Fashionformer: A Simple, Effective and Unified Baseline for Human Fashion Segmentation and Recognition

Shilin Xu*, Xiangtai Li[1*], Jingbo Wang[2], Guangliang Cheng[3], Yunhai Tong[1], Dacheng Tao[4]

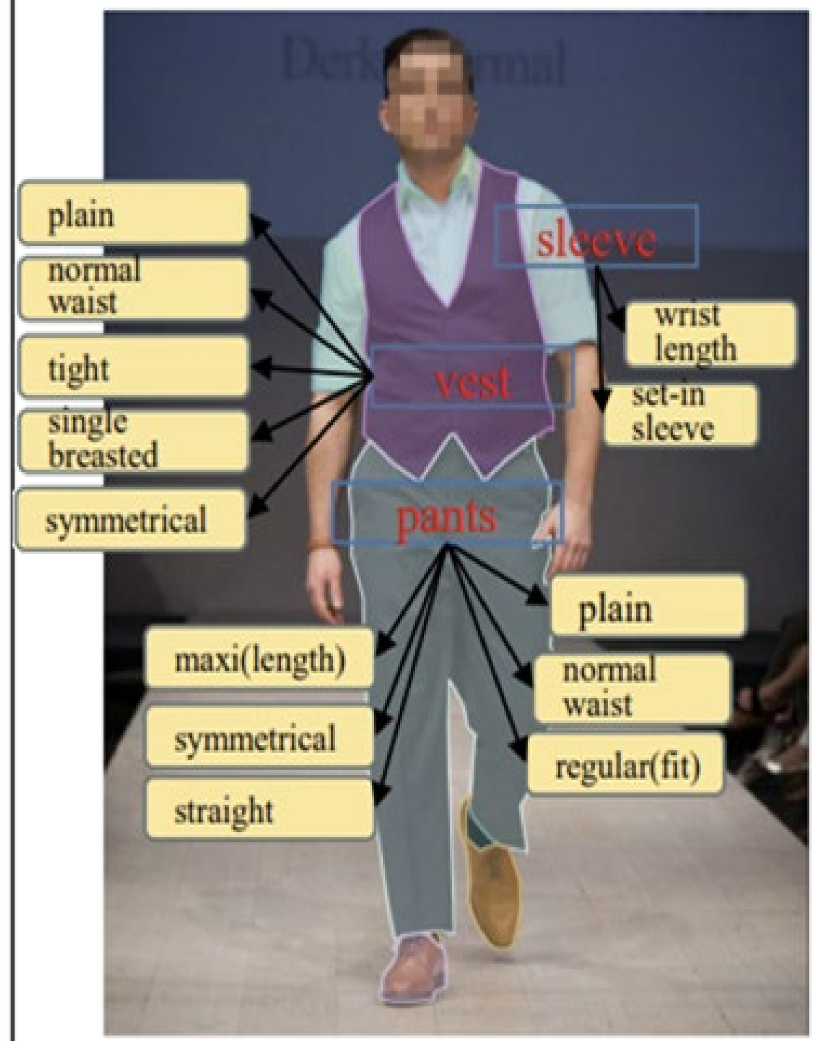
[1]Peking University, [2]The Chinese University of Hong Kong, [3]SenseTime Research, [4]The University of Sydney

Paper

Code

ECCV TEL AVIV 2022

1. Motivation and Introduction



1.1 Fashion Understanding: Fashionpedia

A joint and challenging task that perform **instance segmentation (cloth parts)** and **attribute recognition (cloth attributes)**.

It requires the model to output instance masks and multiple attribute labels.

