DarkSeg: Infrared-Driven Semantic Segmentation for Garment Grasping Detection in Low-Light Conditions

Haifeng Zhong[1,2], Fan Tang[3], Hyung Jin Chang[4], Xingyu Zhu[1,2], Yixing Gao[1,2,*]

Abstract*— Garment grasping in low-light environments remains a critical yet underexplored challenge for domestic service robots. Insufficient illumination leads to sparse visual features, causing ambiguous similarities across garment categories and impairing reliable recognition. While conventional approaches employ infrared–visible multimodal fusion to mitigate this issue, their heavy computational overhead limits real-time deployment on resource-constrained robotic platforms. To overcome these limitations, we propose DarkSeg, a student–teacher model designed for low-light garment detection. Unlike multimodal fusion methods, DarkSeg leverages an indirect feature alignment mechanism, where the student model learns illumination-invariant structural representations from infrared features provided by the teacher model. This effectively compensates for structural deficiencies in low-light imagery while maintaining computational efficiency. To further validate DarkSeg in practical robotic applications, we introduce a depth-perceptive grasping strategy and construct DarkClothes, a low-light multimodal garment dataset. Experiments on a Baxter robot demonstrate that DarkSeg improves the garment grasping success rate by 22%, while reducing parameters by 99.08M compared to traditional methods, highlighting its effectiveness and feasibility for real-world deployment. The code and dataset are available at *https://github.com/Zhonghaifeng6/Darkseg

I. INTRODUCTION

Garment grasping is a fundamental capability for domestic service robots in household cleaning [1], garment storage [2], and dressing assistance [3], [4]. As the core of physical interaction, effective grasping requires comprehensive perception to answer two critical questions: "What to grasp" and "Where to grasp." Current research [5]–[7] primarily adopts semantic segmentation, leveraging its robust scene parsing capability as the main perception mechanism for garment recognition.

The central challenge of applying semantic segmentation in household robots lies in achieving reliable perception under low-light conditions [8]. Low-light imagery exhibits sparse structural features, such as blurred edges and loss of fine details, leading to ambiguous garment similarities and frequent misclassification. A conventional solution employs infrared–visible multimodal fusion [9], which enhances structural representations by incorporating infrared features into low-light images. However, such preprocessing is computationally expensive, making it impractical for service robots with real-time constraints and limited resources. Additionally, annotating low-light datasets is challenging, as manual labeling under poor visibility is error-prone.

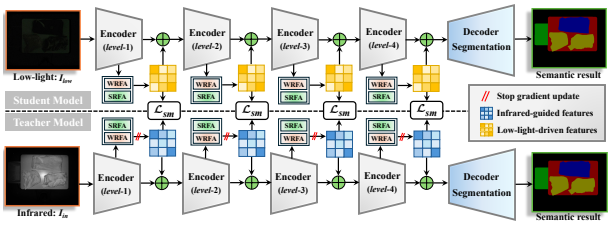

Fig. 1. The overview of proposed DarkSeg. Due to the influence of low-light conditions, the input image exhibits highly entangled and blurred features between the shirts (yellow) and the pants (blue). Nevertheless, DarkSeg can leverage the clear structural representations in infrared features to enhance the distinction of these blurred features.

To address these challenges, we propose DarkSeg—an infrared structure-driven, lightweight semantic segmentation model tailored for embedded platforms in low-light environments. DarkSeg employs a teacher–student paradigm: infrared images are input to the teacher model, while low-light images are input to the student model. Structural features extracted by the teacher act as constraints guiding the student's feature modeling. This enables the student model to learn structural representations that compensate for incomplete details in low-light conditions. Unlike multimodal methods, DarkSeg relies on teacher guidance only during training, avoiding image fusion at inference and ensuring efficiency.

Furthermore, we propose a garment grasping strategy that integrates geometric constraint completion with depth-aware mechanisms. Based on DarkSeg's segmentation outputs, this strategy refines candidate grasp regions through mask optimization and determines precise grasp points via depth-guided search, substantially improving grasp stability and robustness. Finally, we introduce the Darkclothes dataset, a large-scale benchmark for low-light garment grasping with comprehensive ground truth. Each sample includes three aligned modalities: a low-light image, its corresponding well-lit version, and an infrared counterpart. This design enables (1) accurate annotation using well-lit references and (2) effective exploitation of infrared structures as supplementary training data to enhance low-light perception.

