D2FP: Learning Implicit Prior for Human Parsing

Junyoung Hong Hyeri Yang Ye ju Kim
Haerim Kim Shinwoong Kim Euna Shim Kyungjae Lee

Yong In University

Abstract

Human parsing aims to segment human images into fine-grained semantic parts. Considering the underlying structure of the human body, state-of-the-art methods typically depend on prior assumptions to represent intrinsic relationships. However, using the same structural prior knowledge across various scenarios poses challenges in achieving consistent prediction and requires additional network design efforts. To address these issues, we introduce a novel approach, the Dynamic Dual Transformer for Parsing (D2FP), which dynamically learns the implicit prior structures of the human body. Specifically, we derive input-dependent prior information from the learnable semantics of human images, generating prior-embedded object queries accordingly before feeding them into the Transformer decoder. Our model includes three major components to effectively learn prior object queries: a prior extraction module, a prior embedding module, and a multi-scale dual Transformer decoder. Furthermore, a novel prior enhancement strategy is introduced, where the final decoded object queries provide clues, enabling enhanced prior features embedding. Experimental results demonstrate the superiority and effectiveness of the proposed method across two well-known human parsing benchmarks: LIP and CIHP. Code and models are available at https://github.com/cvlab-yongin/D2FP.

1. Introduction

The objective of human parsing is to divide images of humans into distinct regions corresponding to anatomical body parts or clothing items. Understanding human instances at the pixel-level is crucial role in several domains spanning human-centric analysis, autonomous driving, and virtual reality. Human parsing can be considered a fine-grained semantic segmentation task. Consequently, numerous studies [4, 23, 24, 32, 48, 51, 53] have been proposed to achieve high-quality human part segmentation by leveraging key properties such as contextual embeddings. However, unlike other dense prediction tasks, human parsing is restricted to segmenting only the human within an image, following a unique physical structure compared to other typical objects. To incorporate prior knowledge of the natural body structure, several approaches [14, 20, 25, 38, 39] employ geometric data structures like trees or graphs to depict inherent human configuration and understand semantic relationships between encapsulated body parts. However, traditional structure-based approaches rely excessively on consistent prior structures across diverse human scenarios. This over-reliance results in obstacles to achieving accurate and stable predictions in certain cases, such as when a specific body part is occluded by another, causing the designed structure to collapse. Additionally, modeling the human structure requires extra architecture design effort and a complex network architecture. In response, we focus on dynamically learning implicit structural information from distinct human images rather than depending on fixed prior assumptions. However, similar to traditional representations, learning relationships embedded in tree or graph structures

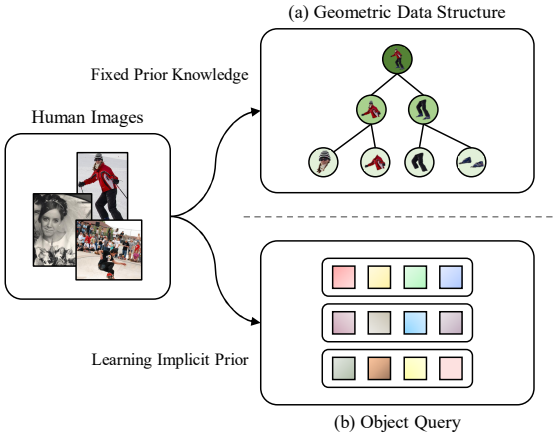

Figure 1. **Learning implicit prior with object query.** Compared to (a) existing methods that adopt complex geometric data structures, (b) our methods learn distinct structural priors from human images and represent body regions with a simple object query representation. These dynamic queries are subsequently employed for mask classification.

presents challenges due to constraints on the static number of nodes and the hurdle of forming a complex topology beforehand.

Recently, object queries in Transformers [1, 8, 9] have demonstrated excellent scalability by representing specific segments within an image as C-dimensional feature vectors in mask classification. Specifically, object queries can encapsulate each body regions without limitations on their number and seamlessly depict complex structural relationships through the self-attention mechanism [13, 27, 33, 36]. From this perspective, M2FP [46] is a seminal approach that introduced query-based Transformer architecture into the human parsing domain with a fixed object query design. They divide object queries into background queries, body part queries, and human instance queries to represent hierarchical relationships.

In this paper, a novel approach, the Dynamic Dual Transformer for Parsing (D2FP), is introduced. The proposed method dynamically captures implicit prior structures from human images and embeds the learned prior designs into object queries. Figure 1 illustrates a visual comparison between traditional methods based on prior assumptions and our methods. Conventional approaches leverage prior information and complex data structures to represent the hierarchical human body. In contrast, we incorporate implicit prior information extracted adaptively from human images with a simple object query schema. To learn effective prior-embedded object query, the proposed network consists of three key modules: 1) a prior extraction module, 2) a prior embedding module, and 3) a multi-scale dual Transformer decoder. Additionally, we propose a simple yet powerful prior enhancement strategy to obtain refined prior features in proposed embedding module. Our method can be viewed as a dynamic query generation approaches since it utilizes input-dependent object queries rather than randomly initialized ones.

Quantitative and qualitative experiments were conducted on the public human parsing benchmarks: the LIP [16] and CIHP [15] datasets, to validate the proposed method. As a result, our method demonstrated superior performance compared to previous state-of-the-art approaches. Furthermore, we demonstrated the effectiveness of our method through an ablation study and visual analysis. Our major contributions can be summarized as follows:

- The introduction of the Dynamic Dual Transformer for Parsing (D2FP), which adapts object queries via robust structural prior learning.
- A novel prior enhancement strategy that refines prior features with simple yet effective implementation.
- Demonstrated superiority of our method on the LIP and CIHP datasets, underpinned by comprehensive experimental validation and visual analysis.

