

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

Supplementary Material

This supplementary material presents additional experimental results that are omitted from the main paper due to the space limit. In this supplementary material, we provide:

- **1.** Explanation of using RGB in extremely low-light conditions (in Sec. 1);
- **2.** Explanation of the PLC (parametric luminance curve) generation process (in Sec. 2);
- **3.** Analysis of the parameters N and R in PLC (in Sec. 3);
- **4.** Real-world experiment deployment details and real-world images display (in Sec. 4);
- **5.** Our solution strategy for dealing with low-quality depth maps (in Sec. 5);
- **6.** Performance analysis for the proposed grasping strategy (i.e., depth-optimal search strategy) (in Sec. 6);
- **7.** Analysis for the parameter α of the EMA (in Sec. 7);
- **8.** Validation of GraspALL’s generalization for other deformable objects (in Sec. 8);
- **9.** Statistical analysis of grasping performance (in Sec. 9);
- **10.** Robustness analysis of our GraspALL (in Sec. 10).

1. Explanation of using RGB in extremely low-light condition

In extremely low-light conditions, our GraspALL still uses RGB images to provide faint but crucial visual information, rather than relying solely on depth map information. Depth mainly captures geometry but lacks discriminative cues such as color, material (Fig. 1 upper left). Since garments are highly deformable, different categories exhibit similar shapes, so using depth alone leads to class confusion. In contrast, low-light RGB, though weak, still retains useful semantic cues that complement depth feature by semantically consistent fusion. In Fig. 1, training with depth alone leads to severe misclassification.

2. Explanation of the PLC Generation Process

In GraspALL, we propose a parametric luminance curve (PLC), which learns a set of curves capable of representing arbitrary illumination from multiple luminance inputs, thereby accurately estimating the luminance of the input.

The PLC forces curve IDs to indicate more accurate positions by calculating the feature consistency between the features generated by the encoder and those in the response library guided by curve IDs. This further guides the model to adjust curve parameters to generate more accurate IDs. However, considering that curve ID generation is discrete and non-differentiable, to enable end-to-end differentiable

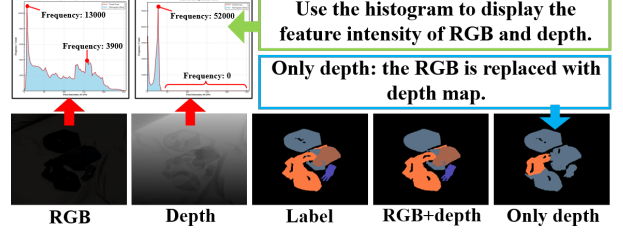


Figure 1. Explanation of using RGB in low-light conditions.

Method	Luminance<20	Luminance<40	#Params
$N = 6, R = 256$	mIoU:80.8%	mIoU:81.5%	34.53
$N = 12, R = 256$	mIoU: 82.2%	mIoU: 83.1%	47.83
$N = 18, R = 256$	mIoU:82.5%	mIoU:83.7%	61.13
$R = 128, N = 12$	mIoU:80.3%	mIoU:80.9%	47.83
$R = 256, N = 12$	mIoU: 82.2%	mIoU: 83.1%	47.83
$R = 512, N = 12$	mIoU:81.4%	mIoU:82.0%	47.83

Table 1. The effect of different N and R on model performance. learning for the PLC and allow gradients to flow to PLC parameters, we adopt a differentiable processing mechanism in practice. First, we calculate the negative distance for each curve and perform temperature-scaled softmax:

$$w_n = \frac{\exp(-\|\mathbf{H} - \mathbf{C}(\mathbf{P}_n)\|/\tau)}{\sum_m \exp(-\|\mathbf{H} - \mathbf{C}(\mathbf{P}_m)\|/\tau)}, \quad (1)$$

where τ is the temperature coefficient, which controls the smoothness of the soft-hard distribution.