1.2 Metric

Joint Fine-grained classifications and instance segmentation metric.

$$\mathrm{AP}^{\mathrm{mask}}_{\mathrm{IoU+F_1}}$$

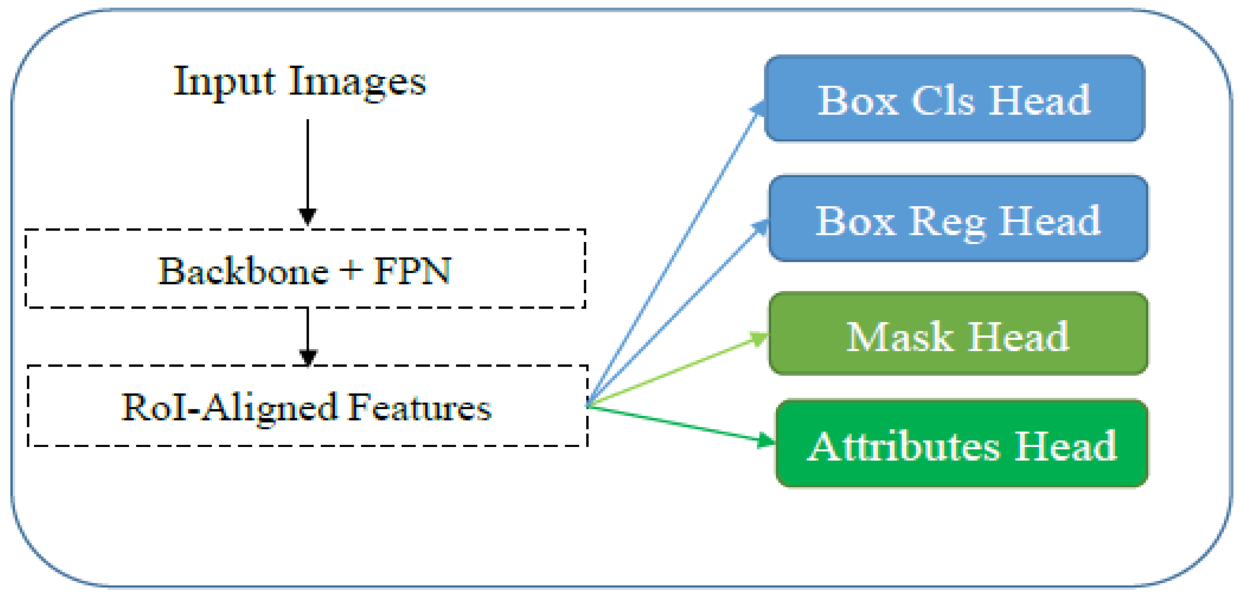
1.3 Why this task?

1.Fine grained understanding of human part mask and labels.

2.Application: Can be used for image retrieval for online shopping.

3.Application: 3D Human Reconstruction with Cloths.

2. Limitation of Existing Methods



Previous Solutions:

