# *GraspALL*: Adaptive Structural Compensation from Illumination Variation for Robotic Garment Grasping in Any Low-Light Conditions

Haifeng Zhong[1], Wenshuo Han[1], Zhouyu Wang[1], Runyang Feng[1], Fan Tang[2], Tong-Yee Lee[3], Zipei Fan[1], Ruihai Wu[4], Yuran Wang[4], Hao Dong[4], Hechang Chen[1,6], Hyung Jin Chang[5], Yixing Gao[1,6*]

[1] School of Artificial Intelligence, Jilin University, [2] Chinese Academy of Sciences, [3] National Cheng-Kung University, [4] Peking University, [5] University of Birmingham [6] Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MoE, China,

zhonghf23@mails.jlu.edu.cn, gaoyixing@jlu.edu.cn

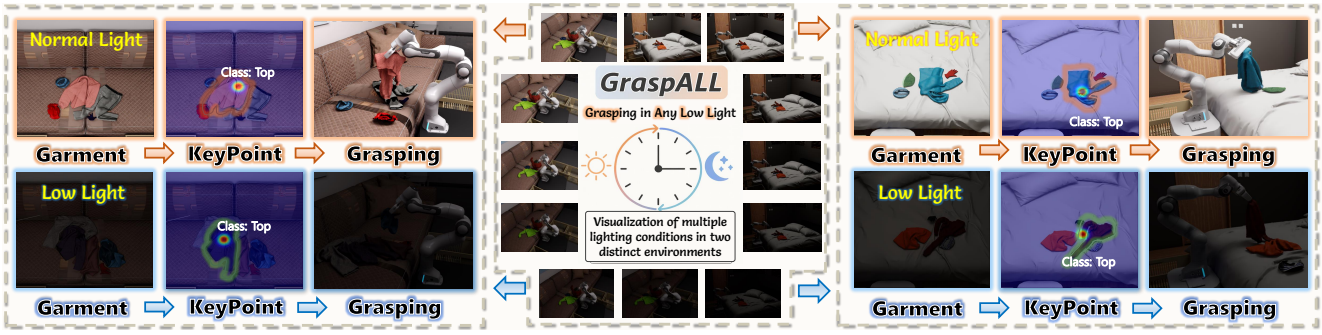


Figure 1. The grasping performance of our GraspALL in the living room (left) and bedroom (right) scenes under different illumination conditions. Given the variability of household scenes and lighting, service robots should possess all-day garment perception capabilities.

## Abstract

*Achieving accurate garment grasping under dynamically changing illumination is crucial for all-day operation of service robots. However, the reduced illumination in low-light scenes severely degrades garment structural features, leading to a significant drop in grasping robustness. Existing methods typically enhance RGB features by exploiting the illumination-invariant properties of non-RGB modalities, yet they overlook the varying dependence on non-RGB features under varying lighting conditions, which can introduce misaligned non-RGB cues and thereby weaken the model's adaptability to illumination changes when utilizing multimodal information. To address this problem, we propose GraspALL, an illumination-structure interactive compensation model. The innovation of GraspALL lies in encoding continuous illumination changes into quantitative references to guide adaptive feature fusion between RGB and non-RGB modalities according to varying lighting intensities, thereby generating illumination-consistent grasping representations. Experiments on the self-built garment grasping dataset demonstrate that GraspALL improves grasping accuracy by 32-44% over baselines under diverse illumination conditions. The code is available at https://github.com/Zhonghaifeng6/GraspALL*

## 1. Introduction

Garment grasping is a fundamental capability for service robots in daily tasks such as cleaning [27, 31] and dressing assistance [7, 8, 33]. While existing methods [4, 30, 31, 33] achieve high grasping accuracy under normal illumination, lighting conditions in real household environments are often dynamic. For instance, in patients, elderly and infant care scenarios, robots are frequently required to operate in low-light or even unlit environments to avoid disturbance. As shown in Fig. 1, this necessitates that robots possess all-day perceptual capability, ensuring reliable performance under any illumination. However, reduced illumination severely degrades garment texture, wrinkles, and edge details, thereby diminishing the robustness of grasping.

To address the perceptual degradation caused by illumination variations, existing methods [15, 35, 39, 44] generally adopt multimodal fusion [40–42] to enhance RGB features by leveraging the structural illumination invariance of non-RGB modalities (e.g., depth map). Although the above

* Corresponding author

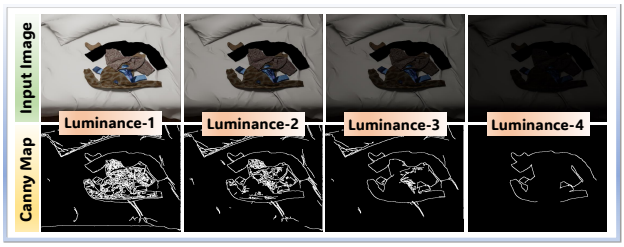

Figure 2. Structure maps generated by the Canny operator from identical scenes under different luminance levels.

methods are effective, they often overlook the differences in the model's requirements for structural features of non-RGB under varying illumination conditions. As shown in Fig. 2, illumination changes can significantly distort the geometry of garments, resulting in inconsistent structural cues for the same garment. Such illumination-induced discrepancies cause the feature response of RGB under illumination variations are suppressed by the stronger structural signals of non-RGB, leading the model to prioritize non-RGB cues over subtle yet critical RGB luminance information, thereby reducing robustness to illumination changes.

Based on the above analysis, we hold that a more rational paradigm is to enable the model to perceive the input illumination levels and then extract appropriate structural compensation from non-RGB modalities according to different illumination levels, thereby conditionally enhancing garment representations [34]. This paradigm facilitates the collaborative fusion of cross-modal features to cope with illumination variations, while we also identify two challenges it poses: **(1)** how to accurately estimate the input illumination level to provide quantitative guidance for cross-modal feature fusion; **(2)** based on the estimation of illumination levels, how to induce non-RGB to generate structural compensation adaptive to illumination changes.

To address the above challenges, as shown in Fig. 3, we propose a novel model capable of adapting to illumination variations to enable garment **Grasp**ing in **A**ny **L**ow-**L**ight (named **GraspALL**). Unlike methods [15, 35, 39, 44] that treat non-RGB as static supplements, the innovation of our GraspALL lies in pioneering a parametric luminance representation method and an illumination-adaptive structural compensation strategy to guide non-RGB in adaptively enhancing garment features in RGB according to different illumination levels. Specifically, we first propose a parametric luminance curve (PLC), which fits representative luminance patterns of inputs under different illumination via learnable parameter sets, enabling general representation of any illumination level. Based on the input luminance estimated by the PLC, we derive the required luminance compensation features during the luminance restoration process to drive depth maps to generate corresponding structural compensation features. We then calculate the feature correlation scores between depth maps and luminance compensation features to suppress the weights of features incompatible with the current illumination level, thus obtaining illumination-adaptive structural compensation features for supporting grasp point modeling process.

Moreover, considering the scarcity of multi-illumination garment grasping datasets, we construct a dataset for grasping tasks under diverse illumination conditions. Unlike previous works [20, 43, 49], the novelty of our dataset lies in that it consists of diverse garment categories, covers typical layouts including sofa and bed furniture, and incorporates diverse illumination variations from bright to dim—thus simulating household scenes with dynamic illumination changes. Experiments conducted under different illumination levels show that compared with baselines, our GraspALL can improve grasping accuracy by 32%–44%.

