FDTDNET: LENSLESS OBJECT SEGMENTATION VIA FEATURE DEMULTIPLEXING AND TASK DECOUPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Camera-based vision systems pose privacy risks, whereas lensless cameras present a viable alternative by omitting visual semantics from their measurements due to the absence of lenses. However, these captured lensless measurements pose challenges for existing computer vision tasks such as object segmentation that usually require visual input. To address this problem, we propose a lensless object segmentation network via feature demultiplexing and task decoupling (FDTDNet) to perform object segmentation for lensless measurements. Specifically, we propose an optical-aware feature demultiplexing mechanism to get meaningful features from lensless measurements without visual reconstruction and design a multi-task learning framework decoupling the lensless object segmentation task into two subtasks, i.e., the reason for contour distribution maps (CDM) and body distribution maps (BDM), respectively. Extensive experiments demonstrate that our FDTDNet achieves highly accurate segmentation effect, which sheds light on privacy-preserving high-level vision with compact lensless cameras.

1 INTRODUCTION

Lensless cameras (Tan et al. (2019); Pan et al. (2021b;a); Salman et al. (2022)) utilize simple, planar optics to convert light into complex patterns, rendering the images unintelligible without knowledge of the mask configurations. Their enhanced privacy features make them promising for privacy-focused applications (Pan et al. (2021b); Yin et al. (2022)). For object segmentation in Fig. 1(a), traditional systems use converging lenses to capture clear images before applying segmentation algorithms, making them vulnerable to network attacks. Lensless cameras produce ambiguous measurements that help safeguard sensitive information (Pan et al. (2021b); Yin et al. (2022)). The typical lensless method (You et al. (2022)) involves restoring the image using the mask, followed by conventional segmentation, as illustrated in Fig. 1(b). However, this method has drawbacks: it prevents data hijacking but is still susceptible to software attacks on reconstructed images. Additionally, segmentation accuracy suffers from blurry reconstructions and suboptimal mask designs Yin et al. (2022), while the reconstruction adds computational overhead, making it less suitable for edge computing.

To enhance segmentation accuracy while ensuring privacy, we propose a one-step method for lensless object segmentation. Unlike the classical two-step process, our method directly segments objects from lensless measurements without intermediate reconstructions, as shown in Fig. 1(c). However, extracting sufficient semantic features in the absence of visual input presents a severe challenge.

To overcome this limitation, we propose an optical-aware feature demultiplexing (OFD) mechanism aimed at refining the features obtained from lensless measurements. This concept is underpinned by the observation that lensless measurements exhibit a direct linear correlation with visual images through the measurement matrix. Similarly, the semantic features corresponding to these measurements can be represented through a linear relationship based on the semantic attributes of the visual images. Considering this, we define Y as the lensless measurement, A as the measurement matrix, and X as the original image. Correspondingly, Y_{θ_i}, A_θ, and X_{θ_i} denote the associated semantic features. Inspired by (Dong et al. (2021)), the above relationship can be succinctly articulated by the following linear equation,

$$Y = A \circ X \Longrightarrow Y_{\theta_i} = A_\theta \circ X_{\theta_i}. \tag{1}$$

Building upon this correlation, we propose a feature demultiplexing and task decoupling network (FDTDNet) for lensless object segmentation. Our approach reconstructs desired features X_{θ_i} from

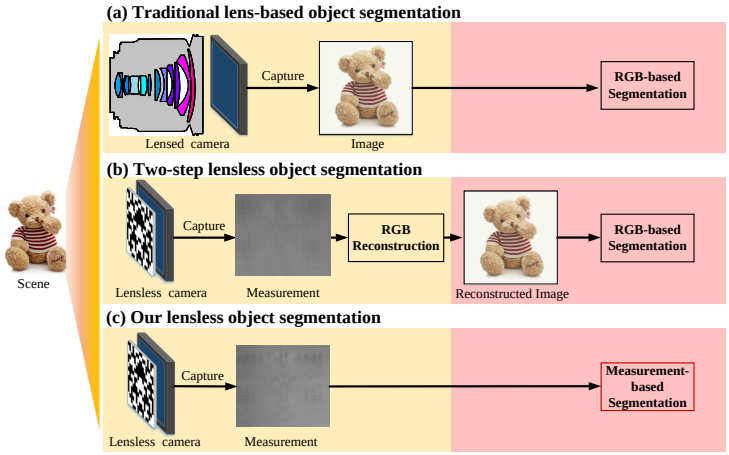

Figure 1: Comparison of object segmentation methods: traditional lens-based (top), two-step lensless (middle), and our optimized one-step lensless methods. Ours improves privacy protection against data interception and software vulnerabilities while maintaining robust segmentation effect.

semantic features Y_{θ_i} while preserving privacy, utilizing the OFD mechanism. By integrating OFD with a Pyramid Vision Transformer (PVT), we enhance long-range feature extraction to tackle segmentation challenges. We further decouple segmentation labels into a contour distribution map (CDM) and a body distribution map (BDM) to mitigate imbalanced pixel distribution issues. To facilitate effective aggregation of CDM and BDM, we introduce a mutual learning strategy using the contour-body interaction (CBI) module. Our main contributions are as follows:

• To our best knowledge, we investigate direct object segmentation from lensless measurements and propose a high-accuracy lensless object segmentation method, which verifies the potential of applying lensless imaging directly to various high-level tasks.

• We model the linear equation between the semantic features bound to lensless measurements and those corresponding to visual inputs. By the proposed OFD, we obtain the expected semantic features to enhance prediction performance.

• We decouple the segmentation task into CDM and BDM inference by contour-/body-distribution learning branches. And a contour-body interaction (CBI) module is proposed for reasoning segmentation results from correlations between CDM and BDM.

• Extensive experiments on two datasets ($i.e.$, directly captured (DIRC) dataset and display captured (DISC) dataset) indicate that our FDTDNet outperforms state-of-the-art methods by a large margin.

2 Related Works

2.1 Lensless Imaging

Lensless imaging (M. Salman et al. (2017); Nick et al. (2018); Pan et al. (2022); Jiachen et al. (2020)) provides an effective way to handdle size constraints in areas like smartphone photography and micro-robotics, relying on masks with amplitude and phase encoding as key components. Different mask architectures have driven the creation of prototypes such as the Fresnel Zone Aperture (FZA) camera (Jiachen et al. (2020); Wu et al. (2021)), FlatCam (M. Salman et al. (2017)), PhlatCam (Nick et al. (2018)), and DiffuserCam (Vivek et al. (2020)). These prototypes have proven valuable in areas such as hyperspectral imaging (Monakhova et al. (2020)), fluorescence microscopy (Alok et al. (2017)), light field encoding Tajima et al. (2017); Cai et al. (2020), and depth sensing (Nick et al. (2018); Tian & Yang (2022)). Recently, researchers have expanded lensless imaging to high-level semantic tasks, successfully achieving recognition (Pan et al. (2021a); Tan et al. (2019); Zhang et al. (2022); Aschenbrenner et al. (2024)), face verification (Tan et al. (2019); Cai et al. (2024)), and object segmentation (Yin et al. (2022; 2024)), showing its potential for high-level inference tasks.

2.2 RECONSTRUCTION-FREE SEMANTIC INFERENCE

Reconstruction-free semantic inference has attracted significant attention, finding applications in fields such as biomedicine, agriculture, and non-visual recognition (Lei et al. (2019); Isogawa et al. (2020); Qiu et al. (2024)). This kind of method offers key benefits in terms of privacy-preserving and reduced computational costs, especially in image recognition (Dave et al. (2022); Hinojosa et al. (2022)). In single-pixel cameras, it enhances computational efficiency (Ji et al. (2022); Liu et al. (2023)), and in lensless cameras, it enables tasks like classification directly from raw measurements (Cai et al. (2024); Perez et al. (2024); Yang et al. (2024)). Recent research has focused on pixel-level reasoning tasks like image segmentation (Yang et al. (2022)). In (You et al. (2022)), human eye segmentation was studied using a reconstruction-before-segmentation approach, but the high computational cost limited practical use. The works (Yin et al. (2022; 2024)) introduced an end-to-end network for segmenting objects from lensless imaging data, but its performance was constrained by the need for original scene supervision. Thus, achieving high-precision segmentation in lensless imaging remains challenges.

3 METHODOLOGY

3.1 MOTIVATION AND OVERVIEW

Among various lensless camera prototypes, FlatCam (M. Salman et al. (2017)) stands out for its wide range of applications due to its high luminous flux, lightweight setups, and cost-effectiveness. We investigate object segmentation based on the FlatCam imaging model, although our method is also easily adaptable to other lensless camera models. FlatCam utilizes a separable mask pattern, $i.e.$, the 2-D mask pattern, which can be represented by the outer product of two 1-D patterns. The imaging model is formulated as

$$Y = A_{\mathrm{L}} X A_{\mathrm{R}}^{\top} + \xi, \tag{2}$$

where $A_{\mathrm{L}} \in \mathcal{R}^{V \times M}$ and $A_{\mathrm{R}}^{\top} \in \mathcal{R}^{N \times W}$ denote matrices that correspond 1D convolutions along the rows and columns, respectively. The ξ repsents the additive noise.

Lensless cameras produce multiplexed measurements Y devoid of visual information, complicating object segmentation. To tackle this, we propose FDTDNet, a lensless object segmentation method leveraging feature demultiplexing and task decoupling. We first develop an optical-aware feature demultiplexing (OFD) mechanism integrated with a Pyramid Vision Transformer (PVT) for semantic feature decoupling. Then, the segmentation task is divided into contour distribution map (CDM) and body distribution map (BDM) subtasks, enhancing edge feature learning and reducing interference. We implement a contour-distribution learning branch with a dual-path attention (DPA) and a body-distribution learning branch with contextual exploration (CE) and hierarchical information fusion (HIF) for CDM and BDM predictions, respectively. A cross-branch learning strategy via the contour-body interaction (CBI) module further improves segmentation by exploiting the correlation between CDM and BDM.