Then, we use the softmax weight w_n to perform a weighted sum of the corresponding features in the luminance response library \mathbf{M}_L :

$$\mathbf{M}_L = \sum_{n=1}^N w_n \mathbf{M}_L^n. \quad (2)$$

Finally, we calculate the spectral consistency loss (\mathcal{L}_{sc}):

$$\mathcal{L}_{sc} = \|\mathbf{F}_{en}^n - \mathbf{M}_L^n\|_1, \quad (3)$$

where $\|\cdot\|_1$ represents the L1 loss. Through the above differentiable processing, gradients can be backpropagated to \mathbf{F}_n^{en} and parameters \mathbf{P}_n in the PLC via \mathbf{M}_L .

Fig. 2 visually presents the set of curves generated by the parametric luminance curve (PLC) after training. As can be seen from Fig. 2, different curves have distinct focuses in representing luminance, enabling a comprehensive evaluation of the input luminance intensity.

3. Analysis of the Parameters N and R in PLC

The PLC setup includes two key parameters: one is the number of luminance curves, denoted as N , and the other

Luminance	BiFCNet [12]	SAM-M [4]	ReKep [3]	DarkSeg [10]	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 2. Real-world grasping success rate of different methods.

Luminance	SegMiF [5]	MRFS [9]	AMDA [11]	Ours
Lu: 0 - 20	51.2%	52.9%	55.1%	70.5%
Lu: 20 - 40	55.8%	58.3%	59.5%	71.3%
Lu: 40 - 60	59.4%	62.7%	63.8%	72.4%
Lu: 60 - 80	64.3%	65.6%	66.3%	74.3%

Table 3. Accuracy rate of mask generation by different methods.

is the number of nodes per curve, denoted as R . For N , we set $N = 12$, considering the symmetric illumination changes from dark to bright and then bright to dark over 24 hours a day. For R , we set $R = 256$ to learn sufficient yet non-redundant luminance features from the input.

To verify the rationality of the settings for N and R , we compared the impacts of different N and R values on model performance. The experimental results are presented in Tab. 1. As shown in Tab. 1, when $N < 12$, the model accuracy degrades. This is because a smaller number of curves leads to vague illumination estimation by the PLC, failing to provide accurate guidance for the subsequent interactive enhancement of luminance and structural features. When $N > 12$, although the model performance improves slightly, the increased number of curves requires more storage space to expand the luminance and structural response libraries, resulting in a significant increase in model complexity. Thus, $N = 12$ achieves a better balance between accuracy and model complexity. Regarding the setting of R , Tab. 1 indicates that model performance declines when $R < 256$ or $R > 256$. This is because an excessive R causes the curves to overfit to a specific luminance pattern, while an insufficient R makes it difficult for the curves to form a unified representation for diverse inputs. Therefore, $R = 256$ enables the sufficient learning of representative patterns from different luminance inputs.

4. Real-World Experiment Deployment

To verify the performance of GraspALL in real-world scenarios, we collected a dataset named RealData containing 1013 real-world captured images, as shown in Fig. 3. To enhance the sample diversity across different illumination levels, we systematically acquired garment images under varying illumination conditions (illumination range: 10–70) in a controlled environment by adjusting the curtain opening. The garment categories in RealData are consistent with those in the synthetic dataset MIGG.

To achieve domain adaptation from the synthetic dataset MIGG to the real-image dataset RealData, we adopted the Fourier Domain Adaptation method proposed by Yang et al [8]. As illustrated in Fig. 4, we performed Fast Fourier

	GraspALL (mGSR)		DarkSeg (mGSR)	
Luminance	with ST	without ST	with ST	without ST
0 - 20	80.0%	66.7%	46.7%	33.3%
Luminance	GraspALL (mIoU)		DarkSeg (mIoU)	
0 - 20	70.5%	61.4%	50.7%	36.9%

Table 4. Analysis for our transfer strategy (sim-to-real).

Transform (FFT) on synthetic and real images respectively, replaced the corresponding regions of the source domain (synthetic) images with parts of the target domain (real) images, and then reconstructed the style-converted synthetic images through Inverse Fast Fourier Transform (iFFT). Subsequently, these synthetic-real hybrid images and their original labels were used to train the model, reducing the differences in low-level features (e.g., texture, shape) between the two domains. Meanwhile, to further improve the model’s generalization ability on real images, we introduced an entropy minimization-based regularization term and a pseudo-label self-supervised training strategy, thereby achieving efficient cross-domain adaptation.