The contributions of this work can be summarized as:

- We present the first systematic analysis of how dynamic illumination variations affect garment grasping, uncovering critical challenges overlooked by existing methods and offering new insights for designing illumination-robust garment grasping models.
- We propose a GraspALL model for garment grasping in varying illumination, which can guide non-RGB to conditionally enhance RGB garment features from a illumination adaptation perspective.
- We introduce a new task of garment classification and grasping under illumination variation, and establish a benchmark comprising a large-scale dataset and diverse household scenarios, providing a unified testbed for evaluating garment grasping under illumination changes.

## 2. Related work

**Garment Grasping.** Garment grasping has broad applications in household scenarios. Existing methods commonly rely on object detection [3, 14, 50], relation detection [6, 18, 19, 50], or learning-based strategies [3, 9, 26, 47] to achieve grasping; some further incorporate semantic segmentation [1, 3, 21, 23] to enhance perception of garments. While the above methods have achieved notable progress, they generally assume stable image quality and thus struggle to cope with the dynamic illumination variations that naturally occur across time and space in household scenes. When the illumination changes from light to dark, garment features often undergo severe degradation, leading to a significant drop in the robustness of grasp point prediction. In contrast, our work provides a detailed analysis of the challenges posed by illumination variation for garment grasping, and addresses these challenges by reinforcing cross-modal features to capture critical garment representations, thereby enabling more accurate grasping.

**Garment Grasping of Illumination Variations.** Faced with dynamically changing illumination, existing methods [16, 36, 37, 43, 48] focus on incorporating non-RGB modalities such as depth maps, exploiting their

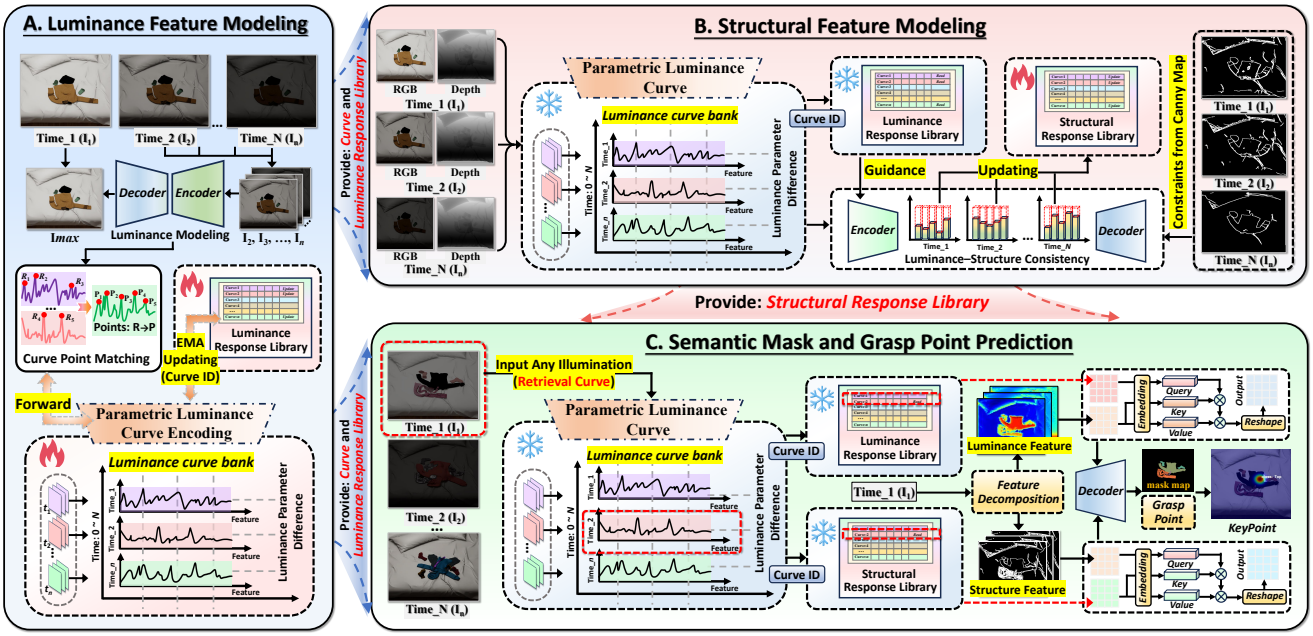


Figure 3. Overview of our GraspALL. (A) describes the generation process of the Parametric Luminance Curve and luminance compensation features (in Sec. 3.1). (B) illustrates the generation process of structural compensation features adaptive to the input luminance (in Sec. 3.2). (C) outlines the generation process of grasp points based on the luminance and structural compensation features (in Sec. 3.3).

illumination-invariant structural properties to complement RGB, thereby improving grasping performance under low light. Although the above methods are partially effective, they largely overlook how to achieve dynamic complementarity between RGB and non-RGB features under illumination variation. Since illumination changes dynamically affect garment structural characteristics, the model's reliance on non-RGB structural features should vary across different lighting conditions. In contrast, our work is the first to systematically address the impact of illumination variation on garment grasping, and the proposed GraspALL enables grasping across arbitrary illumination conditions by guiding depth maps to adapt to illumination-caused degradation.

## 3. Method

To tackle the challenges of illumination variation in garment grasping, as shown in Fig. 3, we propose GraspALL, a grasp point recognition model built on luminance–structure interactive compensation. GraspALL consists of three core components: the parametric luminance curve, the luminance response library, and the structural response library.

### 3.1. Luminance Feature Modeling

Traditional luminance estimation methods [10, 29] typically rely on histograms [11, 34], but the non-learnable nature of histograms makes it difficult to adapt to diverse illumination variations. To address this, we present a learnable Parametric Luminance Curve (PLC), whose distinctiveness lies in adopting a learnable parameter set to uniformly represent the representative luminance patterns across various illuminations, thereby generating more robust luminance interpretations. The performance of the PLC is verified in Sec.4.6.

Firstly, we need to define a luminance curve bank $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_N\}, N = 12$, each curve $\mathbf{C}_n \in \mathbb{R}^R$ consists of $R$ discrete sampling points parameterized by learnable raw parameters $\mathbf{P}_n = \{\mathbf{P}_{n,1}, \mathbf{P}_{n,2}, ..., \mathbf{P}_{n,R}\}, R = 256$. $\mathbf{P}_n$ is a learnable parameter that enables the model to adaptively learn the optimal brightness curve of different illumination conditions. The analysis of $N$ and $R$ in luminance curve bank $\mathbf{C}$ is provided in the supplementary materials.