3.2 OPTICAL-AWARE FEATURE DEMULTIPLEXING (OFD) MECHANISM

Unlike previous works that extract features from natural images, our encoder derive cues from lensless measurements, making conventional encoders ineffective. Thus, we propose a feature demultiplexing mechanism by integrating the OFD at the end of the PVT encoder to mine high-level information. First, we utilize the PVT for feature extraction in lensless object segmentation, generating outputs across four stages:

$$Y_{\theta_1}, Y_{\theta_2}, Y_{\theta_3}, Y_{\theta_4} = \mathrm{PVT}\left(Y\right). \tag{3}$$

Based on Eq. (1) and the lensless imaging model in Eq. (2), the above semantic features Y_{θ_i} $(i = 1, 2, 3, 4)$ is modeled as:

$$Y_{\theta_i} = A_{\mathrm{L},\theta_i} X_{\theta_i} A_{\mathrm{R},\theta_i}^{\top} + \xi, \tag{4}$$

where X_{θ_i}, A_{L,θ_i}, and A_{R,θ_i} denote the X, A_{L}, and A_{R} in the feature space. Therefore, the task of reasoning about X_{θ_i} from Y_{θ_i} can be modeled as an inverse problem. To obtain X_{θ_i} for boosting the lensless object segmentation task, inspired by (Salman et al. (2022)), our OFD-based extractor is designed as the Tikhonov regularization problem as:

$$\arg\min_{X_{\theta_i}} \left\| Y_{\theta_i} - A_{\mathrm{L},\theta_i} X_{\theta_i} A_{\mathrm{R},\theta_i}^{\top} \right\|_2^2 + K_{\theta_i} \left\| X_{\theta_i} \right\|_2^2, \tag{5}$$

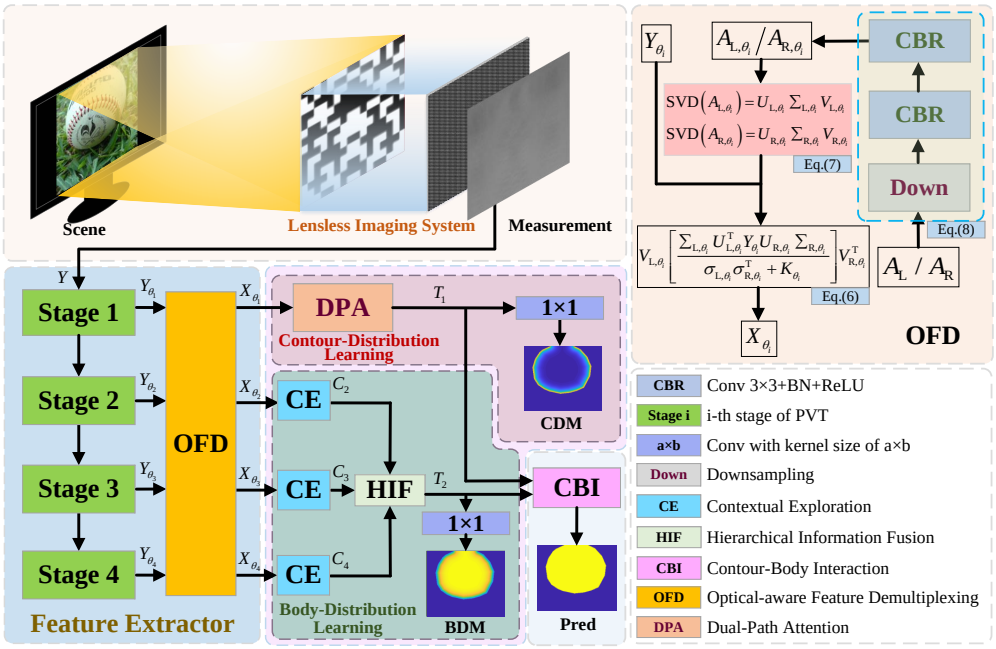

Figure 2: The proposed FDTDNet framework includes: (1) a PVT and OFD-based extractor for reconstructing semantics, (2) a contour-distribution learning branch with the DPA, and a body-distribution learning branch with the CE and HIF for inferring CDM and BDM, respectively, and (3) a CBI-based mutual learning strategy to derive segmentation results from CDM and BDM.

where K_θ is the learnable regularization parameter. The Eq. (5) can be sovlved by Wiener deconvolution (Haywood & Younes (2023)) as:

$$\begin{aligned}\hat{X}_{\theta_i} &= \mathrm{OFD}(Y_{\theta_i}; A_{\mathrm{L},\theta_i}, A_{\mathrm{R},\theta_i}) \\ &= V_{\mathrm{L},\theta_i}[(\Sigma_{\mathrm{L},\theta_i} U_{\mathrm{L},\theta_i}^\top Y_{\theta_i} U_{\mathrm{R},\theta_i} \Sigma_{\mathrm{R},\theta_i})./(\sigma_{\mathrm{L},\theta_i}\sigma_{\mathrm{R},\theta_i}^\top + K_{\theta_i})]V_{\mathrm{R},\theta_i}^\top,\end{aligned} \tag{6}$$

satisfies with

$$A_{\mathrm{L},\theta_i} \overset{\mathrm{SVD}}{=} U_{\mathrm{L},\theta_i}\Sigma_{\mathrm{L},\theta_i}V_{\mathrm{L},\theta_i}^\top, \quad A_{\mathrm{R},\theta_i} \overset{\mathrm{SVD}}{=} U_{\mathrm{R},\theta_i}\Sigma_{\mathrm{R},\theta_i}V_{\mathrm{R},\theta_i}^\top, \tag{7}$$

where SVD is the singular value decomposition (SVD). The A_{L,θ_i} and A_{R,θ_i} are updated by

$$A_{\mathrm{L},\theta_i} = f_{A_\mathrm{L}}(A_\mathrm{L}), A_{\mathrm{R},\theta_i} = f_{A_\mathrm{R}}(A_\mathrm{R}), \tag{8}$$

where $f_{A_\mathrm{L}}(\cdot)$ and $f_{A_\mathrm{R}}(\cdot)$ represent the 3×3 convolution layer + batch normalization (BN) + ReLU + down-sampling operator. We denote each side output as X_{θ_i} for $\hat{X}_{\theta_i}$. The detailed derivation of the OFD is illustrated in Appendix A.1. Importantly, the OFD mechanism facilitates back-end tasks without visual reconstruction, mitigating sensitive privacy leakage.

3.3 Task Decoupling

We assume the contour distribution follows a Gaussian distribution with zero mean and a standard deviation of σ. The ideal contour distribution and body distribution are defined as follows:

$$\mathcal{P}_{\mathrm{cdm}}(\mathbf{p}) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\mathrm{DT}(\mathbf{p}))^2}{2\sigma^2}}, \quad \mathcal{P}_{\mathrm{bdm}}(\mathbf{p}) = 1 - \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\mathrm{DT}(\mathbf{p}))^2}{2\sigma^2}}, \tag{9}$$

where $\sigma = \frac{1}{\sqrt{2\pi}}$ to keep $\mathcal{P}_{\mathrm{cdm}}$ and $\mathcal{P}_{\mathrm{bdm}}$ in the range [0,1]. $\mathrm{DT}(\mathbf{p})$ represents the shortest Euclidean distance from pixel $\mathbf{p}$ to the boundary. $\mathrm{DT}(\mathbf{p})$ is pixel-dependent, varying with classification (foreground or background) and relative position. Pixels closer to the object's center receive higher values, while those farther away or in the background have lower values. We multiply the generated $\mathcal{P}_{\mathrm{cdm}}$ and $\mathcal{P}_{\mathrm{bdm}}$ with the original binary image I to remove the background interference as

$$I_{\mathrm{cdm}} = I \odot \mathcal{P}_{\mathrm{cdm}}, \quad I_{\mathrm{bdm}} = I \odot \mathcal{P}_{\mathrm{bdm}}, \tag{10}$$

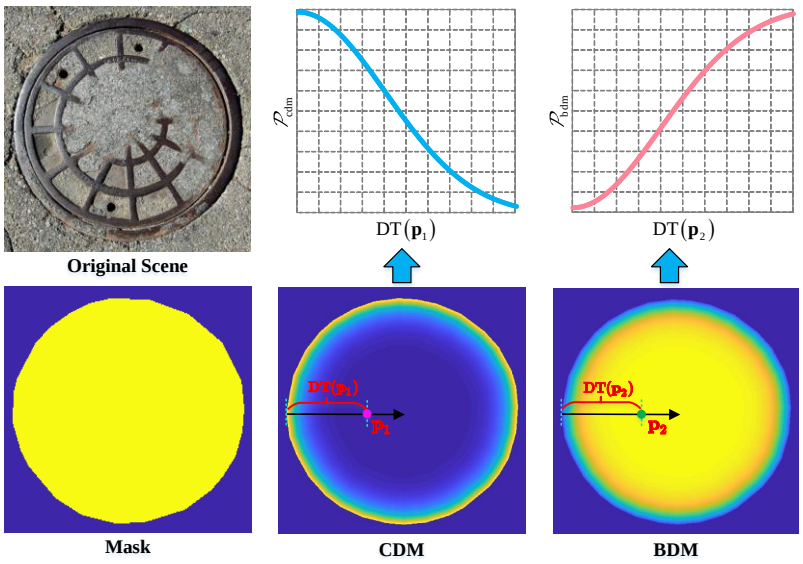

Figure 3: Example of label decoupling: the label is decomposed into a contour distribution map (CDM) and a body distribution map (BDM).

where $\odot$ represents element-wise multiplication. I_{cdm} and I_{bdm} mean the CDM and BDM as shown in Fig. 3, respectively. Accordingly, the segmentation task is decoupled into the inference subtasks for CDM and BDM.

Body-Distribution Learning. The body information is critical for determining the overall segmentation effect. We design a body-distribution learning branch to mine the accurate main region. We feed the multi-level side outputs of OFD into the designed contextual exploration (CE) module for extracting contextual information. Then, we introduce the hierarchical information fusion (HIF) module to aggregate the outputs from the multi-layer CE modules to obtain the BDM results. The details of CE and HIF are explained in Fig. 11 of Appendix A.3.

Contour-Distribution Learning. The contour is usually used as a prime cue to refine the object morphology for accurate segmentation. We design the contour-distribution learning branch consisting of the dual-path attention (DPA) to focus on learning contour information. The details of DPA are explained in Fig. 12 of Appendix A.4.

3.4 Contour-Body Interaction (CBI)

Considering the correlation between CDM and BDM, we propose the CBI in combination with graph convolutional neural networks. As shown in Fig. 4, the CBI consists of three main components: cross-layer correlation, polishing gate, and mask generation.