2. Related Work

Human Parsing. Human parsing involves understanding human images at the pixel-level. Traditional methods leverage hand-crafted features [7, 26, 37, 44, 45], human keypoints [26,44,45], and human configuration [3,11,12,17,34, 40, 42] to partition anatomical body parts and clothes. The advancements in Convolutional Neural Networks (CNNs) have been pivotal in the progress of the human parsing domain. Liang *et al.* [24] introduced a Co-CNN network that aggregates global and local context information into a unified architecture. Ruan *et al.* [32] proposed the CE2P framework, investigating useful factors for human part segmentation. Li *et al.* [23] designed a novel self-correction training strategy to refine the noisy labels. Zhang *et al.* [51] generated adaptive context features for various human appearances. To represent the inherent structure of human body, numerous geometrical structure-based methods have been proposed. Gong *et al.* [14] introduced a graph-based parsing network to represent semantic coherency via graph transfer learning. Ji *et al.* [20] proposed a novel semantic neural tree to encapsulate distinct body parts and encode the hierarchical structure of the human body. Wang *et al.* [38] introduced a novel neural information fusion approach for tree-based hierarchical human structure. Wang *et al.* [39] defined three kinds of relations: decomposition, composition, and dependency based on graph neural network. Gong *et al.* [16] proposed structure-sensitive learning to guide human joint awareness. Liu *et al.* [25] established the horizontal and vertical class distribution to represent the structural knowledge of human body. Yang *et al.* [46] proposed Transformer-based human parsing framework and designed group object queries for hierarchical human representation.

Object Query in Transformers. Object queries are considered as slots that represent specific object regions. They learn distinct modes through an iterative decoding process during training and operate based on their implicit roles to generate final predictions. The object query scheme was first proposed in DETR [1], which considers object detection as a set prediction task. For explicit physical object query embeddings, Wang *et al.* [41] proposed Anchor DETR to ensure each query focuses on objects near the corresponding anchor points. Meng *et al.* [31] proposed a spatial object query from the reference point. Zhang *et al.* [49] demonstrated higher accuracy by transforming conventional sparse queries into a dense and distinct set. Cheng *et al.* [9] defined mask classification using object queries, representing each corresponding segment. For object query enhancement, Chen *et al.* [2] introduced a selective query recollection strategy, which accumulates meaningful queries and recollects them in the subsequent Transformer decoder. Cui *et al.* [10] introduced dynamically combining learned coefficients from input images with object queries.

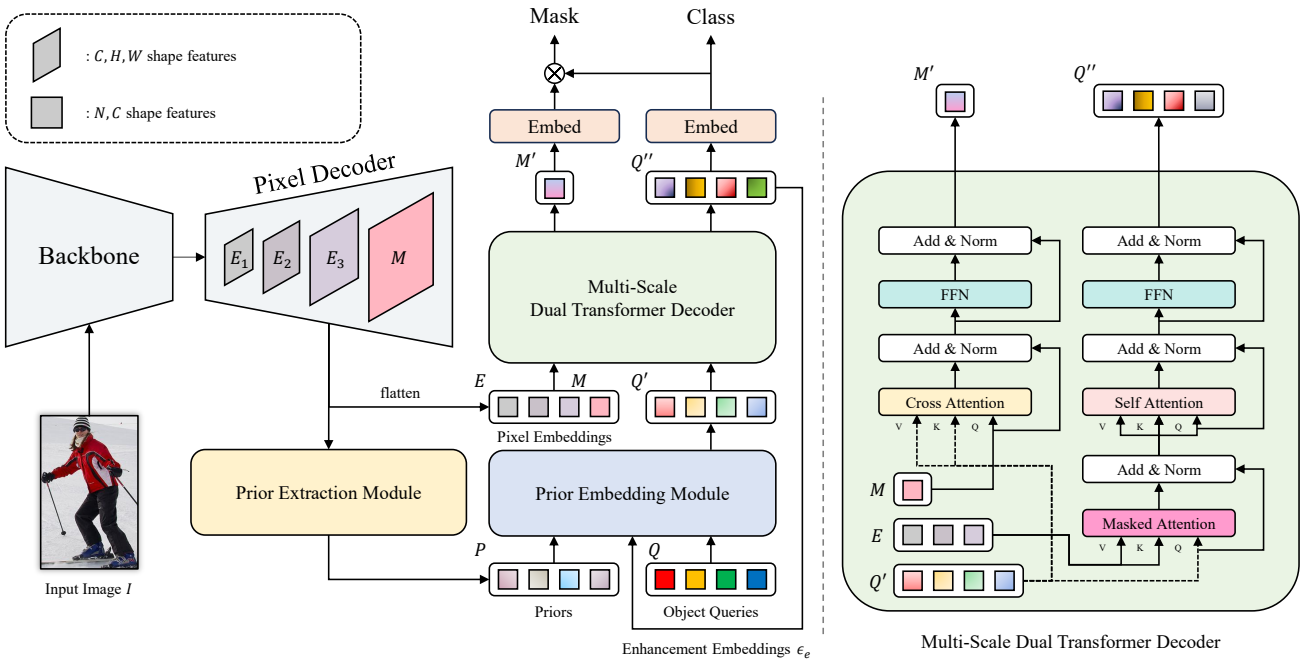

Figure 2. **The overall architecture of D2FP.** The prior features are extracted based on the high-level context captured from the input images. The prior embedding module generates the prior object queries with a prior enhancement strategy. In multi-scale dual Transformer decoder, the mask features and prior object queries are jointly optimized. The dotted lines within the multi-scale dual Transformer decoder depict the path of the prior object queries.