The real-world experimental results are presented in Tab. 2. As can be seen from Tab. 2, even in real scenarios, our method can achieve more accurate grasping precision under different illumination conditions, demonstrating the reliability and practicality of our method. In addition, we compared the semantic mask generation accuracy of different multimodal methods on the Real-world dataset. As indicated in Tab. 3, compared with other methods, our method not only achieves higher mask generation accuracy but also exhibits greater stability with smaller fluctuations when facing varying illumination conditions.

Furthermore, to verify the effectiveness of our proposed domain transfer strategy, we conduct an ablation analysis, with the experimental results presented in Tab. 4. As can be seen from Tab. 4, if our transfer strategy is removed (i.e., training relies solely on real-world data), the performance of both GraspALL and DarkSeg degrades significantly. This demonstrates that our transfer strategy can effectively transfer GraspALL and all baseline models from their MIGG-trained weights to the real-world environment, and leverage the stable feature knowledge from the simulation environment to enhance the model’s generalization ability and robustness in the real world.

5. Strategy for Low-Quality Depth Maps

Our GraspALL relies heavily on the structural information of depth maps. However, depth maps are often affected by factors such as sensor noise, leading to unstable depth values or local missing regions. This makes it difficult to provide accurate geometric structural features for GraspALL. To address this, we designed a two-stage depth enhancement strategy (TSD-En). Unlike previous methods [2, 7]