Cross-layer Correlation. For the input feature map $\{T_1, T_2\} \in \mathcal{R}^{C\times H\times W}$, we apply two 1×1 convolutional layers ($\mathcal{G}_{\text{edge}}$ and $\mathcal{G}_{\text{node}}$) to transform $\{T_1, T_2\}$ into two independent representations, and then extract the transformed feature patches into two groups, *i.e.*, $G' = \{\mathbf{p}'_i \mid 1 \le i \le K\}$ and $G'' = \{\mathbf{p}''_i \mid 1 \le i \le K\}$, via the unfolding operation f_{unfold} (shown in Fig. 4). The feature patches in G' and G'' have the following feature representations:

$$\mathbf{p}'_i = f_{\text{unfold}}\left(\mathcal{G}_{\text{edge}}\left(T_1\right)\right), \quad \mathbf{p}''_i = f_{\text{unfold}}\left(\mathcal{G}_{\text{node}}\left(T_2\right)\right), \tag{11}$$

where G' is used to build graph connections and G'' is assigned as the graph nodes. Given a set of feature patches G', we flatten each patch into a feature vector and compute feature similarity using the dot product, resulting in a similarity matrix $\mathbf{S} \in \mathcal{R}^{K\times K}$, defined as:

$$\mathbf{S} = \text{FC}\left(\text{Flatten}(G'_i)\right) \otimes \text{FC}\left(\text{Flatten}(G'_i)\right), \tag{12}$$

where $\otimes$ denotes the matrix multicaption. $\text{Flatten}\left(\cdot\right)$ is the flatten operator, $\text{FC}\left(\cdot\right)$ is the full connected layer. Consider $\mathbf{S}_{i,:}$, the i-th row of $\mathbf{S}$, representing the similarity of the i-th node to other nodes. We employ a dynamic number of neighbors for each node based on the nearest principle. This is achieved through a dynamic KNN (DKNN) module, which generates an adaptive threshold

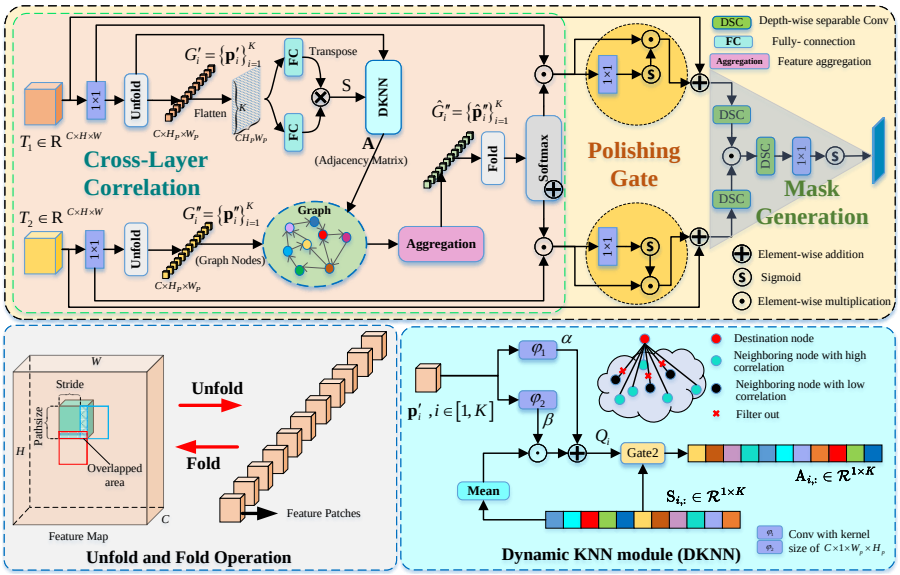

Figure 4: Illustration of the CBI. It consists of three main components: cross-layer correlation, polishing gate, and mask generation.

for each node, selecting neighbors with similarities above this threshold as candidates. The average value of $\mathbf{S}_{i,:}$ represents the average importance of different nodes to the i-th node, denoted as Q_i. As shown in Fig. 4, to improve the adaptability, we apply the node-specific affine transformation to calculate Q_i as:

$$Q_i = \frac{\varphi_2\left(p_i'\right)}{K}\sum_{k=1}^{K}\mathbf{S}_{i,k} + \varphi_1\left(p_i'\right) = \frac{\beta}{K}\sum_{k=1}^{K}\mathbf{S}_{i,k} + \alpha, \tag{13}$$

where $\alpha = \varphi_1\left(p_i'\right)$ and $\beta = \varphi_2\left(p_i'\right)$. φ_1 and φ_2 are two distinct $W_p \times H_p$ convolutional layers, embedding each node into specific affine transform parameters, *i.e.*, α and β. To achieve a different threshold truncation, we utilize the ReLU function to truncate input features and normalize the similarity of all connected nodes by the softmax function to calculate the attention weights by

$$\alpha_{i,j} = \frac{\exp\left(\mathbf{A}_{i,j}\right)}{\sum_{j\in\mathcal{N}_i}\exp\left(\mathbf{A}_{i,j}\right)}, j \in \mathcal{N}_i, \quad \mathbf{A}_{i,:} = \mathrm{ReLU}\left(\mathbf{S}_{i,:} - Q_i\right), \tag{14}$$

where $\mathbf{A} \in \mathcal{R}^{K\times K}$ is the adjacency matrix in which $\mathbf{A}_{i,j}$ is assigned the similarity weight if $\mathbf{p}_j'$ connect to $\mathbf{p}_i'$, otherwise equal to zero. $\mathcal{N}_i$ is the set of indexes of neighboring nodes. The feature aggregation process is a graph described as a weighted sum of all connected neighbors:

$$\hat{\mathbf{p}}_i = \sum_{j\in\mathcal{N}_i}\alpha_{i,j}\times\mathbf{p}_j'' = \sum_{j\in\mathcal{N}_i}\alpha_{i,j}\times\mathcal{G}_{\mathrm{node}}\left(\mathbf{p}_j\right). \tag{15}$$

We extract feature patches from the graph to aggregate into a feature map via folding operation. Overlapping regions are handled by averaging to suppress blocking effects. Global residual connectivity in the cross-layer correlation module further enhances the result. The output of this module is expressed as $r = f_{\mathrm{fold}}\left(\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, ..., \hat{\mathbf{p}}_i, ..., \hat{\mathbf{p}}_K\right)$, $f_{\mathrm{fold}}\left(\cdot\right)$ is a folding operation, as shown in Fig. 4. The extracted cross-layer correlation matrix r is normalized by a softmax operator along the rows and columns, respectively, to locate the object regions involved in the high-level semantics by

$$F_{\mathrm{corr}}^1 = \mathrm{Rp}\left(\mathrm{Rp}\left(\mathcal{G}_{\mathrm{edge}}\left(T_1\right)\right)\odot\mathcal{S}(r)\right), F_{\mathrm{corr}}^2 = \mathrm{Rp}\left(\mathrm{Rp}\left(\mathcal{G}_{\mathrm{node}}\left(T_2\right)\right)\odot\mathcal{S}\left(r^\top\right)\right), \tag{16}$$

where $\mathrm{Rp}(\cdot)$ is the reshape operation and $\mathcal{S}(\cdot)$ is the softmax operator. $F_{\mathrm{corr}}^1, F_{\mathrm{corr}}^2 \in \mathcal{R}^{C\times H\times W}$ are features containing rich location information. Since we perform matrix-based cross-layer correlation operations on F_{corr}^1 and F_{corr}^2 with low computational cost.

Polishing Gate. To address redundancy in $\{F_{\mathrm{corr}}^1, F_{\mathrm{corr}}^2\}$, we introduce an effective gating mechanism to refine location information. Using a 1×1 convolution, we generate response maps in

$[0,1]^{1\times H\times W}$ for $\{F^1_{\text{corr}}, F^2_{\text{corr}}\}$. These maps filter out redundant information by gating mechanism as $F^1_{\text{gate}} = \text{Sigmoid}\left(\text{Conv}_{1\times1}\left(F^1_{\text{corr}}\right)\right) \odot F^1_{\text{corr}}, F^2_{\text{gate}} = \text{Sigmoid}\left(\text{Conv}_{1\times1}\left(F^2_{\text{corr}}\right)\right) \odot F^2_{\text{corr}}$, $\left\{F^1_{\text{gate}}, F^2_{\text{gate}}\right\} \in \mathcal{R}^{C\times H\times W}$ are the polished features. $\text{Conv}_{1\times1}$ is the 1×1 convolution layer.

Mask Generation. Moreover, we adopt the residual connection to merge F^1_{gate} and T_1 as well as F^2_{gate} and T_2, respectively, resulting $\hat{F}^1_{\text{gate}}$ and $\hat{F}^2_{\text{gate}}$ by $\hat{F}^1_{\text{gate}} = \text{DSConv}\left(F^1_{\text{gate}} + T_1\right), \hat{F}^2_{\text{gate}} = \text{DSConv}\left(F^2_{\text{gate}} + T_2\right)$, $\text{DSConv}(\cdot)$ is the 3×3 depth-wise separable convolution layer. The generated $\hat{F}^1_{\text{gate}}$ and $\hat{F}^2_{\text{gate}}$ are fused to generate the segmenatation map by $P_{\text{seg}} = \text{Sigmoid}\left(\text{Conv}_{1\times1}\left(\text{DSConv}\left(\hat{F}^1_{\text{gate}} \odot \hat{F}^2_{\text{gate}}\right)\right)\right)$. We completely extract location information from $\hat{F}^1_{\text{gate}}$ and $\hat{F}^2_{\text{gate}}$ to accurately determine the object regions.

3.5 Loss Function

To well train the FDTDNet, we combine the weighted BCE loss ℓ_{wBCE} (Wei et al. (2020)) and weighted IoU loss ℓ_{wIOU} (Wei et al. (2020)), that is, $L_{\text{s}} = \ell_{\text{wBCE}} + \ell_{\text{wIOU}}$ to perform supervised learning on the CDM, BDM, and final segmentation maps. Thus the total loss function is:

$$L_{\text{All}} = L_{\text{s}}(P_{\text{CDM}}, G_{\text{CDM}}) + L_{\text{s}}(P_{\text{BDM}}, G_{\text{BDM}}) + L_{\text{s}}(P_{\text{seg}}, G_{\text{seg}}), \tag{17}$$

where $P_{\text{CDM}}, P_{\text{BDM}}$, and P_{seg} are the predicted CDM, BDM, and final segmentation maps, respectively. $G_{\text{CDM}}, G_{\text{BDM}}$, and G_{seg} are the true CDM, BDM, and segmentation maps, respectively.

4 Experiments

4.1 Setups

Datasets. We use the datasets (Yin et al. (2024)) for lensless object segmentation named directly captured (DIRC) dataset and display captured (DISC) dataset (Yin et al. (2024)). The DIRC dataset is used for testing, and it consists of 30 natural scene images directly captured from 10 different scenes. The DISC dataset is collected from Display, including 5.2K paired data for training (DISC-Train) and 0.7K for testing (DISC-Test). Note that the measurements are captured by FlatCam.

Evaluation Metrics. To quantitatively evaluate the performance of each method, we use six evaluation matrices, including mean absolute error ($\mathcal{M}$), mean E-measure (E_ξ) (Fan et al. (2021)), weighted F-measure(F^w_β) (Margolin et al. (2014)), S-measure (S_α) (Fan et al. (2017)), mean Dice (mDice), and mean IoU (mIoU).

Implementation Details. In our FDTDNet, the PVT pre-trained on ImageNet initializes the backbone. We train the FDTDNet by the Adam optimizer with "cosine" learning rate policy as $lr = 0.5 \times init\_r \times (1 + \cos(\pi * epoch/max\_epoch))$, where the initial learning rate $init\_r$ is set to 5×10^{-4} and training epoch $epoch \in [1, max\_epoch]$, $max\_epoch = 100$. The whole network is trained with a batch size of 8. All experiments are implemented in Pytorch 1.8.0 and trained on a Linux 20.04 server with a single GPU of NVIDIA RTX 3090.