3. Approach

3.1. Overall Architecture

Figure 2 illustrates the overall architecture of the proposed Dynamic Dual Transformer for Parsing (D2FP). Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, the backbone network extracts three low-resolution features. The pixel decoder processes projected backbone features and generates refined multi-scale pixel embeddings: D_1, D_2, and D_3. The high-resolution mask features M are obtained by the element-wise addition of upsampled D_3 and the highest-resolution backbone features. To effectively learn the implicit prior design, our proposed method consists of three major components: a prior extraction module, a prior embedding module, and a multi-scale dual Transformer decoder. Specifically, contextual prior features $\mathbf{P} \in \mathbb{R}^{N \times C}$, where N and C represent the number of feature vectors and the dimension of each vector, respectively, are extracted from decoded multi-scale feature maps and mask features. Given the prior features, the prior embedding module generates prior object queries $\mathbf{Q} \in \mathbb{R}^{N \times C}$ with initial object queries $\mathbf{Q}_{init} \in \mathbb{R}^{N \times C}$ and enhancement embeddings $\mathbf{E} \in \mathbb{R}^{N \times C}$. In the dual Transformer decoder, we co-optimize mask features M and the prior object queries Q with flattened multi-scale pixel embeddings. Given updated mask features and decoded object queries, we calculate the final segmentation mask and class prediction via embedding layers.

3.2. Prior Extraction Module

To effectively capture contextual priors, we extract learnable C-dimensional prior features. Inspired by object-contextual representations [48], we introduce a multi-scale fusion-based prior extraction module. Figure 3 depicts an illustration of the proposed prior extraction module. We leverage decoded multi-scale pixel embeddings D and high-resolution mask features M. Specifically, we compress concatenated multi-scale features D with a simple convolutional neural network. After that, we generate learnable object regions and obtain flattened mask features. We compute spatial attention maps $\mathbf{A} \in \mathbb{R}^{N \times HW}$ of learnable object regions with a pixel-wise softmax function. Final prior features P are generated through matrix multiplication between the attention maps and the flattened mask features. The overall process is defined as follows:

$$\begin{aligned} \mathbf{A} &= \mathrm{softmax}(\mathrm{conv}(\mathrm{concat}(\mathbf{D}))), \\ \mathbf{P} &= \mathbf{A} \cdot \mathbf{M}^{\mathrm{T}}. \end{aligned} \tag{1}$$

Unlike the previous method [48], we do not require additional supervision for the learnable regions. Specifically, we utilize decoded pixel embeddings, which contain relatively semantic information, instead of low-level backbone features, allowing for a certain degree of appropriateness in soft region prediction.

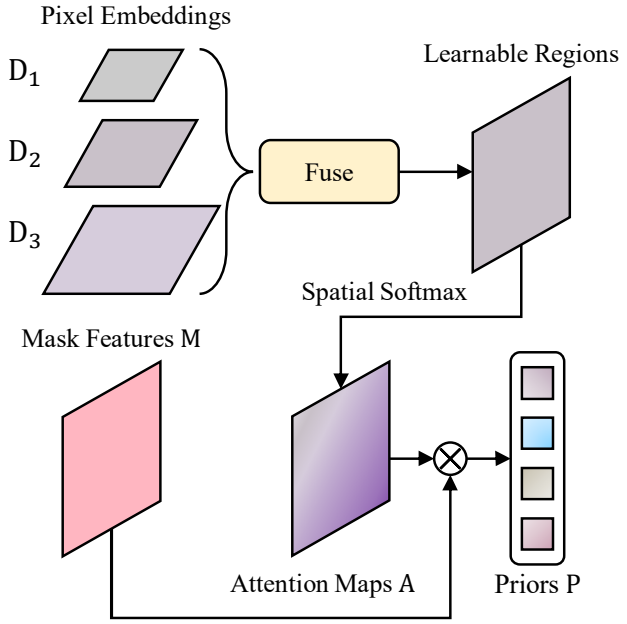

Figure 3. **The prior extraction module.** We fuse multi-scale pixel embeddings E and generate learnable spatial regions. The fusion module is a simple convolutional neural network consisting of convolutional layers, normalization layers, and activation layers. We generate the attention maps A with spatial-aware softmax function to the learnable regions. The prior features are calculated through matrix multiplication between the attention maps and the mask features.

3.3. Prior Embedding Module

We introduce the prior embedding module to embed the extracted adaptive prior features into the object queries. Figure 4 illustrates the proposed prior embedding module, which consists of three iterative attention modules. To ensure sufficient interaction between the priors and queries, we adopt a slot attention mechanism [28]. The prior embedding module inputs obtained prior features P, randomly initialized object queries Q_{init}, and enhancement embeddings E. Details on the prior enhancement strategy are discussed in Section 3.4. First, the prior features are enhanced based on the enhancement embedding through a repetitive attention process. As in conventional methods [8, 9], object queries and positional embeddings $\mathbf{E}_{pos} \in \mathbb{R}^{N\times C}$ are randomly initialized with a Gaussian distribution. Second, a similarity matrix between the object queries Q_{init} and refined priors $\mathbf{P}_{enh} \in \mathbb{R}^{N\times C}$ is obtained, and a hidden query set $\mathbf{Q}_{hidd} \in \mathbb{R}^{N\times C}$ is calculated. The hidden object queries and positional embeddings are combined through element-wise addition. Third, the positional-aware object queries are updated based on the enhanced prior features to generate the final prior object queries Q. By updating object queries in two stages, we enable the attention mechanism to focus more effectively on the intrinsic features of the object queries and refined priors.