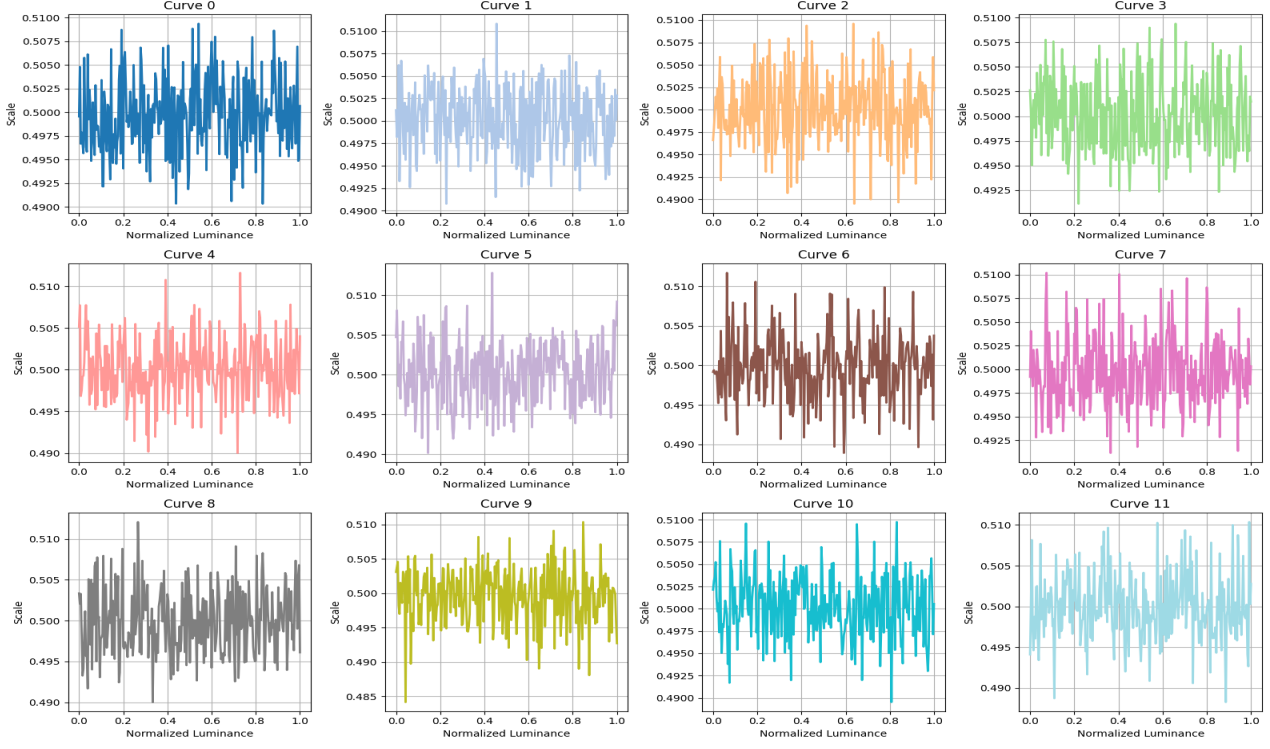


Figure 2. Visualize the different luminance curves in the trained Parametric Luminance Curves.

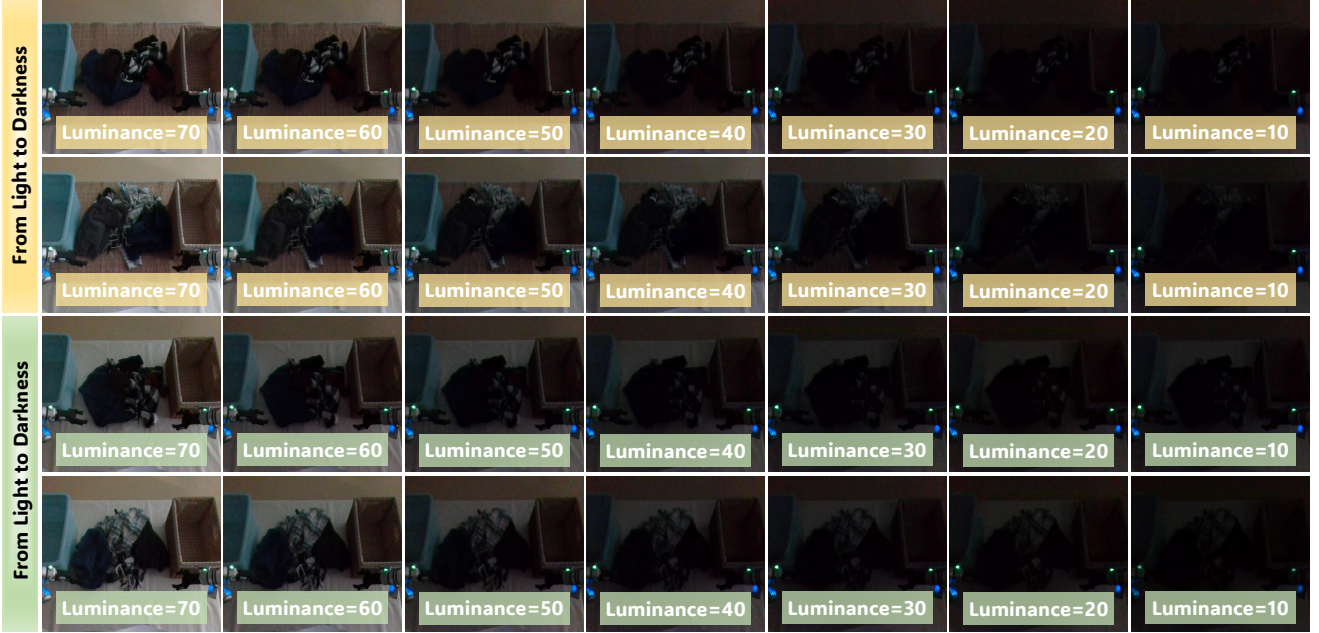


Figure 3. Garment images under different illumination conditions collected in real-world scenarios.

that only perform noise smoothing, our strategy not only emphasizes noise smoothing but also focuses on resolving the problem of local depth value holes caused by noise, thereby improving the structural integrity of depth maps.