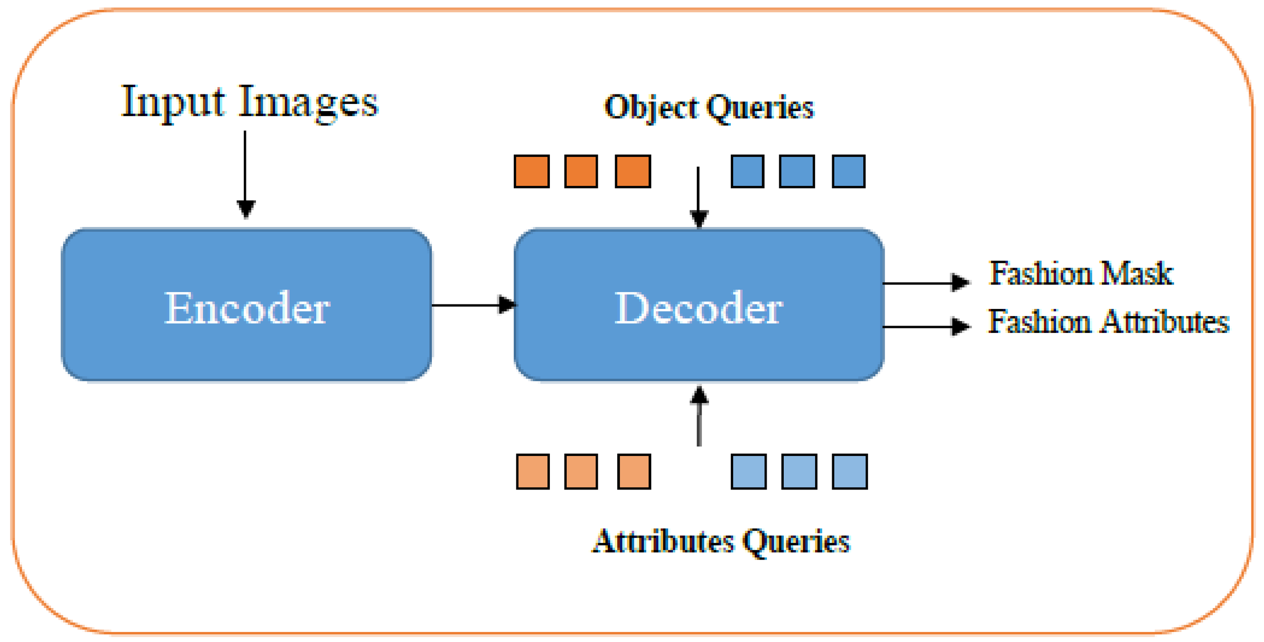
1.Complex pipeline (Two-stage-pipeline and Multiple task heads),

2.No task association. Instance segmentation and Attribution prediction are fully independent.

3.The Mask resolutions are limited. Mask quality is limited by ROI pooling.

4.Attribute predictions do not use the mask information to collect fine-grained features.