3.4. Prior Enhancement Strategy

We propose a novel prior enhancement strategy to retrieve robust prior embeddings for effectively capturing implicit structures. The enhancement embedding is calculated by element-wise summation of the learnable embedding vectors $\epsilon \in \mathbb{R}^{N\times C}$ initialized from a Gaussian distribution and the final decoded object queries Q''. We assume that the decoded object queries has somewhat grasped the characteristics or positional information of specific objects in the image through rich interaction with pixel features in the Transformer decoder. Therefore, we efficiently adjust the information flow of the decoded object queries through learnable embedding vectors. Then, in the first slot attention module, the enhancement embedding improves the prior features by serving as useful clues. Figure 4 illustrates the prior enhancement strategy in prior embedding module. The prior object queries $Q_t^{'}$ at time step t pass through the dual Transformer to generate the final decoded object query embeddings Q''. The final query set is added with learnable embedding vectors ϵ, with gradient updates detached for optimization, to calculate the enhancement embedding ϵ_m. Subsequently, the prior queries $Q_{t+1}^{'}$ are similarly used for the next enhancement embedding.

3.5. Multi-Scale Dual Transformer Decoder

The segmentation mask is predicted through the refined pixel embeddings obtained from the pixel decoder. Therefore, the pixel decoder plays a significant role, and even if the prior object queries are dynamically generated, their impact on the final set prediction may be relatively limited. Following the previous study [18], we introduce a multi-scale dual Transformer decoder that complementarily updates object queries and mask features. The multi-scale dual Transformer decoder processes prior object queries Q', mask features M, and multi-scale decoder features E. Specifically, through the cross-attention layer and feed forward network, the mask features are updated based on the obtained prior queries. After that, the masked attention [8] refines object queries from the multi-scale pixel embeddings, generating the final decoded object queries Q''. For the final prediction, we use a multi-layer perceptron for mask embedding and a linear transformation layer for class embedding and enhanced mask features. The final segmentation mask is predicted through matrix multiplication between the decoded object queries and the refined mask features.

3.6. Implementation Details

The model is implemented using Detectron2 [43]. Settings from previous work [8] are adopted, utilizing a pre-trained ResNet [19] as the backbone network and the Multi-Scale Deformable Attention Transformer (MSDeformAttn) [55] for the pixel decoder. The Hungarian loss [1]

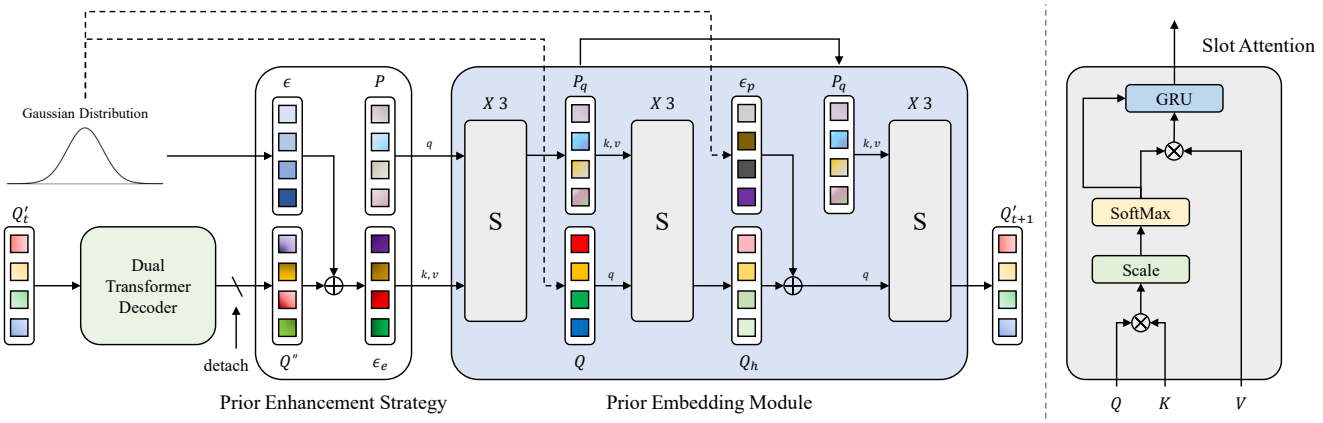

Figure 4. **The prior embedding module with prior enhancement strategy.** The prior embedding module consists of three sub-attention modules. Through each iterative attention mechanism, the final set of prior object queries Q' is generated. For the prior enhancement strategy, the decoded object queries at time t are detached from the gradient update. The dotted lines depict the path of the query-level prior features. The figure on the right depicts a schematic of the slot attention module.

is used to determine optimal bipartite matching between the ground truth set and the prediction set.

Prior Extraction Module. We resize pixel embeddings using bilinear interpolation before channel-wise concatenation. We use a batch normalization layer and the ReLU activation function in the fusion module. The number of priors is set to 100, which is identical to the number of queries.

Prior Embedding Module. We adopt slot attention [28] and the number of iterations is set to 3. We do not use μ and σ for slot initialization. The dimension of the additional multilayer perceptron is set to 384.