Next, we need to define a learning objective for the luminance curve bank $\mathbf{C}$. Given a set of images $\{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_N\}$ under varying illumination, we select the image $\mathbf{I}_{max}$ with the highest luminance based on histogram statistics as the reference luminance image. For the other images $\mathbf{I}_n$ except $\mathbf{I}_{max}$, we compute $R$ representative luminance values using histograms with $R$ intervals to form a set $\mathbf{H}_R$, and then identify the curve that has the most matching points with $\mathbf{H}_R$, obtaining the corresponding curve index $\mathbf{ID}_n$:

$$\mathbf{ID}_n = \mathbf{argmin} \left|\left|\mathbf{H}_i - \mathbf{C}\left(\mathbf{P}_{n,i}\right)\right|\right|, i \in R, n \in N. \quad (1)$$

We take the reference luminance image $\mathbf{I}_{max}$ as the luminance anchor and align the luminance features of other images $\mathbf{I}_n$ to $\mathbf{I}_{max}$ through a shared encoder–decoder:

$$\mathbf{I}^n_{max} = \mathcal{D}\left(\mathcal{E}\left(\mathbf{I}_n\right)\right) \leftarrow \mathcal{L}_{sc}\left(\mathbf{I}_{max} - \mathbf{I}^n_{max}\right)_{\mathbf{L1}}, \quad (2)$$

where $\mathcal{E}(\cdot)$ and $\mathcal{D}(\cdot)$ denote the encoder and decoder [22], and encoder features: $\mathbf{F}^n_{en} = \mathcal{E}\left(\mathbf{I}_n\right)$. "$\leftarrow$" represents the loss supervision. $\mathcal{L}_{sc}$ represents spectral consistency loss.

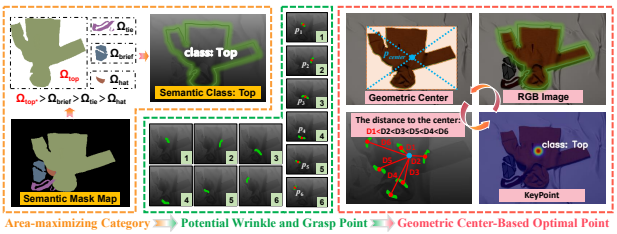

Figure 4. Flowchart of proposed depth-optimal search strategy.

By aligning features with the brightest sample, the model is able to gain the luminance compensation features essential for restoring the luminance of inputs. These features not only enhance garment distinguishability but also implicitly reflect structural deficiencies caused by illumination variations, allowing them to further guide depth maps to generate corresponding structural compensation features. To this end, we construct a Luminance Response Library $\mathbf{M}_L$, to store luminance features corresponding to different illumination conditions: $\mathbf{M}_L = \{\mathbf{M}_L^1, \mathbf{M}_L^2, ..., \mathbf{M}_L^N\}$.

Based on the index $\mathbf{ID}_n$, the encoder feature $\mathbf{F}_{en}^n$ of the alignment process is updated into the corresponding slot of $\mathbf{M}_L$ using an exponential moving average (EMA) [2]:

$$\mathbf{M}_L = (1-\alpha)\,\mathbf{M}_L^n + \alpha \cdot \mathbf{F}_{en}^n, n = \mathbf{ID}_n, \tag{3}$$

where $\alpha = 0.05$ is the EMA momentum. The analysis of the $\alpha$ is provided in the supplementary materials.

To enable the adaptive learning of the luminance curve bank $\mathbf{C}$, we introduce a spectral consistency loss $\mathcal{L}_{sc}$ [45]. $\mathcal{L}_{sc}$ emphasizes clustering consistency by minimizing the L1 distance between the luminance features $\mathbf{F}_{en}^n$ of the current input and the corresponding slot features in $\mathbf{M}_L$, thereby encouraging the curve bank to dynamically adjust its parameters for more accurate luminance indexing. By minimizing $\mathcal{L}_{sc}$, the network optimizes the selected curve parameters $\mathbf{P}_n$ through chained gradient backpropagation:

$$\frac{\partial(\mathcal{L}_{sc})}{\partial(\mathbf{P}_n)} = \frac{\partial(\mathcal{L}_{sc})}{\partial(\mathbf{M}_L)} \cdot \frac{\partial(\mathbf{M}_L)}{\partial(\mathbf{F}_{en}^n)} \cdot \frac{\partial(\mathbf{F}_{en}^n)}{\partial(\mathbf{C})} \cdot \frac{\partial(\mathbf{C})}{\partial(\mathbf{P}_n)}, \tag{4}$$

where $\partial(\cdot)$ denotes the gradient of the current parameters with respect to the output parameters of the previous layer.

Eqs.1–4 uniquely emphasize the learnability of illumination interpretation, and they ultimately generate a luminance curve bank and the luminance response library that are used to guide the generation of the structural response library detailed in Sec.3.2. Although real-illumination involves multiple sources, reflections, and material effects, PLC is primarily designed to capture single and important illumination factor that influence garment appearance, and other factors will be investigated in the future work.

## 3.2. Structural Feature Modeling

Although structurally stable in extreme low light, depth maps lack discriminative appearance details. Because highly deformable garments yield similar geometric structures, relying solely on depth causes class confusion. Conversely, degraded RGB images still preserve complementary semantic cues (detailed analysis in Suppl. Sec.1). However, existing methods [35, 44] are inherently static and struggle with dynamic illumination. To address this, our structural modeling conditionally extracts adaptive depth features guided by explicit illumination understanding.

Given a set of images $\{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_N\}$ ranging from dark to bright, for one of the images $\mathbf{I}_n$, we employ Eq. 1 to match the corresponding curve ID from the luminance curve bank and retrieve the luminance feature $\mathbf{M}_L^n$ from the corresponding slot in $\mathbf{M}_L$. Based on the $\mathbf{M}_L^n$, we guide the model to extract complementary structural features from the depth map $\mathbf{I}_{dep}$. We first encode the $\mathbf{I}_{dep}$: $\mathbf{F}_{en}^{de} = \mathcal{E}(\mathbf{I}_{dep})$.

Then, we apply a linear layer to $\mathbf{M}_L^n$ to compute $Q_{lu}$, and apply another linear layer to $\mathbf{F}_{en}^{de}$ to compute $K_{de}$ and $V_{de}$. Using $Q_{lu}$ to query $K_{de}$ yields the matching scores between the luminance feature $\mathbf{M}_L^n$ and the structural features $\mathbf{F}_{en}^{de}$:

$$\begin{aligned} Q \cdot K &= \underbrace{Q_{lu} \in \mathbb{R}^{HW \times C}}_{\text{luminance}} \times \underbrace{K_{de} \in \mathbb{R}^{HW \times C}}_{\text{structure}}, \\ Score &\in \mathbb{R}^{HW \times C} = \text{Softmax}\,(Q \cdot K)\,. \end{aligned} \tag{5}$$

These scores highlight the varying attention that the luminance feature $\mathbf{M}_L^n$ assigns to valid and invalid information within the depth map features. Then, $V_{de}$ is modulated according to these scores and reshaped to obtain the structural compensation feature $\mathbf{F}_{en}^s$:

$$\mathbf{F}_{en}^s \in \mathbb{R}^{H \times W \times C} = \text{Reshape}\,(Score \times V_{de})\,. \tag{6}$$

To constrain the above structural modeling process, we introduce a Canny map. For the input image set $\{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_N\}$, we first use a histogram to identify the brightest image $\mathbf{I}_{max}$, and process it with a Canny detector to generate a $\mathbf{S}_{can}$, which serves as a reference for the structure modeling. Under the constraint of the $\mathbf{S}_{can}$, $\mathbf{F}_{en}^s$ is decoded using a decoder to produce the structural map:

$$\mathbf{S}_{can}^{dep} = \mathcal{D}(\mathbf{F}_{en}^s) \leftarrow \mathcal{L}_{bce}\left(\mathbf{S}_{can}^{dep}, \mathbf{S}_{can}\right), \tag{7}$$

where $\mathcal{L}_{bce}$ is the binary cross entropy. $\mathbf{S}_{can}$ is the structure map extracted from $\mathbf{I}_{max}$ using the Canny detector.