II. METHOD

A. The Overview of Proposed DarkSeg

The details of the proposed DarkSeg are shown in Fig. 1. DarkSeg consists of a teacher model processing infrared images and a student model handling low-light images. Given a paired input of low-light and infrared images $\{I_{low}$,

[1] School of Artificial Intelligence, Jilin University, China.
[2] Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, Ministry of Education, China.
[3] Institute of Computing Technology, Chinese Academy of Sciences, China.
[4] School of Computer Science, University of Birmingham, U.K.
* Corresponding author. Email: gaoyixing@jlu.edu.cn

$I_{in}\}$, we employ a backbone network [10] to extract multi-scale features $\{$low-light: $\mathbf{F}_{low}$, infrared: $\mathbf{F}_{in}\}$ from both modalities. We leverage the teacher's explicit structural representations derived from infrared features $\mathbf{F}_{in}$ to drive the student's structural modeling of low-light images through two novel components: the Short-Range Structural Perception (SRSP) module for localized pattern enhancement and the Wide-Range Structural Perception (WRSP) module for global contextual modeling.

1) Short-Range Structural Perception Module: The short-range structural perception module (SRSP) operates on the principle that convolutional neural networks inherently extract short-range features [11], enabling the model to perceive prominent high-frequency structural information from localized receptive fields.

Given the output of the encoder $\mathcal{F}_{mid}$, we first split it into four feature subsets along the channel dimension direction. The feature splitting explicitly decouples the initial features into multiple non-overlapping task components. Each task component can independently calculate the structural strength from different perspectives:

$$\mathcal{F}_{mid} = \left(\Delta\mathcal{F}_{1}^{1/4c}, \Delta\mathcal{F}_{1/4c}^{2/4c}, \Delta\mathcal{F}_{2/4c}^{3/4c}, \Delta\mathcal{F}_{3/4c}^{c}\right), \tag{1}$$

where $\Delta\mathcal{F}_{1}^{1/4c}$, $\Delta\mathcal{F}_{1/4c}^{2/4c}$, $\Delta\mathcal{F}_{2/4c}^{3/4c}$ and $\Delta\mathcal{F}_{3/4c}^{c}$ represent the divided feature subsets, and each subset has the same number of channels. Afterward, dilated convolutions are used to compute high-frequency structure information within multi-scale regions to structural intensity values:

$$\nabla\mathcal{F}_{m}^{n} = f_N\left(\Delta\mathcal{F}_{m}^{n} * \mathcal{D}\left(s=r\right)\right) \otimes \Delta\mathcal{F}_{m}^{n}, \tag{2}$$

where f_N is the sigmoid function. $\Delta\mathcal{F}_{m}^{n}$ represents the feature subset. $*$ represents the convolution process. $\mathcal{D}\left(\cdot\right)$ is the dilated convolution, and the dilated rate r=1, 2, 3, 5. After obtaining the strength weights of the structure, it needs to be returned to the initial feature to complete the enhancement of the structure information:

$$\nabla\mathcal{F}_{mid} = \nabla\mathcal{F}_{1}^{1/4c} \cup \nabla\mathcal{F}_{1/4c}^{2/4c} \cup \nabla\mathcal{F}_{2/4c}^{3/4c} \cup \nabla\mathcal{F}_{3/4c}^{c}, \tag{3}$$

where $\cup$ is concatenation. We use $\cup$ to aggregate each subset feature: $\mathcal{F}_{1}^{1/4c}, \mathcal{F}_{1/4c}^{2/4c}, \mathcal{F}_{2/4c}^{3/4c}, \mathcal{F}_{3/4c}^{c}$.