Multi-Scale Dual Transformer Decoder. The proposed multi-scale dual Transformer decoder uses a total of 15 layers. Specifically, 6 layers are used for updating the mask features and 9 layers are used to update the object queries. Auxiliary loss is added to all dual Transformer decoder layers and to the initial object queries before the dual Transformer decoder. The dimension of the refined mask feature embedding layer is set to 256.

3.7. Network Training

All models are trained using two NVIDIA GeForce RTX 4090 GPUs. For training, samples from the LIP dataset were resized to a resolution of 384×512, while those from the CIHP dataset were adjusted to 800×800. We train the model for 150 epochs using the AdamW [29] optimizer. For the LIP dataset, we set the mini-batch size to 16 and the initial learning rate to 0.0002. For the CIHP dataset, we set the mini-batch size to 4 and the initial learning rate to 0.0001. We use a warm-up poly learning rate scheduler, with weight decay set to 0.05. We apply extensive jittering in the range of [0.1, 2.0]. Data augmentation techniques such as fixed random cropping, rotation, color jittering, and horizontal flipping are used. For fair comparison, we use flipping and multi-scale test-time augmentation strategy.

4. Experiments

4.1. Datasets

LIP. The Look Into Person (LIP) dataset [16] is one of the most well-known single human parsing datasets, consisting of 30,462 training images and 10,000 validation images. It provides pixel-level annotations for a total of 19 semantic human parts, including 6 body parts and 13 clothing items, as well as one background class.

CIHP. The Crowd Instance-level Human Parsing (CIHP) dataset [15] is a large-scale multiple human parsing dataset, consisting of 38,280 real-world images with pixel-level annotations for 20 categories. The dataset is split into 28,280 training images, 5,000 validation images, and 5,000 test images, with all samples containing at least two human instances.

4.2. Quantitative Results

The performance of our method was quantitatively evaluated and compared with state-of-the-art approaches. Quantitative experiments were performed on the LIP [16] validation set and the CIHP [15] validation set.

LIP. In Table 1, we evaluate the per-class intersection over union of the proposed method on the LIP validation set. Our method demonstrates superior performance compared to state-of-the-art approaches across the majority of classes. In particular, it recorded notable IoU for small classes such as glasses, gloves, and socks compared to existing methods. Additionally, it demonstrated segmentation performance gain in handling clothing items such as dresses and

Table 1. Quantitative per-class comparison of mIoU on the LIP validation set. The **bold** and underline denote the best and second-best performances, respectively.

Method	hat	hair	glove	glasses	u-cloth	dress	coat	sock	pants	j-suits	scarf	skirt	face	l-arm	r-arm	l-leg	r-leg	l-shoe	r-shoe	bkg	Avg
Attention [5]	58.87	66.78	23.32	19.48	63.20	29.63	49.70	35.23	66.04	24.73	12.84	20.41	70.58	50.17	54.03	38.35	37.70	26.20	27.09	84.00	42.92
DeepLab [4]	56.48	65.33	29.98	19.67	62.44	30.33	51.03	40.51	69.00	22.38	11.29	20.56	70.11	49.25	52.88	42.37	35.78	33.81	32.89	84.53	44.03
PSPNet [53]	63.50	68.00	39.10	23.80	68.10	31.70	56.20	44.50	72.70	28.70	15.70	25.70	70.80	59.70	62.30	54.90	54.50	42.30	42.90	86.10	50.60
MMAN [30]	57.66	65.63	30.07	20.02	64.15	28.39	51.98	41.46	71.03	23.61	9.65	23.20	69.54	55.30	58.13	51.90	52.17	38.58	39.05	84.75	46.81
SS-NAN [54]	63.86	70.12	30.63	23.92	70.27	33.51	56.75	40.18	72.19	27.68	16.98	26.41	75.33	55.24	58.93	44.01	41.87	29.15	32.64	88.67	47.92
JPPNet [16]	63.55	70.20	36.16	23.48	68.15	31.42	55.65	44.56	72.19	28.39	18.76	25.14	73.36	61.97	63.88	58.21	57.99	44.02	44.09	86.26	51.37
CE2P [32]	65.29	72.54	39.09	32.73	69.46	32.52	56.28	49.67	74.11	27.23	14.19	22.51	75.50	65.14	66.59	60.10	58.59	46.63	46.12	87.67	53.10
SNT [20]	66.90	72.20	42.70	32.30	70.10	33.80	57.50	48.90	75.20	32.50	19.40	27.40	74.90	65.80	68.10	60.03	59.80	47.60	48.10	88.20	54.70
CorrPM [52]	66.20	71.56	41.06	31.09	70.20	37.74	57.95	48.40	75.19	32.37	23.79	29.23	74.36	66.53	68.61	62.80	62.81	49.03	49.82	87.77	55.33
SCHP [23]	69.96	73.55	50.46	40.72	69.93	39.02	57.45	54.27	76.01	32.88	26.29	31.68	76.19	68.65	70.92	67.28	66.56	55.76	56.50	88.36	58.62
CDGNet [25]	71.06	74.61	50.13	42.09	71.58	40.00	58.73	55.25	77.92	34.32	30.05	32.97	77.12	71.25	73.35	**70.54**	**69.26**	**58.24**	**58.75**	88.86	60.30
M2FP [46]	71.03	**74.80**	50.18	40.86	71.99	41.56	59.39	56.39	77.90	28.96	27.49	**34.51**	**77.28**	**72.28**	**73.43**	68.79	68.56	57.03	57.82	**89.39**	59.98
Ours	**71.14**	**74.80**	**52.00**	**42.90**	**72.07**	**42.40**	**59.50**	**56.69**	**78.15**	**34.56**	**30.39**	32.27	76.89	71.88	73.42	69.87	68.79	56.91	57.56	89.32	**60.58**

coats, surpassing conventional approaches. Table 2 presents the quantitative evaluation of our method compared to state-of-the-art methods in terms of pixel accuracy, mean accuracy, and mIoU. Our method achieved a state-of-the-art performance of 60.58% mIoU on the LIP dataset. Compared to methods that utilize the inherent structure of humans, such as SNT [20] and CDGNet [25], our method achieved 5.85% and 0.28% higher mIoU scores, respectively. Notably, our D2FP outperforms SOLIDER [6], a human-centric analysis method that requires human prior knowledge for semantic parsing. Furthermore, we achieved 2.03% higher performance in mean accuracy compared to state-of-the-art CDGNet [25].