Similar to Eq. 3, we introduce a Structural Response Library (SRL: $\mathbf{M}_S = \{\mathbf{M}_S^1, \mathbf{M}_S^2, ..., \mathbf{M}_S^N\}$) to store the structural compensation features $\mathbf{F}_{en}^s$ in the corresponding slot of $\mathbf{M}_S$ according to the current curve ID, and update them using the EMA strategy. $\mathbf{M}_S$ will support the structural modeling in the grasp point prediction described in Sec. 3.3.

## 3.3. Semantic Mask and Grasp Point Prediction

We divide grasp point prediction into two closely connected stages: identifying the semantic mask regions of garments (Sec. 3.3.1) and determining the optimal grasp points based on these semantic mask regions (Sec. 3.3.2).

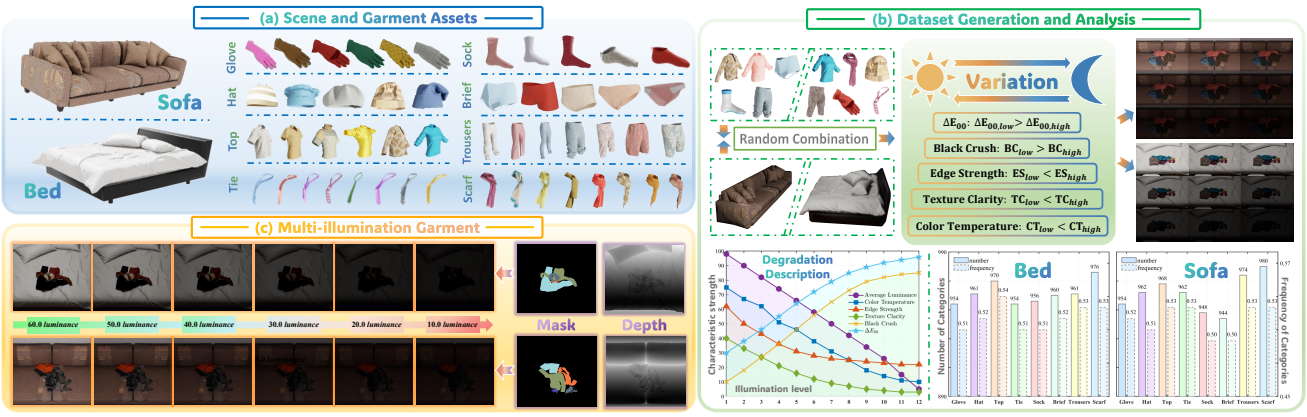


Figure 5. Construction of our MIGG dataset. (a) shows the household scenes and eight categories of garment assets. (b) shows that we obtain images under different illumination conditions by randomly combining the assets with the scenes. We undergo degradation in multiple dimensions such as color difference ($\Delta\mathbf{E}_{00}$), black crush, texture clarity, and color temperature, rather than simple luminance attenuation. The lower part of (b) describes the dataset's degradation status and the number/frequency of each garment category. (c) presents the captured garment images, depth map and semantic mask map. "**Luminance**" is calculated by averaging the histograms.

### 3.3.1 Garment Category Recognition

Garment grasping involves diverse garment types, while existing method [31] primarily relies on point cloud signals to model grasp points, neglecting semantic category information and thus struggling with multi-category grasping. In contrast, our method uses semantic information to enable garment grasping tailored to different categories. For an input captured under arbitrary illumination, we obtain the corresponding curve ID from the curve library in the same manner as Eq. 1. Using the curve ID, we retrieve the associated luminance feature $\mathbf{M}_L$ and structural feature $\mathbf{M}_S$ from the luminance and structural response libraries. To enable complementary enhancement of luminance and structure based on these retrieved features, we perform luminance–structure feature decomposition on the input $\mathbf{I}_n$:

$$\mathbf{I}_L \in \mathbb{R}^{H\times W\times 3}, \mathbf{I}_S \in \mathbb{R}^{H\times W\times 1} = \mathcal{N}_{Retinex}\left(\mathbf{I}_n\right), \quad (8)$$

where $\mathbf{I}_L$ and $\mathbf{I}_S$ denote the decomposed luminance and structural maps, respectively. We use the network from [38] as the image decomposition network $\mathcal{N}_{Retinex}\left(\cdot\right)$.

Then, we use the encoder to encode $\mathbf{I}_L$ and $\mathbf{I}_S$ to obtain the luminance feature $\mathbf{F}_L$ and the structural feature $\mathbf{F}_S$. Based on the feature $\mathbf{F}_L$, we apply a linear layer to compute $\mathbf{F}_L$ as $Q_{lu}$, and compute $K_m$ and $V_m$ from the feature $\mathbf{M}_L$. The same with Eq. 5, we use $Q_{lu}$ to query $K_m$, we obtain the matching scores between the feature $\mathbf{F}_L$ and the features $\mathbf{M}_L$ from the response library. These scores reflect the missing luminance features in the input, which is then modulated by multiplying with $V_m$ to generate the enhanced luminance feature $\mathbf{F}_L^{en}$. Similarly, we compute the matching scores between the structural feature $\mathbf{F}_S$ and the corresponding features $\mathbf{M}_S$ in the structural response library, thereby obtaining the enhanced structural feature $\mathbf{F}_S^{en}$.

Finally, we concatenate the enhanced structural and luminance features, followed by feature mapping and decoding operations to obtain the semantic mask map $\mathcal{M}_m$:

$$\begin{aligned}\mathcal{F} &= \text{MLP}\left(\text{Concatenate}\left(\mathbf{F}_S^{en}, \mathbf{F}_L^{en}\right)\right), \\ \mathcal{M}_m &= \mathcal{D}\left(\text{Reshape}\left(\mathcal{F}\right)\right) \leftarrow \mathcal{L}_{ce}\left(\mathcal{M}_m, gt\right),\end{aligned} \quad (9)$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss used to constrain the decoder in generating the semantic mask, and $gt$ is the label.

### 3.3.2 Category-Specific Grasp Point Generation

In Fig. 3, based on the mask map $\mathcal{M}_m$, our goal is to identify the graspable regions of different garment categories and determine the optimal grasp point using the corresponding depth map. Previous methods [5, 25, 43] typically define the grasp point as the center of each garment in $\mathcal{M}_m$, which reduces dragging during grasp execution. However, due to the high deformability of garments, the geometric center does not always correspond to graspable folds, often leading to unstable grasping. To address this, as shown in Fig. 4, we propose a depth-optimal search strategy that ensures both stability during the grasping process.