First, for the input depth map, we first adopt bilateral filtering to smooth random noise while preserving edge structures. For each pixel p in the depth map, the filtered depth

value $D_b(p)$ is defined as:

$$D_b(p) = \frac{1}{W_p} \sum_q G_\sigma^s(p, q) \cdot G_\sigma^i(p, q) \cdot q, \quad (4)$$

where q denotes a point within the neighborhood window $\mathcal{N}(p)$ of pixel p : $q \in \mathcal{N}(p)$. $\frac{1}{W_p}$ is a normalization coef-

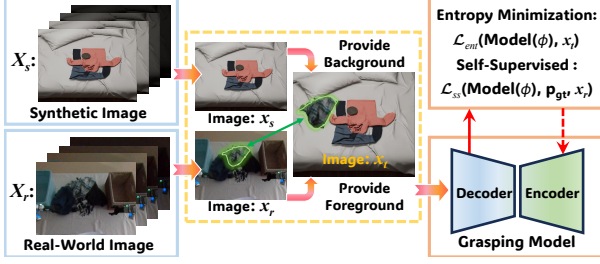


Figure 4. We use the method in [8] to transfer the model trained on the synthetic dataset to the real-world dataset. $\text{Model}(\phi)$ denotes the model trained on synthetic data. \mathbf{p}_{gt} represents the pseudo-labels generated by the model. The domain adaptation process is as follows: first, train the model using synthetic data; then, train the model using both synthetic and real data; finally, train the model using real data.

efficient: $W_p = \sum_q \mathbf{G}_\sigma^s(p, q) \cdot \mathbf{G}_\sigma^i(p, q)$. $\mathbf{G}_\sigma^s(\cdot)$ and $\mathbf{G}_\sigma^i(\cdot)$ represent the spatial and the intensity Gaussian kernel:

$$\begin{aligned} \mathbf{G}_\sigma^s(p, q) &= \exp\left(-\frac{\|p-q\|^2}{2\sigma_s^2}\right), \\ \mathbf{G}_\sigma^i(p, q) &= \exp\left(-\frac{|p-q|^2}{2\sigma_s^2}\right). \end{aligned} \quad (5)$$

Specifically, $\mathbf{G}_\sigma^s(\cdot)$ is used to weight pixels that are closer within the neighborhood and regulate the neighborhood range, and $\mathbf{G}_\sigma^i(\cdot)$ is designed to avoid smoothing across edges. The aforementioned process can efficiently suppress random noise while maintaining critical geometric features such as garment wrinkles and edges.

After filtering, some regions may develop holes due to noise correction or sensor defects. To restore the complete depth structure, we need to further perform hole interpolation completion. For the missing pixel $p \in \Omega_{\text{hole}}$, its completed depth value $\mathbf{D}_c(p)$ is expressed as:

$$\mathbf{D}_c(p) = \frac{\sum_{q \in \mathcal{N}_v(p)} |w(p, q) \cdot \mathbf{D}_b(q)|}{\sum_{q \in \mathcal{N}_v(p)} |w(p, q)|}, \quad (6)$$

where $\mathcal{N}_v(p)$ represents the valid pixel set surrounding pixel p . The weight $w(p, q)$ is defined as follow:

$$w(p, q) = \exp\left(-\frac{|\nabla|\mathbf{D}_b(q)|^2|}{2\sigma_s^2}\right) \cdot \exp\left(-\frac{\|p-q\|^2}{2\sigma_s^2}\right), \quad (7)$$

where, $|\nabla|\mathbf{D}_b(q)|^2|$ denotes the spatial gradient of the depth map at pixel q , which is used to describe the local intensity of depth changes. The introduction of the gradient term can effectively prevent incorrect interpolation across depth edges, making the completion process more consistent with the real geometric structure of the object.

The weight $w(p, q)$ to impose dual constraints on spatial distance and gradient intensity, ensuring the interpolation result is both continuous and non-crossing of edges, thereby preserving the object's geometric contour.