CIHP. In Table 3, we evaluate the mIoU of the proposed method on the CIHP validation set. Although our method showed a slight weakness of 0.05% compared to CDGNet [25], it demonstrated improvements of 6.93% and 4.64% compared to the graph-based Graphonomy [14] and tree-based SNT [20], respectively, which rely on human prior knowledge. The quantitative experimental results denote the superiority of adaptively extracting implicit structural prior based on the human images, rather than relying on predefined relationships, in multiple human parsing scenario.

4.3. Qualitative Results

As illustrated in Figure 5, our method is qualitatively compared with state-of-the-art approaches on the LIP validation set. Compared to M2FP [46], the proposed method captures implicit prior from the input images, generating high-quality segmentation masks even when critical parts of the body are missing or occluded by certain objects. For example, in the second row, where the lower body is occluded, and in the third row, where the hands and arms are occluded by an object, our method qualitatively outperforms fixed-query design method. Additionally, in the fifth row, previous methods misclassified the upper body clothing due to the complex pose and fine details, while our method shows reasonable mask predictions even for challenging scenarios. Notably, our method predicts more accurate segmentation masks in the sixth row compared to SCHP [23], which uses a ground truth self-correction strategy, and even the ground truth itself.

Table 2. Quantitative comparison of pixel accuracy, mean accuracy, and mIoU on the LIP validation set. The **bold** and underline denote the best and second-best performances, respectively.

Method	Pixel Acc.	Mean Acc.	mIoU
Attention [5]	83.43	54.39	42.92
JPPNet [16]	86.39	62.32	51.37
CE2P [32]	87.37	63.20	53.10
CNIF [38]	88.03	68.80	57.74
SNT [20]	88.05	66.42	54.73
CorrPM [52]	87.68	67.21	55.33
BGNet [50]	-	-	56.82
ISNet [22]	-	-	56.96
MCIBISS [21]	-	-	56.99
PCNet [51]	-	-	57.03
HHP [39]	**89.05**	70.58	59.25
CDGNet [25]	88.86	71.49	60.30
SOLIDER [6]	-	-	60.50
Ours	**89.05**	**73.52**	**60.58**

Table 3. Quantitative comparison of mIoU on the CIHP validation set. The **bold** and underline denote the best and second-best performances, respectively.

Method	Backbone	mIoU
PGN [15]	DeepLabV2	55.80
Graphonomy [14]	DeepLabV3+	58.58
CE2P [32]	ResNet101	59.50
Parsing R-CNN [47]	ResNet50	56.30
CorrPM [52]	ResNet101	60.18
SNT [20]	ResNet101	60.87
PCNet [51]	ResNet101	61.05
CDGNet [25]	ResNet101	**65.56**
Ours	ResNet101	65.51

Figure 5. **Qualitative comparison of different methods on the LIP validation set.** Our method demonstrates visually superior segmentation masks compared to previous state-of-the-art methods. Zoom in to see details.

4.4. Analysis

Ablation study. All experiments are conducted on the LIP validation set to demonstrate the effectiveness of each key component of the network design (See Table 4a). Method (a) optimizes the mask features additionally, leading to impressive results compared to existing approaches [46]. Method (b) demonstrates the results of combining the prior extraction module and prior embedding module with the previous architectures [8, 46]. Method (c) shows the results of adopting a dual Transformer decoder for input-dependent object query design and demonstrates superior performance compared to Method (b). Specifically, dual Transformer decoder reduces dependency on the pixel decoder, allowing the dynamically generated object queries to play a significant role in the final mask predictions. The quantitative experimental results with and without enhancement embedding are shown in Methods (c) and (d), respectively. The prior enhancement strategy demonstrated significant performance improvements in mIoU and mean accuracy by effectively enhancing structural priors based on final decoded queries during the embedding process. Consequently, the approach that incorporates all three components showed the best quantitative results compared other network designs.

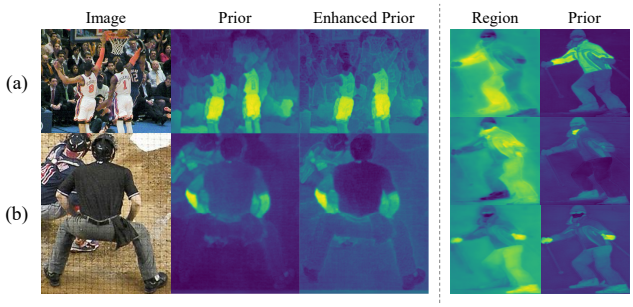

Figure 6. **Visualization of regions, priors, and enhanced priors.** (Left) The enhanced priors incorporate detailed edges across the global region and effectively condition the object queries. (Right) The generated priors extracted from learnable regions encapsulate specific body or clothing parts in the image. Zoom in to see details.