2) Wide-Range Structural Perception Module: The wide-range structural perception module (WRSP) leverages the Transformer's superior long-range modeling ability [12], enabling global perception of salient high-frequency structural features. However, the Transformer's quadratic computational complexity [13] makes it unsuitable for modeling intermediate features with high resolution and multiple channels. To address this, we design a multi-layer structural modeling Transformer (MSMT). A single MSMT provides limited feature extraction, while stacking multiple MSMTs significantly increases computational cost. To overcome this trade-off, we introduce an Invertible Transformer Unit (ITU), built upon invertible neural networks (INN) [14]. By exploiting INN's property of lossless information preservation [15], ITU enhances MSMT's structural feature modeling while improving memory efficiency. The $\mathcal{F}_{mid}$ is first evenly split along the channel dimension into two partitions: $\mathcal{F}\left[1:c\right]$ and $\mathcal{F}\left[c+1:C\right]$, and feeds them into the ITU for processing:

$$\begin{aligned}
\mathcal{T}\left[c+1:C\right] &= \mathcal{F}\left[1:c\right] + \mathcal{I}_1\left(\mathcal{F}\left[c+1:C\right]\right), \\
\mathcal{T}\left[1:c\right] &= \mathcal{F}\left[c+1:C\right] + \mathcal{I}_2\left(\mathcal{T}\left[c+1:C\right]\right), \\
\Delta\mathcal{F}_{mid} &= \mathcal{CAT}\left\{\mathcal{T}\left[1:c\right], \mathcal{T}\left[c+1:C\right]\right\},
\end{aligned} \tag{4}$$

where $\mathcal{I}_{1,2}$ represents the MSMT. $\mathcal{T}\left[\cdot\right]$ indicates the features processed by ITU. $\mathcal{CAT}$ represents the concatenation. We set up two layers of MSMT in series in the ITU, and the weights between different layers are shared to achieve lossless and invertible mapping of features.

For MSMT ($\mathcal{I}_{1,2}$), given input $\mathcal{F}_{mid}$, we employ simple and effective depthwise-separable (DS) convolutions (kernel sizes: 1, 3, 5, and 7) to process the input to minimize computational effort and generate the data of the self-attention from multiple perspectives: ***Init***: ($\mathcal{O} = \mathcal{F}_{mid}|_{k=1}$), ***Query***: ($\mathcal{Q} = \mathcal{F}_{mid}|_{k=3}$), ***Key***: ($\mathcal{K} = \mathcal{F}_{mid}|_{k=5}$), and ***Value***: ($\mathcal{V} = \mathcal{F}_{mid}|_{k=7}$). Considering that content information is directly related to spatial location [16], employing large-kernel convolutions enables self-attention to model contextual associations over more valuable structural regions. After completing the above steps, the self-attention computation proceeds as follows:

$$\textbf{Att} = \mathcal{L}_N * \left(Softmax\left(\left(\mathcal{Q} * \mathcal{K}\right)/d\right) \cdot \mathcal{V} + \mathcal{O}\right), \tag{5}$$

where $\mathcal{L}_N$ is the layer normalization and d is the scaling factor. The purpose of introducing $\mathcal{O}$ is to give it the self-attention results to enhance the ability to express structural information. The results need to be fed into the MLP ($\mathcal{M}_\ell$) to further enhance the nonlinearity of the modeling process:

$$\mathcal{W}_{\textbf{att}} = \mathcal{D}_r * \mathcal{M}_\ell\left(\textbf{Att}\right) + \mathcal{F}_{mid}, \tag{6}$$

where $\mathcal{D}_r$ represents the Dropout operation. Finally, the features $\nabla\mathcal{F}_{mid}$ and $\Delta\mathcal{F}_{mid}$ generated by SRSP and WRSP are fused with the encoder's output features and fed into the segmentation decoder to produce the segmentation results.

B. Darkclothes Dataset

To address the lack of datasets for garment grasping under low-light conditions, we construct Darkclothes, a low-light garment dataset. To enrich illumination diversity, we varied lighting levels (0–50 luminance, maximum 255) by adjusting curtain openness in a controlled environment. Darkclothes includes five categories: background, shirts, pants, tables, and baskets. Data were collected using a Kinect v2 system equipped with visible-light and infrared sensors. Infrared and low-light images were aligned via the Kinect v2's default RGB–IR stereo calibration, followed by joint calibration and cropping to ensure spatial consistency. The dataset contains 1,023 pixel-aligned groups (512×424 resolution), each comprising a low-light image, its normal-light counterpart, an infrared image, and pixel-wise annotations. Darkclothes is partitioned into 921 training and 102 validation groups (9:1 ratio), with the validation set covering the full illumination range to evaluate robustness.

TABLE I

GRASPING PERFORMANCE OF DIFFERENT SEGMENTATION METHODS.