Given a mask $\mathcal{M}_m$ containing multiple garment classes $C$, we first select the class $c^*$ with the largest area $\Omega_{c^*}$:

$$\Omega_{c^*} = \underset{c\in C}{\text{argmax}}\left|\Omega_c\right|, \quad (10)$$

where $\Omega_c$ denotes the pixel set of class $c$ in $\mathcal{M}_m$. This ensures that grasping is performed on the dominant garment region. Within the selected region $\Omega_{c^*}$, we extract $k$ pixels with the smallest depth values (i.e., the closest to the camera) based on the corresponding area of the depth map, representing the most accessible surface points:

$$p_1, p_2, ..., p_k = \text{Depth}_{top}\left(\Omega_{c^*}\right), \quad (11)$$

where $p_1, p_2, ..., p_k$ are the points where the depth value is optimal. The above process effectively identifies geometri-

| Class | Luminance: 90 – 120 | | | | Luminance: 60 – 90 | | | | Luminance: 30 – 60 | | | | Luminance: 0 – 30 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| —— | SegMiF | MRFS | AMDA | **Ours** | SegMiF | MRFS | AMDA | **Ours** | SegMiF | MRFS | AMDA | **Ours** | SegMiF | MRFS | AMDA | **Ours** |
| **Glove** | 74.9% | 77.9% | 76.6% | **84.4%** | 70.1% | 71.1% | 75.5% | **84.2%** | 61.9% | 62.5% | 70.7% | **83.1%** | 62.3% | 61.8% | 68.5% | **81.7%** |
| **Hat** | 82.8% | 80.1% | 81.5% | **86.4%** | 74.6% | 78.7% | 80.3% | **85.8%** | 71.8% | 70.6% | 75.9% | **85.5%** | 65.4% | 68.2% | 71.3% | **84.7%** |
| **Scarf** | 76.0% | 82.2% | 79.1% | **85.7%** | 73.3% | 75.1% | 77.0% | **86.3%** | 65.4% | 65.7% | 73.1% | **85.7%** | 66.3% | 63.5% | 68.6% | **85.6%** |
| **Sock** | 70.9% | 67.2% | 72.8% | **83.5%** | 66.2% | 65.7% | 68.0% | **82.2%** | 62.3% | 58.9% | 65.5% | **80.3%** | 56.9% | 61.8% | 59.3% | **79.2%** |
| **Tie** | 72.5% | 80.3% | 79.9% | **84.7%** | 70.7% | 73.9% | 74.4% | **84.1%** | 64.1% | 66.2% | 71.1% | **82.1%** | 60.8% | 65.3% | 66.9% | **81.4%** |
| **Top** | 73.5% | 70.1% | 75.9% | **86.2%** | 71.6% | 73.2% | 72.8% | **85.8%** | 74.4% | 73.4% | 72.4% | **85.2%** | 71.4% | 70.4% | 72.5% | **84.4%** |
| **Trousers** | 68.8% | 73.9% | 74.1% | **84.3%** | 65.4% | 69.6% | 70.4% | **82.9%** | 61.1% | 63.9% | 67.5% | **82.7%** | 60.4% | 61.7% | 65.0% | **81.5%** |
| **Brief** | 78.2% | 78.4% | 79.1% | **83.5%** | 72.2% | 73.0% | 75.9% | **83.3%** | 67.2% | 67.7% | 76.5% | **82.1%** | 57.1% | 62.5% | 65.7% | **80.2%** |
| **mIoU** | 74.7% | 76.2% | 77.3% | **84.8%** +14% | 70.5% | 72.6% | 74.4% | **84.3%** +14% | 66.0% | 66.3% | 71.5% | **83.4%** +17% | 62.5% | 64.2% | 67.3% | **82.8%** +20% |

Table 1. Quantitative comparison of semantic mask generation accuracy between our GraspALL and other baseline methods under different luminance levels. The best results are highlighted in **bold**, and performance improvements are shown in **red**.

| Class | Luminance: 80 – 120 | | | | | Luminance: 40 – 80 | | | | | Luminance: 0 – 40 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| —— | BiFCNet | SAM-M | ReKep | DarkSeg | **Ours** | BiFCNet | SAM-M | ReKep | DarkSeg | **Ours** | BiFCNet | SAM-M | ReKep | DarkSeg | **Ours** |
| **Glove** | 9/15 | 10/15 | 8/15 | 11/15 | **14/15** | 8/15 | 9/15 | 8/15 | 10/15 | **14/15** | 5/15 | 5/15 | 6/15 | 7/15 | **12/15** |
| **Hat** | 11/15 | 10/15 | 9/15 | 13/15 | **15/15** | 9/15 | 8/15 | 8/15 | 9/15 | **13/15** | 6/15 | 7/15 | 6/15 | 8/15 | **13/15** |
| **Scarf** | 10/15 | 9/15 | 10/15 | 12/15 | **14/15** | 8/15 | 8/15 | 7/15 | 8/15 | **13/15** | 6/15 | 6/15 | 5/15 | 7/15 | **13/15** |
| **Sock** | 9/15 | 7/15 | 10/15 | 11/15 | **13/15** | 7/15 | 7/15 | 7/15 | 9/15 | **13/15** | 5/15 | 5/15 | 6/15 | 7/15 | **11/15** |
| **Tie** | 9/15 | 10/15 | 8/15 | 10/15 | **13/15** | 6/15 | 5/15 | 6/15 | 7/15 | **11/15** | 3/15 | 5/15 | 5/15 | 6/15 | **10/15** |
| **Top** | 10/15 | 8/15 | 11/15 | 13/15 | **15/15** | 8/15 | 9/15 | 9/15 | 11/15 | **14/15** | 8/15 | 8/15 | 9/15 | 10/15 | **14/15** |
| **Trousers** | 9/15 | 11/15 | 11/15 | 13/15 | **14/15** | 10/15 | 10/15 | 10/15 | 12/15 | **14/15** | 9/15 | 10/15 | 8/15 | 12/15 | **14/15** |
| **Brief** | 7/15 | 9/15 | 9/15 | 11/15 | **14/15** | 7/15 | 6/15 | 8/15 | 10/15 | **14/15** | 6/15 | 6/15 | 6/15 | 7/15 | **13/15** |
| **mGSR** | 61.6% | 59.2% | 63.4% | 78.3% | **93.3%** +32% | 52.4% | 51.6% | 52.4% | 63.3% | **88.3%** +36% | 39.9% | 43.3% | 42.4% | 53.3% | **84.2%** +44% |

Table 2. Comparison of garment grasping accuracy between our GraspALL and other baseline methods under different luminance levels.

cally salient points on the garment surface, such as wrinkles or protrusions, thereby enhancing grasping stability.

Following the [43], we compute the geometric center $p_{center}$ of the current semantic region $\Omega_{c^*}$ by fitting the minimum bounding rectangle. Then, among the candidate points $p_1, p_2, ..., p_k$, we select the pixel closest to $p_{center}$ as the optimal grasping point $p_o$:

$$p_o = \underset{p \in p_k}{\text{argmin}} \, ||p - p_{center}||_2, \tag{12}$$

where $|| \cdot ||_2$ is the euclidean distance. By searching for the grasping point across the entire semantic region, our method avoids dragging issues caused by off-center grasping while ensuring that the selected point lies in a structurally stable, wrinkled area. After completing this process, we iteratively repeat the above steps to grasp the remaining garment.