Compared Methods. For a fair evaluation, we compare our method with following methods: (1) Current advanced object segmentation methods, including CDMNet (Song et al. (2023)), SINetV2 (Fan et al. (2022)), C2FNet (Chen et al. (2022)), OCENet (Liu et al. (2022)), ZoomNet (Pang et al. (2022)), TransUnet (Chen et al. (2021)), and BDG-Net (Qiu et al. (2022)); And (2) Existing object inference methods for lensless imaging: LLI_T (Pan et al. (2021a)), Raw3dNet (Zhang et al. (2022)), EyeCoD (You et al. (2022)), LOINet (Yin et al. (2022)), and RecSegNet (Yin et al. (2024)). We employ open-source codes from public repositories to implement established comparison methods. To ensure consistency, all methods are retrained on a shared training dataset.

4.2 Comparison with State-of-the-Arts

Evaluation on DISC-Test Dataset. Figure 5 displays segmentation results from our FDTDNet and various state-of-the-art methods (CDMNet, C2FNet, SINetV2, BDG-Net, OCENet, TransUnet, and

Table 1: Comparison of our FDTDNet and other 12 state-of-the-art methods. ↑ means that the more prominent, the better, and ↓ means that the more minor, the more remarkable. The first, second, and third-ranked performances are highlighted in **red**, **green**, and **blue**, respectively.

Methods	DISC-Test						DIRC					
	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$S_\alpha \uparrow$	mDice ↑	mIoU ↑	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$S_\alpha \uparrow$	mDice ↑	mIoU ↑
CDMNet	0.535	0.241	0.688	0.652	0.618	0.473	0.739	0.117	0.805	0.738	0.756	0.679
C2FNet	0.493	0.291	0.639	0.562	0.557	0.405	0.713	0.119	0.793	0.763	0.826	0.704
SINetV2	0.363	0.360	0.502	0.357	0.365	0.399	0.658	0.126	0.754	0.732	0.785	0.697
BDG-Net	0.508	0.261	0.665	0.582	0.553	0.405	0.645	0.151	0.746	0.714	0.768	0.679
OCENet	0.585	0.222	0.711	0.628	0.632	0.499	0.767	0.116	0.829	0.794	0.835	0.726
TransUNet	0.551	0.242	0.743	0.678	0.593	0.463	0.764	0.117	0.817	0.781	0.833	0.708
ZoomNet	0.661	0.177	0.811	0.753	0.716	0.605	0.773	0.115	0.815	0.787	0.840	0.752
LLI_T	0.721	0.137	0.802	0.748	0.764	0.669	0.742	0.115	0.821	0.759	0.817	0.732
Raw3dNet	0.749	0.118	0.827	0.752	0.777	0.674	0.779	0.105	0.834	0.778	0.836	0.749
EyeCoD	0.755	0.127	0.808	0.756	0.782	0.679	0.785	0.097	0.838	0.786	0.833	0.752
LOINet	**0.763**	**0.129**	**0.832**	**0.764**	**0.799**	**0712**	**0.791**	**0.103**	**0.844**	**0.792**	**0.858**	**0.779**
RecSegNet	**0.866**	**0.067**	**0.907**	**0.861**	**0.879**	**0.818**	**0.854**	**0.078**	**0.858**	**0.891**	**0.867**	**0.824**
FDTDNet	**0.902**	**0.056**	**0.916**	**0.875**	**0.902**	**0.841**	**0.918**	**0.047**	**0.923**	**0.903**	**0.907**	**0.874**

ZoomNet) on the DISC-Test dataset. Many comparison methods struggle with low-contrast (3rd, 4th, 6th rows) and cluttered backgrounds (1st, 2nd, 5th rows), failing to segment objects accurately due to their limited capacity to extract details from lensless measurements. In contrast, FDTDNet employs feature demultiplexing, yielding superior segmentation. Further analysis in Fig. 13 in Appendix A.6 compares our method with existing lensless segmentation techniques (LLI_T, Raw3dNet, EyeCoD, LOINet, and RecSegNet). While these methods perform well, FDTDNet achieves results closest to ground truths. Tab. 1 quantifies lensless object segmentation performance, showing FDTDNet outperforms all competitors across metrics. Specifically, it reduces $\mathcal{M}$ by 12.5% and improves F_β^w, E_ξ, S_α, mDice, and mIoU by 3.3%, 0.4%, 2.0%, 2.0%, and 2.8%, respectively, compared to RecSegNet. Note that our method advances task decomposition by using CDM and BDM with a CBI module for mutual learning, tailored to lensless imaging's ambiguous boundaries. Unlike CDMNet's edge-based focus, our dual-branch design achieves more comprehensive segmentation.

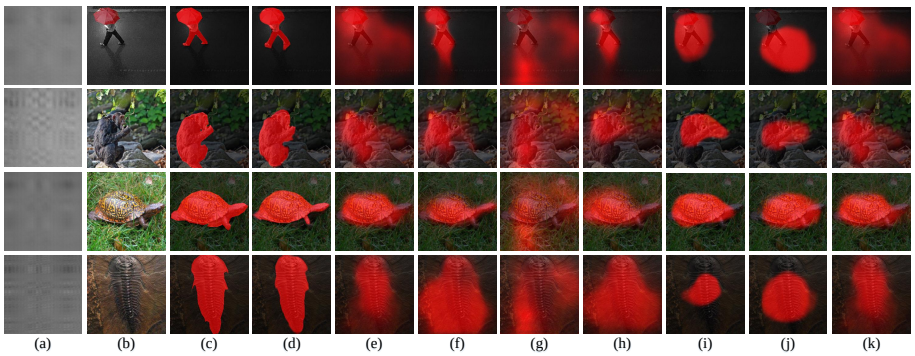

Figure 5: Comparison with state-of-the-art methods on the DISC-Test dataset. The (a) is the lensless measurements corresponding to real images (b); The (c) is the real segmentation maps corresponding to (b); The (d)–(k) are the segmentation results by our FDTDNet, CDMNet, C2FNet, SINetV2, BDG-Net, OCENet, TransUnet, and ZoomNet.

Evaluation on DIRC Dataset. With limited visual input causing failures in most comparison methods on DIRC dataset, we adopt the training setups in (Yin et al. (2024)) to obtain results, as shown in Tab. 1. While comparison methods with the setups in (Yin et al. (2024)) perform well on DIRC dataset due to simpler objects and uniform backgrounds, our FDTDNet consistently outperforms them across all metrics. Notably, it reduces $\mathcal{M}$ by 34.7% and enhances F_β^w, E_ξ, S_α, mDice, and mIoU by 7.1%, 6.7%, 0.9%, 4.0%, and 5.3%, respectively, compared to RecSegNet. Fig. 6 show-

cases segmentation results, revealing that many methods struggle with accurate segmentation due to the lack of visual semantics in lensless measurements. In contrast, FDTDNet excels, demonstrating its robust generalization capability.

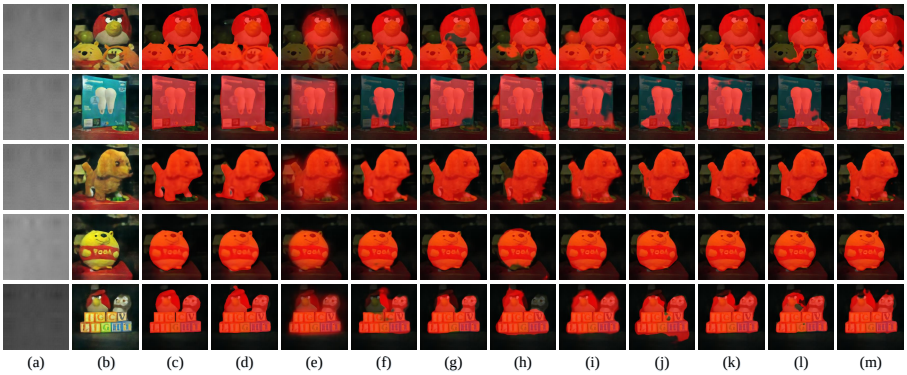

Figure 6: Comparison with state-of-the-art methods on the DIRC dataset. The (a) is lensless measurements; (b) is the restored images by FlatNet (Salman et al. (2022)); The (c) is the real segmentation maps corresponding to (b); The (d)–(m) are the segmentation results by our FDTDNet, CDMNet, C2FNet, SINetV2, BDG-Net, OCENet, TransUnet, ZoomNet, LOINet, and RecSegNet.

4.3 Ablation Studies

Ablation Studies on Tasks. Table 2 presents the comparison results obtained through various task supervisions. Note that "Segm" refers to the direct segmentation map supervision, "Edge" denotes the edge supervision, "CDM" represents the CDM supervision, and "BDM" symbolizes the BDM supervision. The configuration incorporating CDM outperforms the configuration involving edge, suggesting that CDM supervision is more effective than edge supervision. Furthermore, the amalgamation featuring BDM exhibits superior performance compared with the configuration incorporating segmentation maps. This validates that a more effective feature representation could be learned for the body regions without interfering with edges.

Table 2: Comparison of different task supervision on DISC-Test dataset. The first-ranked result is highlighted in **red**.

Task	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$S_\alpha \uparrow$	mDice ↑	mIoU ↑
CDM + Segm	0.889	0.065	0.881	0.864	0.887	0.814
Edge+ Segm	0.882	0.066	0.882	0.861	0.885	0.815
BDM + Edge	0.885	0.066	0.885	0.861	0.883	0.814
BDM + Segm	0.889	0.067	0.886	0.863	0.884	0.814
Full model	**0.902**	**0.056**	**0.916**	**0.875**	**0.902**	**0.841**

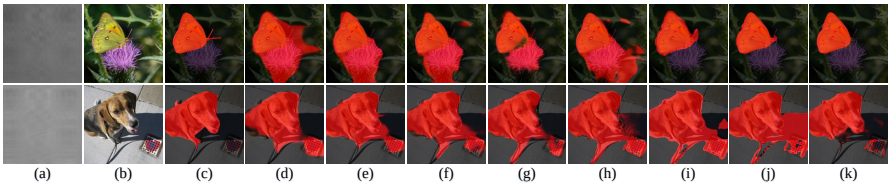

Figure 7: Ablation studies on the DISC-Test dataset. The (d)–(k) corresponding to Conf_1–Conf_8. The (a) is the lensless measurements of the underlying scenes (b) by the lensless camera; The (c) is the ground truth segmentation maps corresponding to (b).

Ablation Studies on Components. We explore the effectiveness of each component in our FDTDNet. Note that removed HIF is replaced by upsampling + concatenation + convolution, removed DPA is replaced by 3×3 convolution layer, and removed CBI is replaced by concatenation. For the above configuration, we obtain the corresponding evaluation results, as illustrated in Fig. 7 and

Tab. 3. From the results in Fig. 7 (d)–(h) and Conf_1–Conf_5 in Tab. 3, the removal of each component ($i.e.$, CE, OFD, DPA, HIF, and CBI) results in a drop of segmentation performance. These results demonstrate the effectiveness of the individual components.