3. Our Method

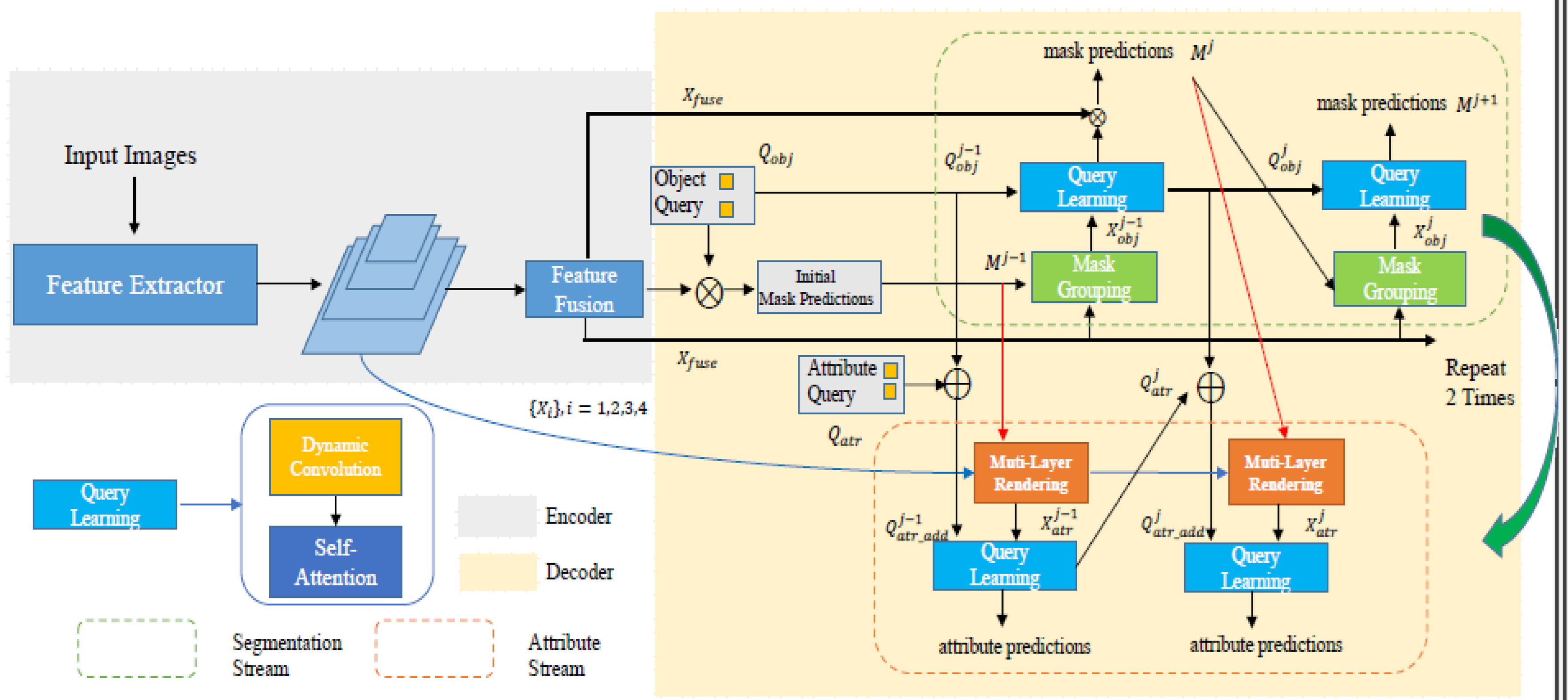


3.1 Key Design of our Fashionformer (DETR-like method)

1.Simpler Pipeline. No RoI-Align, No RPN and Single Stage.

2.Task association via Object Queries and Attributes Queries.

3.Joint Object Queries and Attribute Queries learning benefits Instance Segmentation.



3.2 Object Queries (above, green box):

1, Each object query performs Query Learning including a dynamic convolution with a self-attention.

2, The dynamic convolution design can found in previous works which weights query with corresponding query features.

3.3 Attribute Queries (below, origin box) :

1, Each attribute query first sum the corresponding object query for task association.

2,Attributes query features are obtained via **Multi-Layer Rendering (MLR, A shared MLP to absorb muti-scale features via mask grouping).**

3,Perform Query Learning as object query with Rendered attribute queries.

4. Experiments

RD	DC	MLR	N=1	I=3	$\mathrm{AP}^{\mathrm{mask}}_{\mathrm{IoU}}$	$\mathrm{AP}^{\mathrm{mask}}_{\mathrm{IoU+F_1}}$
✓	✓	✓	-	✓	**33.2**	**29.2**
-	✓	✓	-	✓	33.0	27.0
✓	-	✓	-	✓	32.1	27.5
✓	✓	-	-	✓	33.0	26.5
✓	✓	✓	✓	-	29.1	25.8
-	✓	-	-	✓	30.1	-

(a) Effect of each component.

Setting	$\mathrm{AP}^{\mathrm{mask}}_{\mathrm{IoU}}$	$\mathrm{AP}^{\mathrm{mask}}_{\mathrm{IoU+F_1}}$
NL=1	32.2	27.5
NL=2	33.1	28.6
NL=4	33.2	**29.2**