# 4. Experiments

## 4.1. Experimental Setup

**Dataset:** To overcome the scarcity of garment grasping datasets under varying illumination, as shown in Fig. 5, we construct a **M**ulti-**I**llumination **G**arment **G**rasping (**MIGG**) dataset using NVIDIA Isaac Sim. Unlike previous datasets [20, 43, 49] with limited garment types, uniform lighting, and simplified scenes, MIGG features controllable illumination and realistic household scenes. We create two representative household scenes—a living-room sofa and a bedroom bed. A garment asset library with eight categories (top, brief, glove, hat, tie, trousers, skirt, and sock) is developed, each modeled with realistic fabric properties and deformable dynamics. Illumination is physically controlled through variations in intensity, direction, and color temperature, generating multiple brightness levels from normal to extreme low-light. For each configuration, Isaac Sim captures synchronized RGB image, depth map, and semantic mask map at 512×512 resolution, yielding **15384** image triplets, split into **13008** for training and **2376** for testing.

**Metrics:** We employ two complementary metrics to evaluate both semantic mask accuracy and mask-based garment grasping performance. For semantic mask evaluation, **mIoU** (mean intersection over union) [28] is used to measure accuracy. For grasping evaluation, we design a multi-category garment grasping protocol, where each garment category is assigned a corresponding target basket. The model predicts grasp points from the semantic mask and executes grasp-and-place operations. A trial is considered successful if the garment is correctly grasped and placed into its target basket. Each method is tested 15 times under every illumination level, and the **mGSR** (mean grasping success rate) [17] is reported as the final evaluation metric.

**Baselines:** GraspALL is trained using an NVIDIA 4090 GPU. We compare GraspALL with representative baselines across both semantic mask generation and garment grasping. For comparison of semantic mask, we select three multimodal fusion methods — SegMiF [15], MRFS [35], and AMDA [44]. For garment grasping comparison, we select BiFCNet [46], SAM-M [13], ReKep [12], and DarkSeg [43] as baselines. SegMiF, MRFS, AMDA, BiFCNet and DarkSeg will be retrained on our proposed dataset based on their original configuration. SAM-M and ReKep use the official weights, and the prompt of ReKep is modified to: "In low-

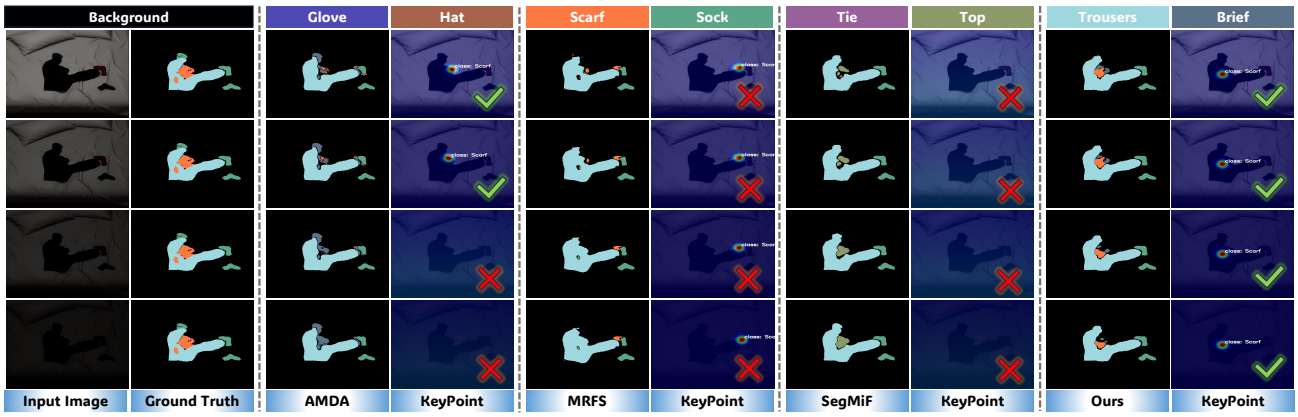


Figure 6. Comparison of semantic mask and grasping key points between our GraspALL and baselines under different luminance levels.

| Model | PLC | LRL | SRL | $\mathcal{L}_{sc}$ | $\mathcal{L}_{bce}$ | mIoU | mGSR |
|---|---|---|---|---|---|---|---|
| **Model-1** | ✗ | | | | | 65.4% | 50.0% |
| **Model-2** | | ✗ | | | | 71.3% | 72.5% |
| **Model-3** | | | ✗ | | | 68.5% | 57.5% |
| **Model-4** | | | | ✗ | | 64.9% | 50.2% |
| **Model-5** | | | | | ✗ | 68.7% | 70.4% |
| **Model-6** | ✔ | ✔ | ✔ | ✔ | ✔ | **82.6%** | **88.3%** |

Table 3. Ablation study of different components (Lum: 0-40).

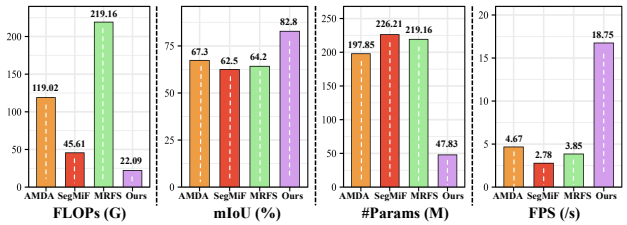


Figure 7. Comparisons of Parameters, FPS, mIoU and FLOPs.

light indoor environment, grasping the garment.."

- BiFCNet is chosen to show the limitations of the traditional garment grasping method for illumination levels.
- SAM-M and ReKep are selected to prove that even with the driving force of powerful large models, it remains difficult to handle complex low-light scenarios. SAM-M is our modification of SAM based on [31].
- DarkSeg is chosen to verify that existing low-light garment grasping methods, though capable of addressing partial low-light conditions, struggle to handle diverse low-light situations under dynamic illumination.

### 4.2. Quantitative Analysis

Tab. 1 presents a quantitative comparison between our method and other multimodal models. Although other methods also leverage depth modality to enhance RGB features, they exhibit significant performance degradation under illumination variation. For instance, the MRFS suffers a 12% drop in mIoU when transitioning from bright to low light. In contrast, GraspALL maintains stable performance across different brightness levels, with fluctuations under 2%. This robustness stems from our proposed parametric luminance representation, which enables dynamic adaptation of feature distributions to illumination shifts, ensuring stable perception under diverse lighting conditions.

Tab. 2 reports the quantitative comparison of garment grasping performance. As shown, our GraspALL consistently outperforms all baselines across luminance ranges, achieving 83.3% mGSR under extreme low-light conditions (0–40), representing a 31% improvement over the second-best method. Even under medium brightness variations (40–120), GraspALL maintains a stable grasping success rate between 88.3%–93.3%, demonstrating strong robustness to illumination changes. These results confirm the effectiveness of our parametric luminance representation and adaptive compensation mechanism in achieving generalized and reliable grasping across complex lighting scenarios.

### 4.3. Qualitative Analysis

Fig. 6 provides qualitative comparisons on semantic mask generation and grasp-point prediction. As illustrated, existing methods (e.g., SegMiF) produce blurred mask boundaries under varying illumination, failing to capture fine differences between garments and resulting in misplaced grasp points. In contrast, GraspALL preserves clear boundaries and consistent structural details even under low and dynamic lighting, with predicted grasp points concentrated on geometrically stable regions (e.g., wrinkles near the center). This demonstrates its superior precision and stability in grasping across challenging illumination conditions.