Ablation Studies on Loss Functions. To further explore the effect of our method, we analyze each loss function, and the corresponding results are shown in Fig. 7(i), (j), as well as Conf_6 and Conf_7 in Tab. 3. Employing only either ℓ_{wBCE} or ℓ_{wIOU} leads to a degradation of predicted effect, while better results are obtained by training our network with the total loss function ($i.e.$, Fig. 7(k) and Conf_8 in Tab. 3). These results indicate that a tailored loss functions are necessary for our FDTDNet.

Table 3: Ablation studies on DISC-Test dataset. The first-ranked result is highlighted in **red**.

$\mathrm{Conf}_{\mathrm{ID}}$	Component					Loss Function		Evaluation Metrics					
	OFD	CE	DPA	HIF	CBI	L_{wIOU}	L_{wBCE}	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$S_\alpha \uparrow$	mDice ↑	mIoU ↑
Conf_1		✓	✓	✓	✓	✓	✓	0.652	0.231	0.694	0.667	0.708	0.588
Conf_2	✓		✓	✓	✓	✓	✓	0.676	0.207	0.726	0.723	0.771	0.613
Conf_3	✓	✓		✓	✓	✓	✓	0.683	0.189	0.747	0.726	0.805	0.621
Conf_4	✓	✓	✓		✓	✓	✓	0.852	0.084	0.867	0.805	0.877	0.805
Conf_5	✓	✓	✓	✓		✓	✓	0.746	0.173	0.776	0.728	0.748	0.698
Conf_6	✓	✓	✓	✓	✓	✓		0.824	0.103	0.857	0.804	0.829	0.783
Conf_7	✓	✓	✓	✓	✓		✓	0.867	0.062	0.901	0.837	0.878	0.803
Conf_8	✓	✓	✓	✓	✓	✓	✓	**0.902**	**0.056**	**0.916**	**0.875**	**0.902**	**0.841**

4.4 LIMITATIONS

While our method performs well in conventional scenarios, it shows performance degradation in unconventional cases, as analyzed in Fig. 8. Specifically, non-uniform or hollow target regions lead to missed detections (first row), small targets result in false positives (second row), and blurred target boundaries cause significant false positives. These issues arise from inherent challenges in lensless imaging, such as optical cross-talk and complex scenes. The method excels with flat, high-intensity, large targets but struggles in more complex conditions. Future work should address these limitations by (1) expanding datasets to include diverse scenarios, (2) applying domain adaptation for improved generalization, and (3) adopting frequency-adaptive techniques to mitigate cross-talk artifacts.

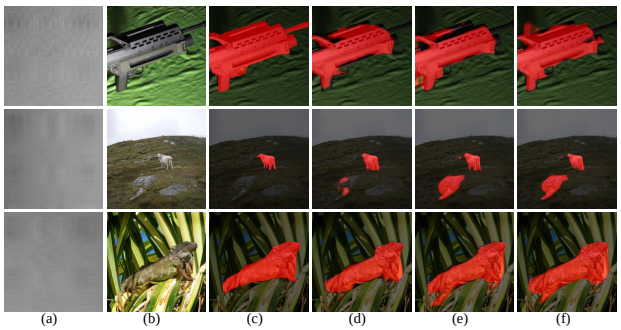

Figure 8: Illustration of failure cases. The (a) is the lensless measurements of the underlying scene (b) by the lensless camera; The (c) is the ground truth segmentation maps corresponding to (b); The (d)–(f) are the segmentation results from our method, LOINet, and RecSegNet, respectively.

5 CONCLUSION

This paper addresses the challenges of lensless object segmentation by the proposed one-step method, FDTDNet, developing an optical-aware feature demultiplexing mechanism and decomposing the task into CDM and BDM inference subtasks. For the former, the FDTDNet applies a new extractor combining OFD and PVT for reconstructing semantic features. Moreover, for the latter, we enhance lensless object segmentation performance by incorporating contour-distribution and body-distribution learning branches and a contour-body interaction strategy. Extensive experiments on the DISC-Test and DIRC datasets show that our FDTDNet outperforms state-of-the-art methods across various evaluation metrics and highlights its potential in advancing the field of lensless imaging.

REFERENCES

Singh Alok, Kumar, Pedrini Giancarlo, Mitsuo Takeda, and Osten Wolfgang. Scatter-plate microscope for lensless microscopy with diffraction limited resolution. *Scientific reports*, 7(10687): 1–8, Sep 2017.

Gregory Aschenbrenner, Peter M Douglass, Timothy O'Connor, Saurabh Goswami, Kashif Usmani, and Bahram Javidi. Lensless imaging systems: a technological and clinical review for automated disease identification. In *Three-Dimensional Imaging, Visualization, and Display 2024*, volume 13041, pp. 68–69. SPIE, 2024.

Xin Cai, Hailong Zhang, Chenchen Wang, Wentao Liu, Jinwei Gu, and Tianfan Xue. Lenslessface: An end-to-end optimized lensless system for privacy-preserving face verification. *arXiv preprint arXiv:2406.04129*, 2024.

Zewei Cai, Jiawei Chen, Giancarlo Pedrini, Wolfgang Osten, Xiaoli Liu, and Xiang Peng. Lensless light-field imaging through diffuser encoding. *Light: Science & Applications*, 143(9):1–9, 2020.

Geng Chen, Sijie Liu, Yujia Sun, Gepeng Ji, Yafeng Wu, and Tao Zhou. Camouflaged object detection via context-aware cross-level fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6981–6993, 2022. doi: 10.1109/TCSVT.2022.3178173.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. Spact: Self-supervised privacy preservation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20164–20173, 2022.

Jiangxin Dong, Stefan Roth, and Bernt Schiele. Dwdn: Deep wiener deconvolution network for non-blind image deblurring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9960–9976, 2021.

Dengping Fan, Mingming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-Measure: A new way to evaluate foreground maps. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4558–4567, 2017. doi: 10.1109/ICCV.2017.487.

Dengping Fan, Gepeng Ji, Xuebin Qian, and Mingming Cheng. Cognitive vision inspired object segmentation metric and loss function. *SCIENTIA SINICA Informationis*, 51(9):1475, 2021.

Dengping Fan, Gepeng Ji, Mingming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042, 2022. doi: 10.1109/TPAMI.2021.3085766.

Charlie Haywood and Rabih Younes. Real-time blind deblurring based on lightweight deep-wiener-network. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2023.

Carlos Hinojosa, Miguel Marquez, Henry Arguello, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Privhar: Recognizing human actions from privacy-preserving lens. In *European Conference on Computer Vision*, pp. 314–332. Springer, 2022.

Mariko Isogawa, Ye Yuan, Matthew O'Toole, and Kris M Kitani. Optical non-line-of-sight physics-based 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7013–7022, 2020.

Jiazhen Ji, Huan Wang, Yuge Huang, Jiaxiang Wu, Xingkun Xu, Shouhong Ding, ShengChuan Zhang, Liujuan Cao, and Rongrong Ji. Privacy-preserving face recognition with learnable privacy budgets in frequency domain. In *European Conference on Computer Vision*, pp. 475–491. Springer, 2022.

Wu Jiachen, Zhang Hua, Zhang Wenhui, Jin Guofan, Cao Liangcai, and Barbastathis George. Single-shot lensless imaging with fresnel zone aperture and incoherent illumination. *Light: Science & Applications*, 9(53), 2020.

Salman Siddique Khan, Adarsh V R, Vivek Boominathan, Jasper Tan, Ashok Veeraraghavan, and Kaushik Mitra. Towards photorealistic reconstruction of highly multiplexed lensless images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7859–7868, 2019. doi: 10.1109/ICCV.2019.00795.

Xin Lei, Liangyu He, Yixuan Tan, Ken Xingze Wang, Xinggang Wang, Yihan Du, Shanhui Fan, and Zongfu Yu. Direct object recognition without line-of-sight using optical coherence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11737–11746, 2019.

Haiyan Liu, Liheng Bian, and Jun Zhang. Image-free single-pixel segmentation. *Optics Laser Technology*, 157:108600, 2023. ISSN 0030-3992. doi: https://doi.org/10.1016/j.optlastec.2022.108600.

Jiawei Liu, Jing Zhang, and Nick Barnes. Modeling aleatoric uncertainty for camouflaged object detection. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2613–2622, 2022. doi: 10.1109/WACV51458.2022.00267.

Asif M. Salman, Ayremlou Ali, Sankaranarayanan Aswin, Veeraraghavan Ashok, and G. Baraniuk Richard. Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging*, 3(3):384–397, 2017.

Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2014. doi: 10.1109/CVPR.2014.39.

Kristina Monakhova, Kyrollos Yanny, Neerja Aggarwal, and Laura Waller. Spectral DiffuserCam: Lensless snapshot hyperspectral imaging with a spectral filter array. *Optica*, 7(10):1298–1307, Sep 2020.

Antipa Nick, Kuo Grace, Heckel Reinhard, Mildenhall Ben, Bostan Emrah, Ng Ren, and Waller Laura. Diffusercam: lensless single-exposure 3d imaging. *Optica*, 5(1):1–9, 2018.

Xiuxi Pan, Xiao Chen, Tomoya Nakamura, and Masahiro Yamaguchi. Incoherent reconstruction-free object recognition with mask-based lensless optics and the transformer. *Optics Express*, 29 (23):37962, 2021a.

Xiuxi Pan, Tomoya Nakamura, Xiao Chen, and Masahiro Yamaguchi. Lensless inference camera: incoherent object recognition through a thin mask with lbp map generation. *Optics Express*, 29 (7):9758–9771, 2021b.

Xiuxi Pan, Xiao Chen, Saori Takeyama, and Masahiro Yamaguchi. Image reconstruction with Transformer for mask-based lensless imaging. *Optica letter*, 47:1843–1846, 2022.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2150–2160, 2022. doi: 10.1109/CVPR52688.2022.00220.

Fabian Perez, Jhon Lopez, and Henry Arguello. Privacy-preserving deep learning using deformable operators for secure task learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5980–5984. IEEE, 2024.

Chenxi Qiu, Tao Yue, and Xuemei Hu. Reconstruction-free cascaded adaptive compressive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2620–2630, 2024.

Zihuan Qiu, Zhichuan Wang, Miaomiao Zhang, Ziyong Xu, Jie Fan, and Linfeng Xu. Bdg-net: boundary distribution guided network for accurate polyp segmentation. In *Medical Imaging*, 2022.

Khan Salman, Siddique, Sundar Varun, Boominathan Vivek, Veeraraghavan Ashok, and Mitra Kaushik. Flatnet: Towards photorealistic scene reconstruction from lensless measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1934–1948, 2022.

Yue Song, Hao Tang, Nicu Sebe, and Wei Wang. Disentangle saliency detection into cascaded detail modeling and body filling. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(1), jan 2023.

Kazuyuki Tajima, Takeshi Shimano, Yusuke Nakamura, Mayu Sao, and Taku Hoshizawa. Lensless light-field imaging with multi-phased fresnel zone aperture. In *2017 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–7, 2017. doi: 10.1109/ICCPHOT.2017.7951485.