Method	Luminance	mIoU	mGSR
GraspALL	0 - 20	82.2%	82.5%
GraspALL + [7]	0 - 20	82.9%	83.3%
GraspALL + TSD-En	0 - 20	84.5%	86.6%
GraspALL	20 - 40	83.1%	84.1%
GraspALL + [7]	20 - 40	83.6%	85.0%
GraspALL + TSD-En	20 - 40	85.2%	87.5%

Table 5. The improvement brought by the TSD-En to GraspALL.

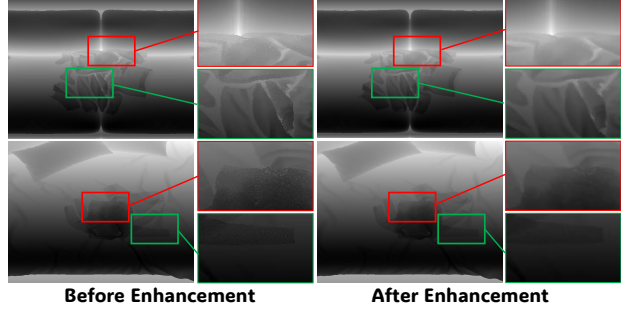


Figure 5. Visualization results with and without our depth map enhancement strategy (TSD-En).

To verify the our strategy, we conduct corresponding ablation analysis on GraspALL. The experimental results are shown in Tab. 5. As can be seen from Tab. 5, when our TSN-En strategy is adopted, both the semantic mask generation accuracy and grasp success rate of the model are improved to a certain extent under different illumination conditions. In addition, compared with methods that only focus on smoothing depth map noise, our strategy also achieves better performance, increasing the grasp accuracy and semantic mask generation accuracy by 5 and 3 percentage points, respectively. This indicates that for the processing of low-quality depth maps, it is not only necessary to smooth the abnormalities caused by noise but also to further fill the depth gaps that may be generated after noise smoothing.

To intuitively demonstrate the effectiveness of our depth map enhancement strategy, we show the impact of using and not using TSD-En on depth maps, as illustrated in Fig. 5. As can be seen from Fig. 5, when depth maps suffer from noise or depth value holes due to sensor noise, our method can effectively suppress noise and fill the holes. The above results prove that when the depth sensor generates low-quality depth maps, the model can further correct the disturbed depth information through the proposed depth map enhancement strategy, thereby providing more accurate structural information for the grasping model.

6. Analysis of the Proposed Grasping Strategy

To obtain grasp points based on garment semantic masks, existing methods usually determine grasp points based on the geometric center of the garment or positions adjacent

Method	Luminance<20	Luminance<40	Average
GraspALL + [6]	17/25 (68%)	18/25 (72%)	70%
GraspALL + [1]	17/25 (68%)	19/25 (76%)	72%
GraspALL + [10]	19/25 (76%)	20/25 (80%)	78%
GraspALL + Ours	21/25 (84%)	22/25 (88%)	86%

Table 6. Grasping performance of different grasping strategies.

Luminance	DarkSeg	BiFCNet	GraspALL
0 – 20	59.2%	46.7%	70.5%
20 – 40	61.6%	54.5%	74.4%

Table 7. Analysis for our Grasping Strategy (in Simulation)

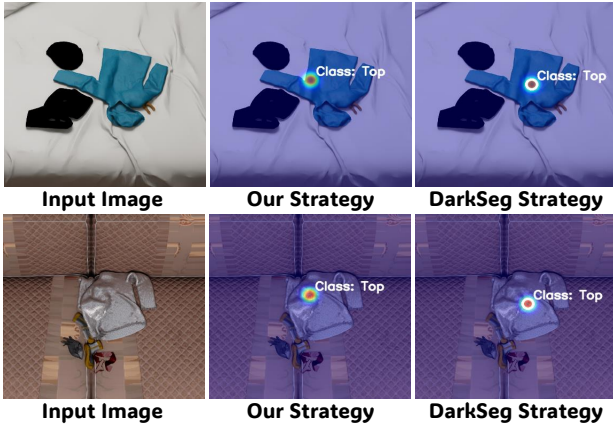


Figure 6. Comparison between our grasping strategy and main-stream strategies that take the geometric center as the grasp point.

