides more details on the performance after each training stage (DMT, SMW and SMPT) in DSMT. The results highlight the critical role of static matching, which optimizes the model towards the proper direction and significantly reduces object penetration.

Method	$Q_1 \uparrow$	$Pen. \downarrow$	$\eta_{np} \uparrow$	$\eta_{tb} \uparrow$
DMT	0.0115	0.869	7.69	96.74
DMT + SMW	0.0100	0.879	6.55	97.25
DMT + SMW + SMPT	0.0278	0.466	52.36	65.10
w/o Static	0.0064	0.600	36.84	56.67
w/o Warm	0.0271	0.482	50.03	67.15

Table 4. Ablation study for three stages in DSMT. *Static* and *Warm* are static matching and warm-up for SMPT. The complete DSMT is colored in gray.

N	$Q_1 \uparrow$	$\delta_t \uparrow$	$\delta_r \uparrow$	$\delta_q \uparrow$
4	0.0392	18.40	21.85	9.60
8	0.0305	28.26	33.64	12.96
16	0.0278	47.77	51.66	27.81
32	0.0275	72.13	65.88	19.48
64	0.0170	89.57	78.41	25.50

Table 5. Analysis of the Number of grasp queries. N is the number of queries.

λ_{pen}	$Q_1 \uparrow$	$Pen. \downarrow$	$\eta_{np} \uparrow$	$\eta_{tb} \uparrow$
0	0.0115	0.869	7.69	96.84
5	0.0203	0.717	22.94	84.79
50	0.0109	0.662	36.76	60.62
500	0.0020	0.207	78.19	16.75
$0 \rightarrow 50$	0.0061	0.651	31.45	59.86

Table 6. Analysis of different object penetration loss weight. λ_{pen} is the penetration weight.

Pen	VDis	GDis	TM	CN	$Q_1 \uparrow$	$\eta_{np} \uparrow$	$\eta_{tb} \uparrow$
✓	✓				0	100	0
✓		✓			0.0125	77.15	28.08
✓			✓		0.0295	75.31	48.56
✓				✓	0.0435	98.54	50.50
✓		✓	✓	✓	0.0491	78.24	64.80
✓		✓	✓		0.0515	75.78	69.62

Table 7. Ablation study for designs in AB-TTA. *Pen* and *Dis* denote penetration and distance loss. *GDis* and *TM* are our generalized tta-distance loss and translation moderation strategy. *CN* refers to ContactNet [12]. Our whole AB-TTA is colored in gray.

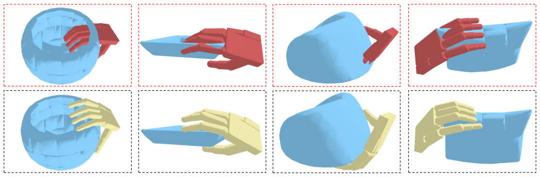

Figure 8. Visualization of grasps **before** and **after** AB-TTA.

4.4.2 Adversarial-Balanced Test-Time Adaptation

We conduct ablation studies on our AB-TTA module and the results are in Table 3, Table 7, and Figure 8. As shown in Table 3, our AB-TTA significantly increases Q_1 by 1.85-fold, and enhances η_{np}, and η_{tb} at the same time. Furthermore, Table 7 shows that the integration of our key designs (*i.e.*, generalized tta-distance loss (GDis) and translation moderation strategy (TM)) are indispensable, while the simple implementation of TTA (*i.e.*, penetration and vanilla distance loss (VDis)) has limited effect. Furthermore, our AB-TTA module demonstrates superior grasp quality compared to ContactNet-TTA [12], and it can even boost the Q_1 and η_{tb} performance of ContactNet-TTA.

4.5. DGTR Analysis

We conduct a series of analytic experiments for DGTR. We discuss object penetration weight and the number of queries below. Please refer to the *Appendix* B for more analysis.

4.5.1 Loss Weight for Object Penetration

The results in Table 6 show that the object penetration decreases as λ_{pen} increase, but a severe non-contact issue occurs concurrently. As illustrated in Figure 4 and Figure 5, the instability of Hungarian matching leads to model collapse when we apply a large penetration loss. And it is worth noting that gradually increasing λ_{pen} from 0 to 50 after several warm-up epochs cannot tackle this problem ($\lambda_{pen} = 0 \rightarrow 50$ in Table 6). We believe that learning to predict multiple grasps simultaneously is a more difficult optimization process compared to the previous one-to-one grasping learning. And our proposed progressive strategies (*i.e.*, DSMT and AB-TTA) tackle this challenge effectively.

4.5.2 Number of Grasping Queries

We conduct experiments to analyze the effect of the number of grasping queries. As shown in Table 5, the grasp quality Q_1 tends to decrease as the number of queries increases, which suggests that simultaneous learning a larger set of grasping poses is a challenge. Furthermore, the diversity increases as the number of queries becomes larger, implying that DGTR can learn a more diverse set of grasp with a greater number of queries.