Jasper Tan, Li Niu, Jesse K. Adams, Vivek Boominathan, Jacob T. Robinson, Richard G. Baraniuk, and Ashok Veeraraghavan. Face detection and verification using lensless cameras. *IEEE Transactions on Computational Imaging*, 5(2):180–194, 2019. doi: 10.1109/TCI.2018.2889933.

Feng Tian and Weijian Yang. Learned lensless 3d camera. *Optics Express*, 30(19):34479–34496, Sep 2022.

Boominathan Vivek, K. Adams Jesse, T. Robinson Jacob, and Veeraraghavan Ashok. Phlatcam: Designed phase-mask based thin lensless camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(27):1618–1629, 2020.

Jun Wei, Shuhui Wang, and Qingming Huang. F3Net: Fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pp. 12321–12328, Apr 2020.

Jiachen Wu, Liangcai Cao, and George Barbastathis. DNN-FZA camera: A deep learning approach toward broadband fza lensless imaging. *Optics Letter*, 46(1):130–133, Jan 2021.

Chenxing Xia, Xinyu Chen, Yanguang Sun, Bin Ge, Xianjin Fang, Xiuju Gao, Kuan-Ching Li, Hanling Zhang, and Yan Zhang. Ceminet: Context exploration and multi-level interaction network for salient object detectionimage 1. *Digital Signal Processing*, 147:104403, 2024. ISSN 1051-2004. doi: https://doi.org/10.1016/j.dsp.2024.104403.

Jingyu Yang, Mengxi Zhang, Xiangjun Yin, Kun Li, and Huanjing Yue. Lensless sensing of facial expression by transforming spectral attention features. *IEEE Transactions on Instrumentation and Measurement*, 73:1–13, 2024.

Zhang Yang, Liu Moyun, He Jingwu, Pan Fei, and Guo Yanwen. Affinity fusion graph-based framework for natural image segmentation. *IEEE Transactions on Multimedia*, 24:440–450, 2022. doi: 10.1109/TMM.2021.3053393.

Xiangjun Yin, Huanjing Yue, Mengxi Zhang, Huihui Yue, Xingyu Cui, and Jingyu Yang. Inferring objects from lensless imaging measurements. *IEEE Transactions on Computational Imaging*, 8: 1265–1276, 2022.

Xiangjun Yin, Huanjing Yue, Huihui Yue, Mengxi Zhang, Kun Li, and Jingyu Yang. A multi-task deep learning framework integrating segmentation and reconstruction for lensless imaging. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

Haoran You, Cheng Wan, Yang Zhao, Zhongzhi Yu, Yonggan Fu, Jiayi Yuan, Shang Wu, Shunyao Zhang, Yongan Zhang, Chaojian Li, Vivek Boominathan, Ashok Veeraraghavan, Ziyun Li, and Yingyan Lin. Eyecod: Eye tracking system acceleration via flatcam-based algorithm and accelerator co-design. In *International Symposium on Computer Architecture (ISCA)*, 06 2022.

Yinger Zhang, Zhouyi Wu, Peiying Lin, Yang Pan, Yuting Wu, Liufang Zhang, and Jiangtao Huangfu. Hand gestures recognition in videos taken with a lensless camera. *Optics express*, 30(22):39520–39533, 2022.

A Appendix

A.1 The Derivation Details of The OFD

Combining Eq. (1) with the lensless imaging model in Eq. (2), the above semantic features Y_{θ_i} $(i = 1, 2, 3, 4)$ is modeled as:

$$Y_{\theta_i} = A_{\mathrm{L},\theta_i} X_{\theta_i} A_{\mathrm{R},\theta_i}^{\top} + \xi, \tag{18}$$

where X_{θ_i}, A_{L,θ_i}, and A_{R,θ_i} denote the X, A_{L}, and A_{R} in the feature space. Therefore, the task of reasoning about X_{θ_i} from Y_{θ_i} can be modeled as an inverse problem. To obtain X_{θ_i} for boosting the lensless object segmentation task, inspired by (Salman et al. (2022)), our OFD-based extractor is designed as the Tikhonov regularization problem as:

$$\arg\min_{X_{\theta_i}} \left\| Y_{\theta_i} - A_{\mathrm{L},\theta_i} X_{\theta_i} A_{\mathrm{R},\theta_i}^{\top} \right\|_2^2 + K_{\theta_i} \left\| X_{\theta_i} \right\|_2^2, \tag{19}$$

where K_θ is the learnable regularization parameter.

The Eq. (19) represents a convex optimization problem, implying the existence of a unique minimum, which corresponds to the function value at the point where its derivative equals zero. To solve this, we set the derivative of Eq. (19) to zero, yielding:

$$A_{\mathrm{L},\theta_i}^{\top}(A_{\mathrm{L},\theta_i} X_{\theta_i} A_{\mathrm{R},\theta_i}^{\top} - Y_{\theta_i}) A_{\mathrm{R},\theta_i} + K_{\theta_i} X_{\theta_i} = 0. \tag{20}$$

Expanding the first term and rearranging yields:

$$A_{\mathrm{L},\theta_i}^{\top} A_{\mathrm{L},\theta_i} X_{\theta_i} A_{\mathrm{R},\theta_i}^{\top} A_{\mathrm{R},\theta_i} + K_{\theta_i} X_{\theta_i} = A_{\mathrm{L},\theta_i}^{\top} Y_{\theta_i} A_{\mathrm{R},\theta_i}. \tag{21}$$

Getting the SVD of A_{L,θ_i} and A_{R,θ_i} as:

$$\begin{aligned} A_{\mathrm{L},\theta_i} &=^{\mathrm{SVD}} U_{\mathrm{L},\theta_i} \Sigma_{\mathrm{L},\theta_i} V_{\mathrm{L},\theta_i}^{\top}, \quad A_{\mathrm{R},\theta_i} \overset{\mathrm{SVD}}{=} U_{\mathrm{R},\theta_i} \Sigma_{\mathrm{R},\theta_i} V_{\mathrm{R},\theta_i}^{\top} \\ A_{\mathrm{L},\theta_i}^{\top} &=^{\mathrm{SVD}} \mathrm{V}_{\mathrm{L},\theta_i} \Sigma_{\mathrm{L},\theta_i}^{\top} U_{\mathrm{L},\theta_i}^{\top}, \quad A_{\mathrm{R},\theta_i}^{\top} \overset{\mathrm{SVD}}{=} V_{\mathrm{R},\theta_i} \Sigma_{\mathrm{R},\theta_i}^{\top} U_{\mathrm{R},\theta_i}^{\top}. \end{aligned} \tag{22}$$

Thus we can further obatain:

$$\begin{aligned} A_{\mathrm{L},\theta_i}^{\top} A_{\mathrm{L},\theta_i} &\overset{\mathrm{SVD}}{=} \mathrm{V}_{\mathrm{L},\theta_i} \Sigma_{\mathrm{L},\theta_i}^{\top} U_{\top,\theta_i}^{\top} U_{\mathrm{L},\theta_i} \Sigma_{\mathrm{L},\theta_i} V_{\mathrm{L},\theta_i}^{\top} = \mathrm{V}_{\mathrm{L},\theta_i} \Sigma_{\mathrm{L},\theta_i}^2 V_{\mathrm{L},\theta_i}^{\top} \\ A_{\mathrm{R},\theta_i}^{\top} A_{\mathrm{R},\theta_i} &\overset{\mathrm{SVD}}{=} V_{\mathrm{R},\theta_i} \Sigma_{\mathrm{R},\theta_i}^{\top} U_{\mathrm{R},\theta_i}^{\top} U_{\mathrm{R},\theta_i} \Sigma_{\mathrm{R},\theta_i} V_{\mathrm{R},\theta_i}^{\top} = V_{\mathrm{R},\theta_i} \Sigma_{\mathrm{R},\theta_i}^2 V_{\mathrm{R},\theta_i}^{\top}. \end{aligned} \tag{23}$$

Combining Eq. (23) with Eq. (23), we can obtain:

$$V_{\mathrm{L},\theta_i} \Sigma_{\mathrm{L},\theta_i}^2 V_{\mathrm{L},\theta_i}^{\top} X_{\theta_i} V_{\mathrm{R},\theta_i} \Sigma_{\mathrm{R},\theta_i}^2 V_{\mathrm{R},\theta_i}^{\top} + K_{\theta_i} X_{\theta_i} = V_{\mathrm{L},\theta_i} \Sigma_{\mathrm{L},\theta_i} U_{\mathrm{L},\theta_i}^{\top} Y_{\theta_i} U_{\mathrm{R},\theta_i} \Sigma_{\mathrm{R},\theta_i} V_{\mathrm{R},\theta_i}^{\top}. \tag{24}$$

Multiplying both sides of the Eq. (24) with $V_{\mathrm{L},\theta_i}^{\top}$ from the left and V_{R,θ_i} from the right yields:

$$\Sigma_{\mathrm{L},\theta_i}^2 V_{\mathrm{L},\theta_i}^{\top} X_{\theta_i} V_{\mathrm{R},\theta_i} \Sigma_{\mathrm{R},\theta_i}^2 + K_{\theta_i} V_{\mathrm{L},\theta_i}^{\top} X_{\theta_i} V_{\mathrm{R},\theta_i} = \Sigma_{\mathrm{L},\theta_i} U_{\mathrm{L},\theta_i}^{\top} Y_{\theta_i} U_{\mathrm{R},\theta_i} \Sigma_{\mathrm{R},\theta_i}. \tag{25}$$