to the center point, aiming to avoid grasping the edge positions of the garment which may lead to dragging problems during the grasping process. However, the deformability of garments means that the center position is not necessarily a point with obvious wrinkles and easy to grasp, which can result in slipping during grasping and moving. In response to this, we propose a depth optimal search strategy to find positions that are as easy to stably grasp as possible on the premise of avoiding garment dragging. To verify the effectiveness of the proposed strategy, we compare its performance with several grasp strategies that take the garment center as the grasp point. The results are shown in Tab. 6. It can be seen from Tab. 6 that using our strategy can improve the grasping accuracy by 7–11%, indicating that our strategy enables the model to achieve better grasping performance. The above results verify the correctness of our grasp strategy: when determining grasp points based on garment semantic mask regions, we should first identify candidate grasp points with obvious wrinkles from a global perspective, and then select the optimal grasp point from the candidate points according to the center point of the garment.

To more intuitively verify the performance of our method, we compare the performance differences between the proposed method and the method that takes the geomet-

Metrics	$\alpha=0.01$	$\alpha=0.03$	$\alpha=0.05$	$\alpha=0.07$	$\alpha=0.09$
mIoU	78.2%	80.9%	82.8%	81.1%	79.6%
mGSR	78.3%	81.6%	84.2%	82.5%	81.6%

Table 8. The effect of different α on model (Luminance:0-40).

Method	mIoU	Method	mGSR
SegMiF [5]	64.6%	BiFCNet [12]	52.4%
MRFS [9]	67.5%	SAM-M [4]	59.2%
AMDA [11]	68.9%	DarkSeg [10]	63.3%
GraspALL	86.7%	GraspALL	88.3%

Table 9. Performance of different methods on other objects.

ric center as the grasp point. As shown in Fig. 6, although DarkSeg—with the central point as its main grasping region—can identify the geometric center of the garment, the wrinkles at the geometric center are very weak. In contrast, the grasp points determined by our method can not only be as close as possible to the geometric center but also better locate regions with prominent wrinkles.

Moreover, in Tab. 7, to isolate the contribution of model perception, we fix the grasping strategy to the geometric center of the perception mask for the garment, without using depth or top-k heuristics. GraspALL still outperforms baselines, indicating that baselines are less robust to illumination changes and thus produce more class errors, while GraspALL can better adapt to illumination changes.

7. Analysis for the Parameter α of the EMA

In GraspALL, we adopt the EMA (Exponential Moving Average) strategy to update the generated luminance and structural compensation features to the corresponding luminance and structural response libraries. For the EMA strategy, we set the update momentum $\alpha = 0.05$. To verify the rationality of this parameter setting, we analyze the impact of different α values on model performance.

The experimental results are shown in Tab. 8. As can be seen from Tab. 8, when α is small (e.g., 0.01 and 0.03), the model absorbs newly generated compensation features too slowly with low update efficiency, making it difficult for the features in the two response libraries to quickly adapt to diverse illumination and structural changes. When α is large (e.g., 0.07 and 0.09), the model is overly sensitive to the update of new features and prone to interference from noise or abnormal sample features, which reduces the stability of features in the response libraries and leads to overfitting. Therefore, when $\alpha = 0.05$, the model can achieve an optimal balance between the update efficiency of the response libraries and feature stability. It can not only timely integrate effective compensation features to adapt to scene changes but also avoid noise interference to ensure the reliability of features in the libraries.



Figure 7. Non-garment images (towels and shopping bags) captured under different illumination conditions in real-world scenarios.

Luminance	BiFCNet [12]	SAM-M [4]	DarkSeg [10]	Ours
Lu: 00 - 40	37.7% (4.8-7.5)	35.5% (3.4-5.5)	48.8% (5.3-8.7)	80.0% (2.9-3.6)
Lu: 40 - 80	44.4% (3.8-4.9)	40.0% (4.5-6.4)	53.3% (2.5-4.4)	84.4% (1.9-3.2)

Table 10. Statistical analysis of different methods. The red font indicates the accuracy fluctuation across three experimental rounds.