Prior enhancement strategy. The goal of the prior enhancement strategy is to enhance the extracted priors using the decoded query set as clues. Figure 6 (Left) qualitatively demonstrates the validity of the proposed strategy. Specifically, in case (a), the enhanced prior provides highly detailed reconstructions of scenes and the other objects (e.g., the audience and basketball hoop) in complex environments. Case (b) also illustrates the effectiveness of the proposed method in embedding the structure of the lower body of a human. It demonstrates that the strategy can capture the relationship between anatomical details and high-level context, refining structural understanding through prior enhancement.

Computational analysis. In Table 4b, we compare the number of model parameters and inference speed. Note that the results are obtained at 473p resolution on the LIP [16] validation set. Our method requires a large number of parameters and demonstrates slower inference compared to existing methods [23, 46], as it relies on a dual Transformer decoder and a prior extraction module to effectively learn implicit structure. However, the extracted priors generate a robust set of object queries, ultimately leading to superior segmentation performance compared to state-of-the-art methods.

Visualization of regions and priors. Figure 6 (Right) shows the visualization of learnable regions and prior features. From the visual experimental results, we can identify two notable points. First, learnable regions can represent semantic areas without auxiliary supervision [48], and the extracted priors can effectively encapsulate specific seg-

Table 4. D2FP ablations. All experiments were measured on the LIP validation set. **bold** and underline denote the best and second-best performances, respectively.

(a) Quantitative comparison of different combinations.

Method	Prior Query	Enhancement Embedding	Dual Decoder	Pixel Acc.	mIoU	Mean Acc.	fwIoU
(a)	-	-	✓	89.02	60.05	73.30	81.03
(b)	✓	-	-	89.01	59.96	72.69	80.99
(c)	✓	-	✓	**89.12**	60.00	72.94	**81.17**
(d)	✓	✓	✓	89.05	**60.58**	**73.52**	81.06

(b) Computational analysis.

Method	Params #	Times (s)	mIoU
SCHP [23]	66.7M	0.0295	58.62
M2FP [46]	63.0M	0.0906	59.98
Ours	79.7M	0.1040	**60.58**

(c) Number of priors ablation.

	mIoU	fwIoU	mAcc	pAcc
50	60.28	80.90	73.16	88.96
100	**60.58**	**81.06**	**73.52**	**89.05**
200	60.43	81.02	73.41	89.00

(d) Number of decoders ablation.

	mIoU	fwIoU	mAcc	pAcc
3	60.07	80.97	72.94	89.02
6	60.24	81.01	73.07	89.02
9	**60.58**	**81.06**	**73.52**	**89.05**

ments of the image. Second, representing a single body region does not necessarily require adherence to any hierarchical or inherent body structure, and accurate parsing can be achieved with learned implicit structural prior. Specifically, Prior in first row predicts upper body clothing, as seen in same row Region, where the left part of the human is semantically activated. In conclusion, by combining prior features that capture unique relationships with a simple object query representation, our method can achieve high-quality prediction in human parsing.

Visualization of prior object queries. One of the most important differences between the proposed method and existing object query-based methods [1,8,9,55] is that the object queries are conditioned on distinct samples. In Figure 7, we visualize the dynamically generated prior object queries using 200 randomly selected samples in the LIP dataset [16]. Specifically, we flattened the object queries and projected them onto a 2-dimensional space using t-SNE [35]. Notably, the object queries for similar scenes are semantically clustered. For example, we can observe that samples where other physiological body parts are occluded and primarily hands or feet are visible tend to have similar similarity. Additionally, scenarios with sports activities involving skis or complex crowd scenes, which share identical semantics, are also clustered together. The prior learning enables object queries to capture structural relationships based on high-level semantics from diverse human samples, extending beyond the internal structure of humans. It indicates that the priors implicitly learn the underlying structure, providing the Transformer decoder with a query set tailored to each specific scene.

Hyperparameter sensitivity. We performed quantitative experiments focusing on two key hyperparameters: the number of priors and Transformer decoders. As shown in Table 4c, the model achieved the best performance across all metrics when the number of priors was set to 100. Table 4d demonstrates that the performance improves in proportion to the number of Transformer decoders.

Figure 7. **t-SNE visualization of object queries.** Samples with similar meanings have the same similarity score in the query set. Zoom in to see details.

5. Conclusion

Conventional methods primarily depend on fixed prior knowledge about the intrinsic structure of the human body. However, using the identical structural design presents challenges in adapting diverse human scenarios. To address these issues, the Dynamic Dual Transformer for Parsing (D2FP) is proposed, which dynamically learns implicit prior structures based on the learnable semantics of human images, providing prior object queries to dual Transformer decoder. Our network includes three main components designed for effective prior learning: a prior extraction module, a prior embedding module, and a multi-scale dual Transformer decoder. Additionally, we introduce a novel prior enhancement strategy to refine the extracted prior features with straightforward implementation. Our approach outperformed state-of-the-art methods in both quantitative and qualitative experiments on two public datasets, supported by extensive ablation studies and visual analysis.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. ii, iv, viii

[2] Fangyi Chen, Han Zhang, Kai Hu, Yu-kai Huang, Chenchen Zhu, and Marios Savvides. Enhanced training of query-based object detection via selective query recollection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23756–23765, 2023. ii

[3] Hong Chen, Zi Jian Xu, Zi Qiang Liu, and Song Chun Zhu. Composite templates for cloth modeling and sketching. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 943–950. IEEE, 2006. ii