### 4.4. Ablation Studies

We perform ablation studies on the Parametric Luminance Curve (PLC), Luminance Response Library (LRL), Structural Response Library (SRL), spectral consistency loss ($\mathcal{L}_{sc}$), and binary cross-entropy loss ($\mathcal{L}_{bce}$) of GraspALL. The study involves six model variants: **Model-1**: PLC is removed, and the encoded features from both luminance and structural modeling are stored in fixed positions of LRL and SRL without luminance distinction. **Model-2**: LRL is removed, and the structural modeling process stores the encoded features in SRL according to the luminance indica-

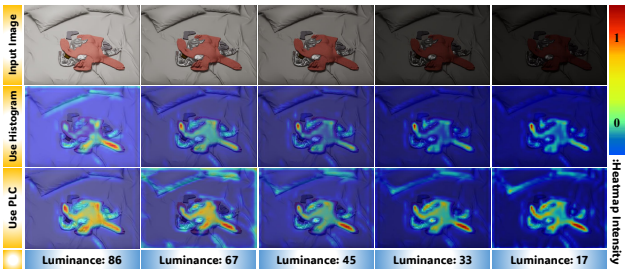


Figure 8. Differences in Grad-CAM at different luminance levels.

tors produced by PLC. **Model-3**: SRL is removed. **Model-4**: $\mathcal{L}_{sc}$ is removed. **Model-5**: $\mathcal{L}_{bce}$ is removed. **Model-6**: Without any modification. As shown in Tab. 3, all variants exhibit varying degrees of performance degradation compared to Model-6, confirming the effectiveness of each component. Model-1 shows the most significant performance drop, indicating that the absence of interpretable luminance estimation prevents explicit guidance for feature fusion between RGB and depth. The removal of SRL and LRL leads to performance decline, demonstrating that the complementary between structural and luminance features restores garment discriminability degraded by illumination changes. Removing $\mathcal{L}_{sc}$ and $\mathcal{L}_{bce}$ results in performance drops. This is because the absence of $\mathcal{L}_{sc}$ eliminates the constraint signal for the PLC learning, while the lack of $\mathcal{L}_{bce}$ deprives the model of supervision for RGB-D fusion.

### 4.5. Complexity Analysis

We compare GraspALL with other models in terms of parameter, frames per second (FPS), and floating point operations per second (FLOPs). As shown in Fig. 7, GraspALL not only achieves a higher mIoU but also delivers faster inference and a smaller model size than comparison methods. This is because we use the structural and luminance response libraries as intermediaries to decouple useful features from complex multimodal fusion, avoiding the intricate semantic alignment and feature interaction computations in multimodal fusion. When processing the input, the model only needs to estimate the illumination level via the PLC, then directly retrieve matching features from the two libraries for grasp point modeling—eliminating the need to repeatedly perform complex cross-modal fusion.

### 4.6. Generalization Analysis of the PLC

To evaluate the performance of PLC for unseen illumination, we employ Grad-CAM [24] to visualize the encoder intermediate features of GraspALL. For comparison, we replace the PLC with a traditional histogram mean method to generate luminance indices for input images. As shown in Fig. 8, GraspALL maintains stable attention across different illumination levels and even exhibits stronger structural focus under low light. In contrast, GraspALL with the histogram method fails to produce more generalized interpretations for diverse inputs due to its non-learnable nature, thus showing significant attention divergence. These results demonstrate that PLC can capture representative luminance patterns from input lighting and form a unified representation for arbitrary illumination, enabling accurate luminance interpretation even under unseen lighting conditions.

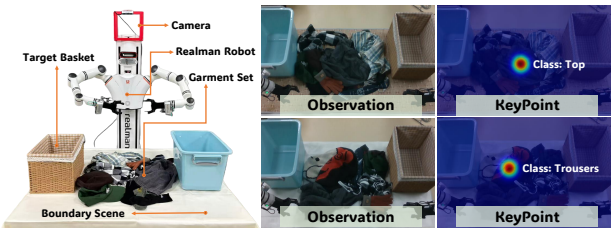


Figure 9. Setup and observation of real-world experiment.

| Luminance | BiFCNet | SAM-M | ReKep | DarkSeg | Ours |
|---|---|---|---|---|---|
| Lu: 0 - 20 | 5 / 15 | 4 / 15 | 6 / 15 | 7 / 15 | **12 / 15** |
| Lu: 20 - 40 | 7 / 15 | 5 / 15 | 7 / 15 | 10 / 15 | **13 / 15** |
| Lu: 40 - 60 | 7 / 15 | 7 / 15 | 9 / 15 | 11 / 15 | **13 / 15** |
| Lu: 60 - 80 | 9 / 15 | 8 / 15 | 10 / 15 | 11 / 15 | **14 / 15** |

Table 4. Real-world success rate under different luminance levels.

### 4.7. Real-World Experiment

To evaluate the real-world performance of GraspALL, we build a dataset of 1013 multi-illumination real-world garment images. We follow the adaptation strategy [32], transferring the weights of all models except ReKep and SAM-M, which were trained on the MIGG simulation dataset, to the real-world dataset, and then deploying them on the Realman robot for evaluation. A RGB-Depth sensor, mounted above Realman's head and facing downward, is used for scene capture. Fig. 9 illustrates the real-world observation process, and Tab. 4 reports the superior grasping success rates achieved by GraspALL. Additional implementation details are provided in the supplementary material.

## 5. Conclusion and Limitation

For garment grasping under varying illumination, we propose GraspALL, an adaptive framework that leverages luminance curve and dual response libraries to achieve illumination-structure feature compensation. By encoding continuous illumination variations into quantifiable references and using these references to guide the feature complementarity between RGB and non-RGB modalities. Experiments in both simulations and real-world deployments demonstrate that GraspALL improves grasping accuracy by 32-44% under diverse illumination conditions. Although our MIGG dataset is simulation-based, it is intentionally designed as a controlled benchmark that allows systematic analysis of illumination variations under physically consistent conditions. To further improve generalization, we plan to expand the real-world data in future work, incorporating more diverse household scenes and garment materials.

## 6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. W2421093) and the International Cooperation Project of Jilin Province (No. 20250205079GH). This work was also supported by the National Science and Technology Council, Taiwan under Grant 114-2221-E-006-114-MY3. This work was supported in part by the National Natural Science Foundation of China (No. 62476110, No. U2341229), and the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2025-RS-2024-00437102).