Let $\sigma_{\mathrm{L},\theta_i}$ and $\sigma_{\mathrm{R},\theta_i}$ denote the diagonal entries of $\Sigma_{\mathrm{L},\theta_i}^2$ and $\Sigma_{\mathrm{R},\theta_i}^2$, respectively, yields:

$$V_{\mathrm{L},\theta_i}^{\top} X_{\theta_i} V_{\mathrm{R},\theta_i} \odot (\sigma_{\mathrm{L},\theta_i} \sigma_{\mathrm{R},\theta_i}^{\top}) + K_{\theta_i} V_{\mathrm{L},\theta_i}^{\top} X_{\theta_i} V_{\mathrm{R},\theta_i} = \Sigma_{\mathrm{L},\theta_i} U_{\mathrm{L},\theta_i}^{\top} Y_{\theta_i} U_{\mathrm{R},\theta_i} \Sigma_{\mathrm{R},\theta_i}, \tag{26}$$

where $\odot$ denotes element-wise multiplication. We further obtain

$$V_{\mathrm{L},\theta_i}^{\top} X_{\theta_i} V_{\mathrm{R},\theta_i} = (\Sigma_{\mathrm{L},\theta_i} U_{\mathrm{L},\theta_i}^{\top} Y_{\theta_i} U_{\mathrm{R},\theta_i} \Sigma_{\mathrm{R},\theta_i}) ./ (\sigma_{\mathrm{L},\theta_i} \sigma_{\mathrm{R},\theta_i}^{\top} + K_{\theta_i}), \tag{27}$$

where $./$ denotes element-wise division. Therefore, the solution of Eq. (19) is written as

$$\begin{aligned} \hat{X}_{\theta_i} &= \mathrm{OFD}(Y_{\theta_i}; A_{\mathrm{L},\theta_i}, A_{\mathrm{R},\theta_i}) \\ &= V_{\mathrm{L},\theta_i} [(\Sigma_{\mathrm{L},\theta_i} U_{\mathrm{L},\theta_i}^{\top} Y_{\theta_i} U_{\mathrm{R},\theta_i} \Sigma_{\mathrm{R},\theta_i}) ./ (\sigma_{\mathrm{L},\theta_i} \sigma_{\mathrm{R},\theta_i}^{\top} + K_{\theta_i})] V_{\mathrm{R},\theta_i}^{\top}, \end{aligned} \tag{28}$$

where the A_{L,θ_i} and A_{R,θ_i} are updated by $A_{\mathrm{L},\theta_i} = f_{A_{\mathrm{L}}}(A_{\mathrm{L}})$ and $A_{\mathrm{R},\theta_i} = f_{A_{\mathrm{R}}}(A_{\mathrm{R}})$. $f_{A_{\mathrm{L}}}(\cdot)$ and $f_{A_{\mathrm{R}}}(\cdot)$ represent the 3×3 convolution layer + batch normalization (BN) + ReLU + down-sampling operator. We denote each side output as X_{θ_i} for $\hat{X}_{\theta_i}$. Note that A_{L} and A_{R} are primarily associated with the system's system function and do not inherently contain scene-specific information, limiting their semantic content. With simpler convolutional operations for A_{L} and A_{R}, network complexity is reduced while maintaining efficiency. Figure 9 shows the output results at different levels of OFD. As seen, OFD focuses on deriving semantically relevant features, such as object contours, to drive downstream tasks, rather than reconstructing visual details. This approach effectively prevents the leakage of visual information.

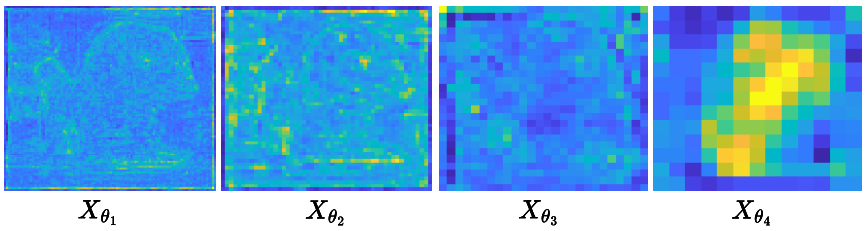

Figure 9: Examples of output results from different levels of OFD. All results are zoomed to the same visualization size for comparison.

A.2 The Details of Datasets

To perform object segmentation tasks for lensless imaging measurements, we construct two datasets named directly captured (DIRC) dataset and display captured (DISC) dataset. DIRC dataset is a testing dataset by directly capturing natural scenes containing 30 images across 10 scenes. DISC dataset is collected from Display-Captured dataset[1] (Khan et al. (2019)) containing 1000 categories of scenarios and the corresponding lensless imaging measurements. By removing unqualified scenes, we obtain 5.9K paired images with 869 categories, which cover flying, aquatic, terrestrial, amphibians, sky, vegetation, and indoor categories. Each category has at least 1 scenario and at most 10 scenarios. The DISC dataset includes 5.2K paired data for training (called DISC-Train) and 0.7K paired data for testing (called DISC-Test). The construction steps of these two datasets are detailed as follows.

First, we use $Eiseg$[2] software (a well-known datasets annotation application) combined with manual refinement to label binary maps I_{mask} for the two datasets.

Next, to perform the multi-task learning strategy with body distribution maps (BDM) I_{bdm} and contour distribution maps (CDM) I_{cdm}, we acquire I_{bdm} and I_{cdm} via Eqs. (9) and (10) of the main manuscript.

Finally, we perform a double-check to ensure the accuracy of the labels, $i.e.$, I_{mask}, I_{bdm}, and I_{cdm}.

Fig. 10 presents some examples showing the reliable annotation of our datasets. Note that the DISC-Train dataset is used to train both our method and the baselines, while the DISC-Test and DIRC datasets are employed for testing to evaluate the performance of each method.

A.3 The Details of Body-Distribution Learning

The body information is critical for determining the overall segmentation effect. Thus, we design a body-distribution learning branch consisting of three contextual exploration (CE) modules and a hierarchical information fusion (HIF) module.

Contextual Exploration (CE). As illustrated in Xia et al. (2024), for the human eye, group receptive fields of different sizes are beneficial for enhancing the perception of tiny areas near the focal point of the retina. Following this strategy, we propose a CE module that simulates the mechanism of the human eye in perceiving external objects to obtain a coarse representation of their bodies. The CE module consists of five branches, denoted as b_k ($k = 1, 2, ..., 5$), as shown in Fig. 11. Except for the first and last branches with only one 1×1 convolution, other branches have four convolutions with a size of 1×1, $1 \times (2k-1)$, $(2k-1) \times 1$, and 3×3. First, the outputs of the first four branches are combined by concatenation, followed by a convolution to adjust the channel number to match that of b_5. Then, the results above are element-wise multiplied with the output of b_5 and then fed into the ReLU activation function to obtain the final features. As shown in Fig. 2, we cascade the CE module at the end of the OFD to get the features C_2, C_3, and C_4, respectively.

Hierarchical Information Fusion (HIF). We aggregate three outputs of the CE modules, $i.e.$, C_2, C_3, and C_4, to refine the object regions embedded. Unlike the way the partial decoder works, the

[1] https://siddiquesalman.github.io/flatnet/
[2] https://github.com/PaddleCV-SIG/EISeg

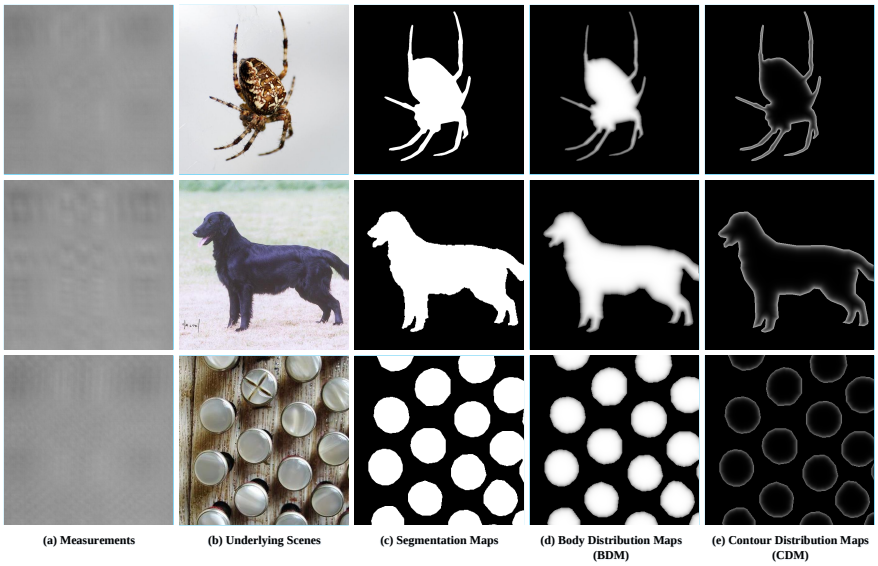

Figure 10: Examples of our dataset. (a)–(e) represent lensless imaging measurements, underlying scenes, ground truth (GT), BDM, and CDM.

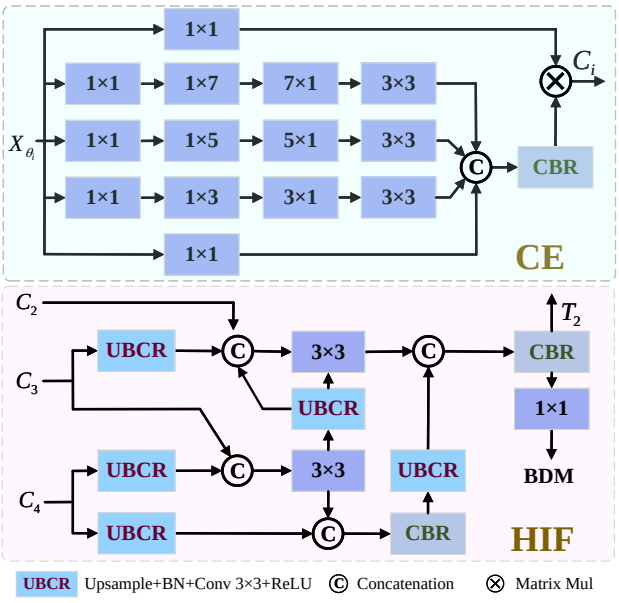

Figure 11: Illustration of CE and HIF.

HIF modifies the skip between different scale features and neighborhood features to sufficiently enhance the bodies of objects and compensate for the details. The detailed structure is shown in Fig. 11. The HIF outputs T_2, fed into one 1×1 convolution layer for obtaining the BDM.

A.4 The Details of Contour-Distribution Learning

The contour information is usually used as a prime cue to refine the object morphology for accurate segmentation. We design the contour-distribution learning branch consisting of the DPA to focus on learning contour information. As shown in Fig. 12, through a 3×3 convolution layer, we first transform the output of the first OFD, *i.e.*, $X_{\theta_1} \in \mathcal{R}^{3\times H\times W}$, into $F_1 \in \mathcal{R}^{C\times H\times W}$, where C,

Table 4: Ablation study on the weight setting of loss functions.

ID	Configuration	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$E_x i \uparrow$	$S_\alpha \uparrow$	mDice ↑	mIoU ↑
#1	$L_{\text{wIOU}} + 0.5 * L_{\text{wBCE}}$	0.898	0.056	0.913	0.872	0.899	0.839
#2	$0.5 * L_{\text{wIOU}} + L_{\text{wBCE}}$	0.893	0.057	0.909	0.864	0.892	0.831
#3	$0.5 * L_{\text{wIOU}} + 0.5 * L_{\text{wBCE}}$	0.899	0.056	0.911	0.873	0.898	0.837
#4	L_{wBCE}	0.867	0.062	0.901	0.837	0.878	0.803
#5	L_{wIOU}	0.824	0.103	0.857	0.804	0.829	0.783
#6	$L_{\text{wIOU}} + L_{\text{wBCE}}$	0.902	0.056	0.916	0.875	0.902	0.841

H, and W are the channel, height, and width of F_1, respectively. Then, we apply global average pooling (GAP) in the spatial dimension of F_1 to calculate channel-wise statistics and a channel downscaling convolution to generate a feature representation. Further, the feature representation is passed through two parallel channel-upscaling convolutions to generate two feature descriptors, *i.e.*, V_1 and V_2, each of dimension is $C \times 1 \times 1$. Moreover, attentional activations W_1 and W_2 after softmax of V_1 and V_2 are generated for calibration and aggregation of F_1. Finally, we add up the two results obtained to get T_1.