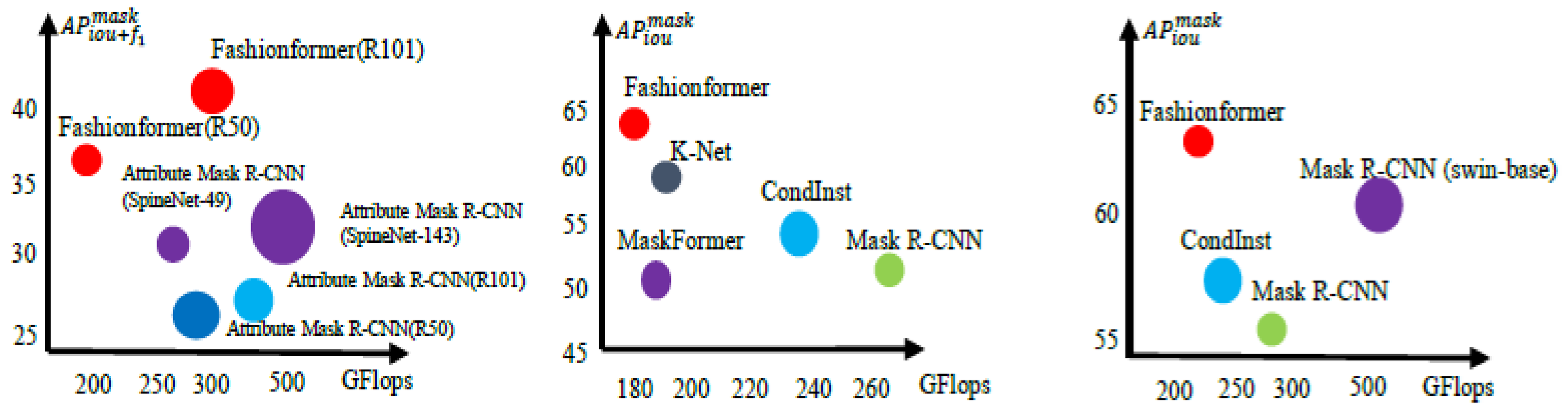
(b) Ablation on design of MLR. NL: number of layers in MLR.

Setting	$\mathrm{AP}^{\mathrm{mask}}_{\mathrm{IoU}}$	$\mathrm{AP}^{\mathrm{mask}}_{\mathrm{IoU+F_1}}$	Param(M)
Shared Query	33.4	27.6	36.2
Individual Query	33.3	**29.1**	37.7

(c) Decoupled query learning. The Shared Query use the same object query via a MLP to generate attribute query.

Table 1: Ablation studies and analysis on Fahsionpedia dataset set with ResNet50 as backbone. DC: Dynamic Convolution. RD: Residual Addition. MLR: Multi-Layer Rendering. N: Number of decoder layers.

4.1 Detailed Ablation Studies show the effectiveness of our Fashionformer design.



(a) Results on Fashionpedia (b) Results on ModaNet (c) Results on DeepFashion

4.2 New state-of-the-art results on three datasets !!!

method	backbone	schedule	GFlops	params(M)	$\mathrm{AP}^{\mathrm{mask}}_{\mathrm{IoU}}$ ↑	$\mathrm{AP}^{\mathrm{mask}}_{\mathrm{IoU+F_1}}$ ↑	G ↓
Attirbute-Mask R-CNN	R50-FPN	1×	296.7	46.4	34.3	25.5	8.8
		2×			38.1	28.5	9.6
		3×			39.2	29.5	9.7
Attirbute-Mask R-CNN	R101-FPN	1×	374.3	65.4	36.7	27.6	9.1
		2×			39.2	29.8	9.4
		3×			40.7	31.4	9.3
Attirbute-Mask R-CNN	SpineNet-49	6×	267.2	40.8	39.6	31.4	8.2
	SpineNet-96		314.0	55.2	41.2	31.8	9.4
	SpineNet-143		498.0	79.2	43.1	33.3	9.8
Fashionformer	R50-FPN	1×	198.0	37.7	**40.3**	**36.6**	3.7
		3×			**42.5**	**39.4**	3.1
Fashionformer	R101-FPN	1×	275.7	56.6	**43.2**	**40.5**	**2.7**
		3×			**45.6**	**42.8**	**2.8**
Attirbute-Mask R-CNN	Swin-b	3×	508.3	107.3	47.5	40.6	6.9
Fashionformer	Swin-b	3×	442.5	100.6	**49.5**	**46.5**	3.0

4.3 Significant Improvements Over Previous Baseline.

Image Ground Truth Attribute Mask R-CNN Fashionformer

4.4 Visual Improvements Over Previous Baseline.

More Details can be found In Our Paper