## References

[1] Stefan Ainetter and Friedrich Fraundorfer. End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 13452–13458, 2021. 2

[2] Dan Busbridge, Jason Ramapuram, Pierre Ablin, Tatiana Likhomanenko, Eeshan Gunesh Dhekane, Xavier Suau, and Russ Webb. How to scale your ema. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 4

[3] Wei Chen, Daniel Lee, Daniel Chappell, and Nicolas Rojas. Learning to grasp clothing structural regions for garment manipulation tasks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023. 2

[4] Wei Chen, Dongmyoung Lee, Digby Chappell, and Nicolas Rojas. Learning to grasp clothing structural regions for garment manipulation tasks. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023. 1

[5] Wei Chen, Dongmyoung Lee, Digby Chappell, and Nicolas Rojas. Learning to grasp clothing structural regions for garment manipulation tasks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4889–4895, 2023. 5

[6] Enric Corona, Guillem Alenyà, Antonio Gabas, and Carme Torras. Active garment recognition and target grasping point detection using deep learning. *Pattern Recognition*, 74:629–641, 2018. 2

[7] Yixing Gao, Hyung Jin Chang, and Yiannis Demiris. User modelling for personalised dressing assistance by humanoid robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1840–1845, 2015. 1

[8] Yixing Gao, Hyung Jin Chang, and Yiannis Demiris. Iterative path optimisation for personalised dressing assistance using vision and force information. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4398–4403, 2016. 1

[9] Daniel F. Gomes, Shan Luo, and Lucas F. Teixeira. Garmnet: Improving global with local perception for robotic laundry folding. *arXiv preprint arXiv:1907.00408*, 2019. 2

[10] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2017. 3

[11] Yi Han, Xiangyong Chen, Yi Zhong, Yanqing Huang, Zhuo Li, Ping Han, Qing Li, and Zhenhui Yuan. Low-illumination road image enhancement by fusing retinex theory and histogram equalization. *Electronics*, 12(4), 2023. 3

[12] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In *Proceedings of The 8th Conference on Robot Learning*, pages 4573–4602. PMLR, 2025. 6

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 6

[14] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015. 2

[15] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8115–8124, 2023. 1, 2, 6

[16] Mingdi Niu, Zhenyu Lu, Lu Chen, Jing Yang, and Chenguang Yang. Vergnet: Visual enhancement guided robotic grasp detection under low-light condition. *IEEE Robotics and Automation Letters*, 2023. 2

[17] Mingjie Pan, Jiyao Zhang, Tianshu Wu, Yinghao Zhao, Wenlong Gao, and Hao Dong. Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17359–17369, 2025. 6

[18] Jin Qian, Tian Weng, Lingni Zhang, Brian Okorn, and David Held. Cloth region segmentation for robust grasp selection. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. 2

[19] Paolo Rabino and Tatiana Tommasi. A modern take on visual relationship reasoning for grasp planning. *arXiv preprint arXiv:2403.12165*, 2024. 2

[20] Arnau Ramisa, Guillem Alenyà, Francesc Moreno-Noguer, and Carme Torras. Learning rgb-d descriptors of garment parts for informed robot grasping. *Engineering Applications of Artificial Intelligence*, 35:246–258, 2014. 2, 6

[21] Rui Ren, Mayank Gurnani, Jordi Sánchez-Riera, Fan Zhang, Ye Tian, Antonio Agudo, Yiannis Demiris, Krystian Mikolajczyk, and Francesc Moreno-Noguer. Grasp-oriented fine-grained cloth segmentation without real supervision. In *International Conference on Machine Vision and Applications (ICMVA)*, 2023. 2

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 3

[23] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d se-

mantic segmentation for indoor scene analysis. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 13525–13531, 2021. 2

[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 8

[25] Hassan Shehawy, Paolo Rocco, and Andrea Maria Zanchettin. Estimating a garment grasping point for robot. In *2021 20th International Conference on Advanced Robotics (ICAR)*, pages 707–714, 2021. 5

[26] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. Imagdressing-v1: Customizable virtual dressing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):6795–6804, 2025. 2

[27] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760, 2012. 1

[28] Changki Sung, Wanhee Kim, Jungho An, Wooju Lee, Hyungtae Lim, and Hyun Myung. Contextrast: Contextual contrastive learning for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3732–3742, 2024. 6

[29] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference (BMVC)*, 2018. Oral. 3

[30] Ruihai Wu, Haoran Lu, Yiyan Wang, Yubo Wang, and Hao Dong. Unigarmentmanip: A unified framework for category-level garment manipulation via dense visual correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16340–16350, 2024. 1

[31] Ruihai Wu, Ziyu Zhu, Yuran Wang, Yue Chen, Jiarui Wang, and Hao Dong. Garmentpile: Point-level visual affordance guided retrieval and adaptation for cluttered garments manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6950–6959, 2025. 1, 5, 7

[32] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[33] Fan Zhang and Yiannis Demiris. Learning grasping points for garment manipulation in robot-assisted dressing. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 9114–9120, 2020. 1

[34] Feng Zhang, Xinran Liu, Changxin Gao, and Nong Sang. Color and luminance separated enhancement for low-light images with brightness guidance. *Sensors*, 24(9), 2024. 2, 3

[35] Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26974–26983, 2024. 1, 2, 4, 6

[36] Kun Zhang, Qinghua Li, Kaiyue Liu, Mengyao Zhang, Xuyang Wang, and Chao Feng. Spanet—sparse convolutional pyramid attention network for grasping detection in low-light conditions. In *2023 China Automation Congress (CAC)*, pages 5674–5679. IEEE, 2023. 2

[37] Kun Zhang, Qinghua Li, Kaiyue Liu, Mengyao Zhang, Zhaoxin Zhu, and Chao Feng. Am-gpd: Manipulator grasping pose detector based on attention mechanism in dark light scene. In *2023 China Automation Congress (CAC)*, pages 2281–2286. IEEE, 2023. 2

[38] Pingping Zhang, Wenhui Wu, Jian Weng, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5891–5900. IEEE, 2022. 5

[39] Guoqiang Zhao, Junjie Huang, Xiaoyun Yan, Zhaojing Wang, Junwei Tang, Yangjun Ou, Xinrong Hu, and Tao Peng. Open-vocabulary rgb-thermal semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1, 2

[40] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5906–5916, 2023. 1

[41] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: Denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8082–8093, 2023.

[42] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[43] Haifeng Zhong, Fan Tang, Hyung Jin Chang, Xingyu Zhu, and Yixing Gao. Darkseg: Infrared-driven semantic segmentation for garment grasping detection in low-light conditions. In *IROS*, 2025. 2, 5, 6

[44] Haifeng Zhong, Fan Tang, Zhuo Chen, Hyung Jin Chang, and Yixing Gao. Amdanet: Attention-driven multi-perspective discrepancy alignment for rgb-infrared image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10645–10655, 2025. 1, 2, 4, 6

[45] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *Advances in Neural Information Processing Systems*, pages 30599–30611. Curran Associates, Inc., 2022. 4

[46] Xingyu Zhu, Xin Wang, Jonathan Freer, Hyung Jin Chang, and Yixing Gao. Clothes grasping and unfolding based on rgb-d semantic segmentation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 9471–9477, 2023. 6

[47] Xingyu Zhu, Yan Wu, Zhiwen Tu, Haifeng Zhong, and Yixing Gao. Generalizable category-level topological structure learning for clothing recognition in robotic grasping. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 21259–21266. IEEE, 2025. 2

[48] Xingyu Zhu, Haifeng Zhong, Yan Wu, Shan Luo, and Yixing Gao. Guiding robotic cloth grasping in darkness: Infrared semantic segmentation and grasping position selection. *IEEE Robotics and Automation Letters*, 2025. 2

[49] Lipeng Zhuang, Shiyu Fan, Yingdong Ru, Florent P. Audonnet, Paul Henderson, and Gerardo Aragon-Camarasa. Flat'n'fold: A diverse multi-modal dataset for garment perception and manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7937–7944, 2025. 2, 6

[50] Guoqiang Zuo, Jing Tong, Hao Liu, Wei Chen, and Jianwei Li. Graph-based visual manipulation relationship reasoning network for robotic grasping. *Frontiers in Neurorobotics*, 15: 719731, 2021. 2