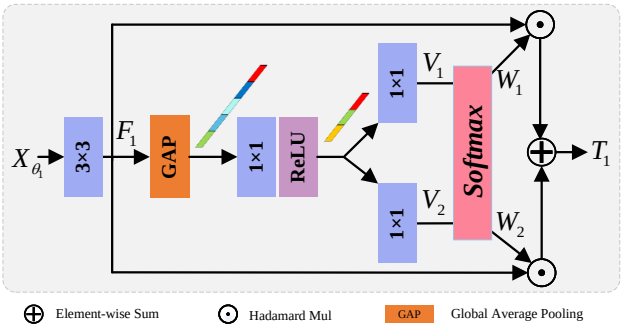

Figure 12: Illustration of the DPA.

A.5 Ablation Study on The Coefficients for Loss Functions

To provide a more comprehensive analysis of our method, we conducted additional experiments on the weight selection of each loss function, building upon the original ablation studies. The quantitative results are presented in Tab. 4. The results demonstrate that variations in performance across different weight configurations are marginal. Given above, we empirically adopted a 1:1 weight ratio.

A.6 The Comparison Results by Our Method and Existing Lensless Segmentation Techniques

For fair evaluation, we also select state-of-the-art method (LLI_T, Raw3dNet, EyeCoD, LOINet, RecSegNet) for comparisons as shown in Fig. 13. Our method demonstrates more precise segmentation results compared to these methods.

A.7 Complexity Analysis

Fig. 14 displays the comparison results for complexity among the aforementioned 9 methods and our FDTDNet, considering parameters (Param), Floating Point Operations (FLOPs), and Frame Per Second (FPS). Our method features 25.87M parameters and 6.82G FLOPs, which are at an intermediate level. Furthermore, our method achieves a frame rate of 35.9 FPS, thereby fulfilling the essential real-time processing requirements. These results highlight that our FDTDNet achieves a favorable balance between performance and complexity.

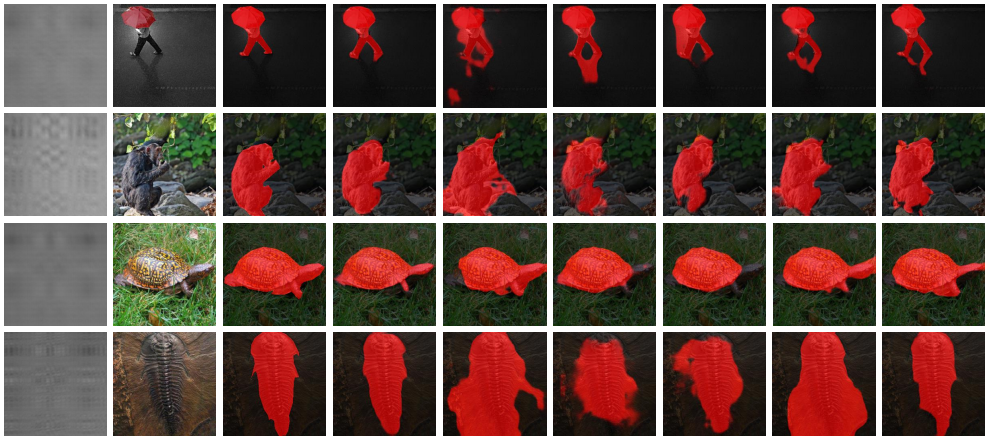

Figure 13: Comparison with state-of-the-art methods on the DISC-Test dataset. The (a) is the lensless measurements corresponding to real images (b); The (c) is the real segmentation maps corresponding to (b); The (d)–(i) are the segmentation results by our FDTDNet, LLI_T, Raw3dNet, EyeCoD, LOINet, and RecSegNet.

A.8 The Comparison Results by Our Method and "Reconstruction + Segmentation" Two-Step Methods

While the focus of this paper is on the architecture and its potential benefits, we acknowledge the importance of comparing our method with traditional reconstruction-based methods. To this end, we employ FlatNet to reconstruct the underlying scene, followed by segmentation using methods such as CDMNet, BDG-Net, and ZoomNet as the "reconstruction + segmentation" two-step methods. Our method, however, retains its original configuration. The comparative results shown in Fig. 15 clearly demonstrate that our method outperforms these state-of-the-art methods in segmentation accuracy.

A.9 Multi-Object Segmentation Results by Comparison Methods and Ours

To further demonstrate the potential of our method in multi-object segmentation, we provide additional visualization results, as shown in Fig. 16. The figure illustrates that while all methods achieve partial segmentation of multiple objects, false positives and missed detections increase as the number of objects grows. Compared to other methods, ours achieves significantly higher segmentation accuracy. However, we acknowledge that there is still room for improvement in multi-object segmentation performance. In future work, we aim to enhance this aspect to increase its practical applicability.

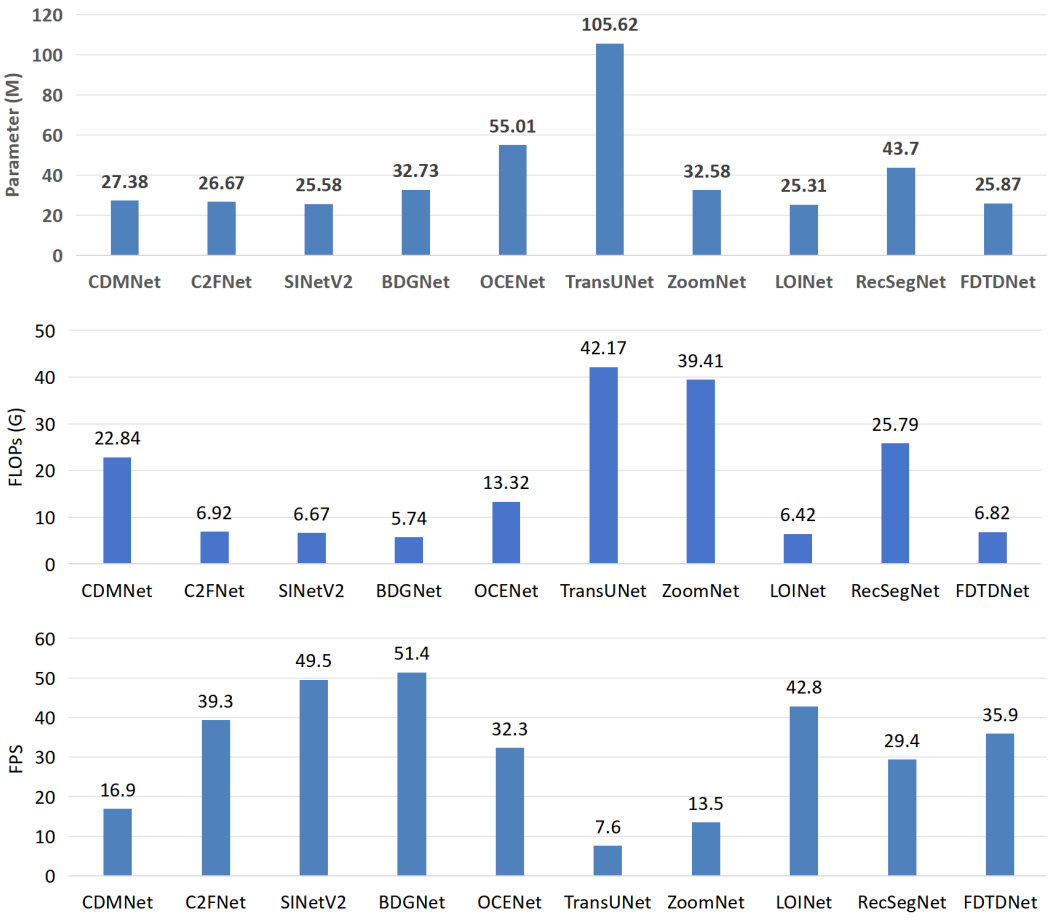

Figure 14: Complexity analysis of our FDTDNet and other state-of-the-art methods in terms of Param, FLOPs and FPS.

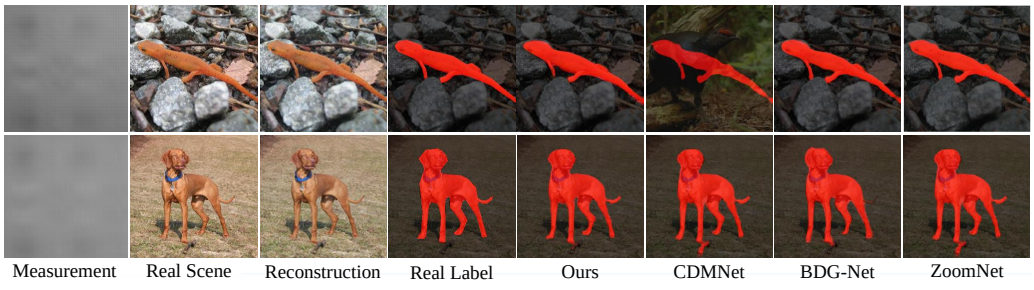

Figure 15: Comparison experiment between the "reconstruction + segmentation" two-step method and ours.

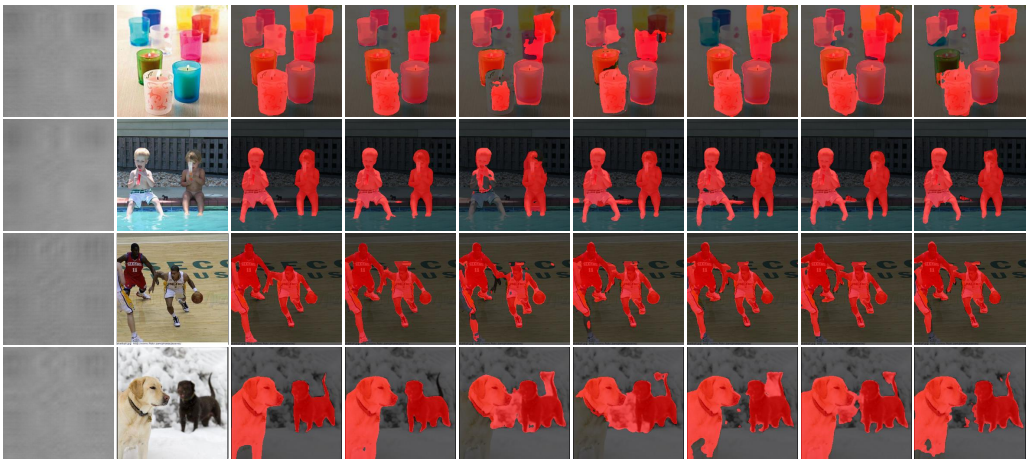

Figure 16: Multi-object segmentation results by comparison methods and ours. The (a) is the lensless measurements corresponding to real images (b); The (c) is the real segmentation maps corresponding to (b); The (d)–(i) are the segmentation results by our FDTDNet, LLI_T, Raw3dNet, EyeCoD, LOINet, and RecSegNet.