Method	Success Rate		
	Pants	Shirts	Average
PIDNet (RGB)	14 / 25 (64%)	18 / 25 (68%)	64%
MFRS (RGB-T)	18 / 25 (56%)	19 / 25 (72%)	74%
DarkSeg w/o TM	15 / 25 (68%)	19 / 25 (72%)	68%
DarkSeg	**22 / 25 (88%)**	**21 / 25 (84%)**	**86%**

TABLE II

GRASPING PERFORMANCE OF DIFFERENT GRASPING STRATEGIES.

Method	Success Rate		
	Pants	Shirts	Average
DarkSeg $w/$ [7]	19 / 25 (76%)	19 / 25 (76%)	76%
DarkSeg $w/$ [17]	17 / 25 (68%)	20 / 25 (80%)	74%
DarkSeg w/o DOSM	19 / 25 (76%)	20 / 25 (80%)	78%
DarkSeg	**22 / 25 (88%)**	**21 / 25 (84%)**	**86%**

C. Garment Grasping Strategy

To address flexible garment grasping, we propose a depth-perceptive grasping strategy. Individual garments are first separated from DarkSeg's masks, and their geometric centers are computed. To handle irregular garment shapes, we introduce a geometric constraint completion method, applying morphological closing with a 5×5 circular structuring element to refine boundaries and extracting the largest connected component for minimum bounding rectangle fitting:

$$\min Area\left(R\right), \forall p \in S, p \in C, C \in R, \tag{7}$$

where S represents the original segmentation area, R is the fitting rectangle, and p is the pixel point.

The geometric center (x_o, y_o) is determined by calculating the intersection of the rectangle's diagonals. To enhance grasping reliability, we propose a depth-optimal search method. Centered around the geometric center, we construct a candidate region N_a based on the garment's area: $N_a = Area\left(S\right)/k$, where k is a scaling factor (default: 50). For candidate point $(x_i, y_i) \in N_a$, we utilize a candidate point normalization function to ensure the selection of the highest point while avoiding interference from isolated noise points:

$$Score\left(z_i\right) = \frac{z(p) - z_{\min}}{z_{\max} - z_{\min}}, \tag{8}$$

where $z\left(p\right)$ is the depth value of a pixel in N_a, z_{min} and z_{max} are the minimum and maximum depth values of all pixels in the N_a, respectively. z_i is the optimal depth. Finally, the robotic arm adopts a layered motion strategy: it first lifts vertically along the Z-axis from the table center to a safe height, then performs spatial linear interpolation based on Eye-to-Hand calibration results to move above the optimal grasp point (x_i, y_i, z_i) and complete the grasping process.

D. Loss Function

The loss function of DarkSeg comprises infrared loss $\mathcal{L}_{in}$ for training the teacher model and low-light loss $\mathcal{L}_{low}$ for training the student model. For $\mathcal{L}_{in}$ and $\mathcal{L}_{vi}$, we adopt a cross-entropy loss to measure discrepancies between prediction and result: $\mathcal{L}_{in}\left(p_i, g_t\right) = -\sum^{n} g_t \times \log\left(p_i\right)$ and $\mathcal{L}_{low}\left(p_l, g_t\right) = -\sum^{n} g_t \times \log\left(p_l\right)$, where n, g_2, p_i, and p_l represent the class, label, infrared result, and low-light result. Notably, to align features between the student and teacher models, we introduce a structural modeling loss ($\mathcal{L}_{sm}$):

$$\mathcal{L}_{sm} = \frac{1}{H \times W} \sum_{h}^{H} \sum_{w}^{W} \left\| \mathcal{F}_{in}^{mid}\left(h, w\right) - \mathcal{F}_{low}^{mid}\left(h, w\right) \right\|_2, \tag{9}$$

where $\|\cdot\|_2$ denotes the mean squared error. $\mathcal{F}_{in}^{mid}$ and $\mathcal{F}_{low}^{mid}$ represent the intermediate features extracted from the encoder outputs of the teacher and student models. $\mathcal{L}_{sm}$ enforces the student model to emulate the teacher model's structural feature representation by aligning their intermediate layer features. The total loss is formulated as: $\mathcal{L}_{total} = \mathcal{L}_{low} + \mathcal{L}_{in} + \mathcal{L}_{sm}$.