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. i, vi

[5] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016. vi

[6] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15050–15061, 2023. vi

[7] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1978, 2014. ii

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. ii, iv, vii, viii

[9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021. ii, iv, viii

[10] Yiming Cui, Linjie Yang, and Haichao Yu. Dq-det: Learning dynamic query combinations for transformer-based object detection and segmentation. *arXiv preprint arXiv:2307.12239*, 2023. ii

[11] Jian Dong, Qiang Chen, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Towards unified human parsing and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 843–850, 2014. ii

[12] Jian Dong, Qiang Chen, Wei Xia, Zhongyang Huang, and Shuicheng Yan. A deformable mixture parsing model with parselets. In *Proceedings of the IEEE international conference on computer vision*, pages 3408–3415, 2013. ii

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. ii

[14] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7450–7459, 2019. i, ii, vi

[15] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 770–785, 2018. ii, v, vi

[16] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 932–940, 2017. ii, v, vi, vii, viii

[17] Jianyang Gu, Kai Wang, Hao Luo, Chen Chen, Wei Jiang, Yuqiang Fang, Shanghang Zhang, Yang You, and Jian Zhao. Msinet: Twins contrastive search of multi-scale interaction for object reid, 2023. ii

[18] Junjie He, Pengyu Li, Yifeng Geng, and Xuansong Xie. Fastinst: A simple query-based model for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23663–23672, 2023. iv

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. iv

[20] Ruyi Ji, Dawei Du, Libo Zhang, Longyin Wen, Yanjun Wu, Chen Zhao, Feiyue Huang, and Siwei Lyu. Learning semantic neural tree for human parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 205–221. Springer, 2020. i, ii, vi

[21] Zhenchao Jin, Tao Gong, Dongdong Yu, Qi Chu, Jian Wang, Changhu Wang, and Jie Shao. Mining contextual information beyond image for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7231–7241, 2021. vi

[22] Zhenchao Jin, Bin Liu, Qi Chu, and Nenghai Yu. Isnet: Integrate image-level and semantic-level context for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7189–7198, 2021. vi

[23] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020. i, ii, vi, vii, viii

[24] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *Proceedings of the IEEE international conference on computer vision*, pages 1386–1394, 2015. i, ii

[25] Kunliang Liu, Ouk Choi, Jianming Wang, and Wonjun Hwang. Cdgnet: Class distribution guided network for human parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4473–4482, 2022. i, ii, vi

[26] Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia*, 16(1):253–265, 2013. ii

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. ii

[28] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020. iv, v

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. v

[30] Yawei Luo, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Macro-micro adversarial network for human parsing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. vi

[31] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3651–3660, 2021. ii

[32] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4814–4821, 2019. i, ii, vi

[33] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. ii

[34] Xiaoguang Tu, Yingtian Zou, Jian Zhao, Wenjie Ai, Jian Dong, Yuan Yao, Zhikang Wang, Guodong Guo, Zhifeng Li, Wei Liu, et al. Image-to-video generation via 3d facial dynamics. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):1805–1819, 2021. ii

[35] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. viii

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. ii

[37] Nan Wang and Haizhou Ai. Who blocks who: Simultaneous clothing segmentation for grouping images. In *2011 International Conference on Computer Vision*, pages 1535–1542. IEEE, 2011. ii

[38] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5703–5713, 2019. i, ii, vi

[39] Wenguan Wang, Hailong Zhu, Jifeng Dai, Yanwei Pang, Jianbing Shen, and Ling Shao. Hierarchical human parsing with typed part-relation reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8929–8939, 2020. i, ii, vi

[40] Yang Wang, Duan Tran, and Zicheng Liao. Learning hierarchical poselets for human parsing. In *CVPR 2011*, pages 1705–1712. IEEE, 2011. ii

[41] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2567–2575, 2022. ii

[42] Zhecan Wang, Jian Zhao, Cheng Lu, Fan Yang, Han Huang, Yandong Guo, et al. Learning to detect head movement in unconstrained remote gaze estimation in the wild. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3443–3452, 2020. ii

[43] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. iv

[44] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *Proceedings of the IEEE international conference on computer vision*, pages 3519–3526, 2013. ii

[45] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *2012 IEEE Conference on Computer vision and pattern recognition*, pages 3570–3577. IEEE, 2012. ii

[46] Lu Yang, Wenhe Jia, Shan Li, and Qing Song. Deep learning technique for human parsing: A survey and outlook. *International Journal of Computer Vision*, pages 1–32, 2024. ii, vi, vii, viii

[47] Lu Yang, Qing Song, Zhihui Wang, and Ming Jiang. Parsing r-cnn for instance-level human analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 364–373, 2019. vi

[48] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020. i, iii, vii

[49] Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo, and Kai Chen. Dense distinct query for end-to-end object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7329–7338, 2023. ii

[50] Xiaomei Zhang, Yingying Chen, Bingke Zhu, Jinqiao Wang, and Ming Tang. Blended grammar network for human pars-

ing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 189–205. Springer, 2020. vi

[51] Xiaomei Zhang, Yingying Chen, Bingke Zhu, Jinqiao Wang, and Ming Tang. Part-aware context network for human parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8971–8980, 2020. i, ii, vi

[52] Ziwei Zhang, Chi Su, Liang Zheng, and Xiaodong Xie. Correlating edge, pose with parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8900–8909, 2020. vi

[53] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. i, vi

[54] Jian Zhao, Jianshu Li, Xuecheng Nie, Fang Zhao, Yunpeng Chen, Zhecan Wang, Jiashi Feng, and Shuicheng Yan. Self-supervised neural aggregation networks for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 7–15, 2017. vi

[55] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. iv, viii