5. Conclusions

In this work, we propose DGTR (Dexterous Grasp Transformer), a novel discriminative framework for dexterous grasp generation. Our progressive strategies, including dynamic-static matching training (DSMT) strategy and adversarial-balanced test-time adaptation (AB-TTA), substantially improve grasping stability and reduce penetration. To the best of our knowledge, DGTR is the first work to introduce set prediction formulation into dexterous grasp domain and achieves both high quality and diversity with one forward pass. We believe that DGTR holds good development potential in robotic dexterous grasping scenarios, such as task-oriented and real-world dexterous grasp generation.

Acknowledgements

We thank Jialiang Zhang for his helpful discussion. This work was supported in part by the National Key Research and Development Program of China (2023YFA1008503), NSFC (U21A20471, U1911401), and Guangdong NSF Project (No. 2023B1515040025, 2020B1515120085).

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, 2020. 2, 3

[2] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 3

[3] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 1, 2

[4] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 3

[6] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 5

[7] Enrico Maria Fenoaltea, Izat B Baybusinov, Jianyang Zhao, Lei Zhou, and Yi-Cheng Zhang. The stable marriage problem: An interdisciplinary review from the physicist's perspective. *Physics Reports*, 2021. 4

[8] Carlo Ferrari, J Canny, et al. Planning optimal grasps. In *Proceedings., 1992 IEEE International Conference on Robotics and Automation, 1992.*, 1992. 2, 6

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2015. 5

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 2

[11] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 7

[12] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the International Conference on Computer Vision*, 2021. 1, 2, 7, 8

[13] Jiayu Jiao, Yu-Ming Tang, Kun-Yu Lin, Yipeng Gao, Jinhua Ma, Yaowei Wang, and Wei-Shi Zheng. Dilateformer: Multi-scale dilated transformer for visual recognition. *IEEE Transactions on Multimedia*, 2023. 3

[14] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 2018. 2

[15] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 1955. 3, 6

[16] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 4

[17] Kelin Li, Nicholas Baron, Xian Zhang, and Nicolas Rojas. Efficientgrasp: A unified data-efficient learning to grasp method for multi-fingered robot hands. *IEEE Robotics and Automation Letters*, 2022. 2

[18] Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 1, 2

[19] Min Liu, Zherong Pan, Kai Xu, Kanishka Ganguly, and Dinesh Manocha. Deep differentiable grasp planner for high-dof grippers. *arXiv preprint arXiv:2002.01530*, 2020. 1, 2, 7

[20] Tengyu Liu, Zeyu Liu, Ziyuan Jiao, Yixin Zhu, and Song-Chun Zhu. Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. *IEEE Robotics and Automation Letters*, 2022. 2

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 3

[22] Jens Lundell, Francesco Verdoja, and Ville Kyrki. Ddgc: Generative deep dexterous grasping in clutter. *IEEE Robotics and Automation Letters*, 2021. 2

[23] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning, 2021. 6

[24] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2, 3

[25] Kieran A Murphy, Carlos Esteves, Varun Jampani, Srikumar Ramalingam, and Ameesh Makadia. Implicit-pdf: Non-parametric representation of probability distributions on the rotation manifold. In *International Conference on Machine Learning*, 2021. 2

[26] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 2021. 2

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 2019. 6

[28] J. Ponce, S. Sullivan, J.-D. Boissonnat, and J.-P. Merlet. On characterizing and computing three- and four-finger force-closure grasps of polyhedral objects. In *ICRA*, 1993. 2

[29] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 2017. 3

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 2

[31] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[32] ShadowHand. Shadowrobot. https://www.shadowrobot.com/dexterous-hand-series/, 2005. 3, 6

[33] Lin Shao, Fabio Ferreira, Mikael Jorda, Varun Nambiar, Jianlan Luo, Eugen Solowjow, Juan Aparicio Ojea, Oussama Khatib, and Jeannette Bohg. Unigrasp: Learning a unified model to grasp with multifingered robotic hands. *IEEE Robotics and Automation Letters*, 2020. 2

[34] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 2015. 2

[35] Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In *European Conference on Computer Vision*, 2022. 2

[36] Julen Urain, Niklas Funk, Jan Peters, and Georgia Chalvatzaki. Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 2

[37] Jacob Varley, Jonathan Weisz, Jared Weiss, and Peter Allen. Generating multi-fingered robotic grasps via deep learning. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 2015. 2

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017. 3

[39] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 2, 5, 6, 7

[40] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 2021. 3

[41] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 5, 7

[42] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. *arXiv preprint arXiv:2303.09833*, 2023. 1

[43] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 3

[44] Jiaming Zhou, Kun-Yu Lin, Yu-Kun Qiu, and Wei-Shi Zheng. Twinformer: Fine-to-coarse temporal modeling for long-term action recognition. *IEEE Transactions on Multimedia*, 2023. 3

[45] Tianqiang Zhu, Rina Wu, Xiangbo Lin, and Yi Sun. Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[46] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020. 3