8. Validation of GraspALL’s Generalization

To verify the generalization performance of GraspALL on other deformable objects, we collected images of shopping bags and towels (commonly seen in household scenarios) under different illumination conditions for generalization experiment validation. Fig. 7 shows some of the shopping bag and towel images, and their collection process is consistent with that of RealData in Sec. 4. The experimental results are presented in Tab. 9. As can be seen from Tab. 9, compared with other methods, even when facing non-garment deformable objects such as shopping bags and towels with the interference of illumination variations, our GraspALL can still maintain high semantic mask generation accuracy and grasp success rate. The above experimental results indicate that our method not only performs excellently in garment grasping tasks but also possesses strong cross-deformable object category generalization ability, which can effectively adapt to different common deformable objects and complex illumination environments in household scenarios.

9. Statistical Analysis

During the grasping tests, our setup involves 15 grasping attempts per test, with a successful grasp defined as picking up the garment of the target category and placing it into the corresponding category-specific basket. Considering the experimental randomness, we conduct multiple repeated experiments to ensure statistical significance of the results. Specifically, we perform three consecutive rounds of grasping tests for each method in the real world, with 15 attempts per round, and finally compared the average grasp success rate and standard fluctuations across the three rounds. As shown in Tab. 10, our GraspALL achieves the highest average grasp success rate among all methods, and the fluctua-

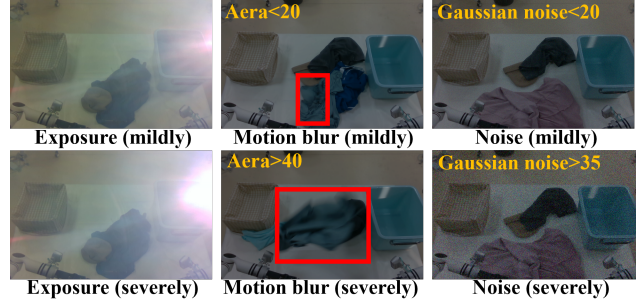


Figure 8. The test samples we collected for different types of image degradation (exposure, noise, and motion blur).

No Corruption		Mildly		Severely	
mGSR	mIoU	mGSR	mIoU	mGSR	mIoU
80%	73.4%	Exposure	70%	69.7%	70%
		Motion blur	60%	67.2%	40%
		Noise	70%	70.8%	50%
				64.3%	53.8%
				49.5%	

Table 11. Validation of GraspALL Robustness.

tion in success rate is smaller than that of the comparative methods. The above results indicate that our method has obvious advantages in terms of result stability, and further proves that the experimental results are not affected by random factors and have reliable statistical significance.

10. Robustness analysis of our GraspALL

GraspALL exhibits a certain degree of robustness to real-world exposure, noise and motion blur. First, since the PLC Curve ID is derived from 256 RGB histogram points, even under mild corruption that degrades some details, sufficient features remain for stable matching. In addition, our transfer strategy leverages stable priors learned in simulation to partially compensate for corrupted RGB features. For validation, in Fig. 8, we collect mildly and severely corrupted samples (10 images each) for exposure, noise, and motion blur. As shown in Tab. 11, compared to no corruption, GraspALL’s performance changes slightly under mild corruption, but drops noticeably under severe corruption. We attribute this potential drawback to severe corruption disrupting too many effective RGB features, making PLC matching and the cross-attention unreliable. Overall, our Gras-

pALL has certain robustness in dealing with image quality degradation problems such as motion blur.

References

- [1] 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
- [2] Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, and Yao Zhao. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9229–9238, 2022. 2
- [3] 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. 2
- [4] 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. 2, 5, 6
- [5] 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. 2, 5
- [6] 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
- [7] Silvan Weder, Johannes L. Schönberger, Marc Pollefeys, and Martin R. Oswald. Neurfusion: Online depth fusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3162–3172, 2021. 2, 4
- [8] 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. 2, 4
- [9] 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. 2, 5
- [10] 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
- [11] 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. 2, 5
- [12] 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. 2, 5, 6