III. EXPERIMENTS

A. Experiments on the Baxter Robotic Platform

To evaluate DarkSeg's performance in garment grasping (pants and shirts) under low-light conditions, we deploy DarkSeg trained on the Darkclothes to a Baxter robot for grasping assessment. The Kinect v2, used for scene capture, is mounted above the Baxter's head, oriented vertically toward the table. The grasping process employs the depth-perceptive strategy to target our specific garments, with 25 trials per round. A successful grasp is defined as the garment being grasped and released into the basket.

1) We compared the grasping performance of different methods under low-light conditions (0-20 luminance), as shown in Tab. I. Compared with DarkSeg, conventional semantic segmentation models like PIDNet and IDRNet exhibited a decrease in average grasping success rates by 20% and 18%, respectively. This is due to the lack of infrared features to correct the sparse structural features, resulting in misclassification of the model, which affects the selection of grasping points. Similarly, without (w/o) the teacher model (TM), DarkSeg also exhibited significantly reduced grasping accuracy under these conditions. This underscores the role of infrared feature guidance in mitigating low-light perception issues, thereby enabling robot to make accurate decisions.

2) To validate the effectiveness of the proposed depth-perceptive grasping strategy (DPG) and its depth-optimal search method (DOSM), we conduct the following ablation studies: 1) substituting the DPG with $(w/)$ alternative grasping strategies ([7] and [17]), and 2) removing DOSM. As shown in Tab. II, the elimination of DOSM resulted in an 8% decline in DarkSeg's grasping accuracy. This degradation occurs because, without DOSM, the model defaults to considering the geometric centroid of candidate regions as grasping points, without evaluating whether the depth values facilitate arm grasping. When replacing our proposed grasping strategy with [7] and [17], the accuracy decreased by 10% and 12%, respectively. Although these baseline methods adopt similar region-optimal point search strategies, they fail to identify depth coordinates most suitable for robotic grasping, resulting in failed grasps.

REFERENCES

[1] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*, pp. 746–760, 2012.

[2] F. Zhang and Y. Demiris, "Learning grasping points for garment manipulation in robot-assisted dressing," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9114–9120, 2020.

[3] Y. Gao, H. J. Chang, and Y. Demiris, "Iterative path optimisation for personalised dressing assistance using vision and force information," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4398–4403, 2016.

[4] Y. Gao, H. J. Chang, and Y. Demiris, "User modelling for personalised dressing assistance by humanoid robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1840–1845, 2015.

[5] J. Qian, T. Weng, L. Zhang, B. Okorn, and D. Held, "Cloth region segmentation for robust grasp selection," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[6] X. Zhu, X. Wang, J. Freer, H. J. Chang, and Y. Gao, "Clothes grasping and unfolding based on rgb-d semantic segmentation," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9471–9477, 2023.

[7] W. Chen, D. Lee, D. Chappell, and N. Rojas, "Learning to grasp clothing structural regions for garment manipulation tasks," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4889–4895, 2023.

[8] M. Niu, Z. Lu, L. Chen, and J. Yang, "Vergnet: Visual enhancement guided robotic grasp detection under low-light condition," *IEEE Robotics and Automation Letters*, vol. 8, no. 12, pp. 8541–8548, 2023.

[9] H. Zhang, X. Zuo, J. Jiang, C. Guo, and J. Ma, "Mrfs: Mutually reinforcing image fusion and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26974–26983, 2024.

[10] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077–12090, Curran Associates, Inc., 2021.

[11] X. Xu, R. Wang, and J. Lu, "Low-light image enhancement via structure modeling and guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9893–9903, 2023.

[12] L. Peng, C. Zhu, and L. Bian, "U-shape transformer for underwater image enhancement," *IEEE Transactions on Image Processing*, vol. 32, pp. 3066–3079, 2023.

[13] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12302–12311, 2023.

[14] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," in *International Conference on Learning Representations*, 2017.

[15] X. Zhu, Z. Li, X.-Y. Zhang, C. Li, Y. Liu, and Z. Xue, "Residual invertible spatio-temporal network for video super-resolution," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 5981–5988, 2019.

[16] L. Tang, J. Ma, H. Zhang, and X. Guo, "Drlie: Flexible low-light image enhancement via disentangled representations," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022.

[17] H. Shehawy, P. Rocco, and A. M. Zanchettin, "Estimating a garment grasping point for robot," in *2021 20th International Conference on Advanced Robotics (ICAR)*, pp. 707–714, 2021.