Unconstrained Fashion Landmark Detection via Hierarchical Recurrent Transformer Networks

Sijie Yan[1] Ziwei Liu[1] Ping Luo[1] Shi Qiu[2] Xiaogang Wang[1] Xiaoou Tang[1]
[1]The Chinese University of Hong Kong [2]SenseTime Group Limited
{ys016,lz013,pluo,xtang}@ie.cuhk.edu.hk,sqiu@sensetime.com,xgwang@ee.cuhk.edu.hk

ABSTRACT

Fashion landmarks are functional key points defined on clothes, such as corners of neckline, hemline, and cuff. They have been recently introduced [18] as an effective visual representation for fashion image understanding. However, detecting fashion landmarks are challenging due to background clutters, human poses, and scales as shown in Fig. 1. To remove the above variations, previous works usually assumed bounding boxes of clothes are provided in training and test as additional annotations, which are expensive to obtain and inapplicable in practice. This work addresses unconstrained fashion landmark detection, where clothing bounding boxes are not provided in both training and test. To this end, we present a novel *Deep LAndmark Network (DLAN)*, where bounding boxes and landmarks are jointly estimated and trained iteratively in an end-to-end manner. DLAN contains two dedicated modules, including a *Selective Dilated Convolution* for handling scale discrepancies, and a *Hierarchical Recurrent Spatial Transformer* for handling background clutters. To evaluate DLAN, we present a large-scale fashion landmark dataset, namely *Unconstrained Landmark Database (ULD)*, consisting of 30K images. Statistics show that ULD is more challenging than existing datasets in terms of image scales, background clutters, and human poses. Extensive experiments demonstrate the effectiveness of DLAN over the state-of-the-art methods. DLAN also exhibits excellent generalization across different clothing categories and modalities, making it extremely suitable for real-world fashion analysis.

KEYWORDS

Visual fashion understanding; landmark detection; deep learning; convolutional neural network

1 INTRODUCTION

Recently, interest in visual fashion analysis has been growing in the community and extensive research have been devoted to style discovery [1, 5, 12, 22], attribute prediction [2, 17, 26], and clothes retrieval [6, 9, 15–17, 28]. The reasons behind it are two-fold. On the one hand, visual fashion analysis brings enormous values to the industry, which is estimated to be a $2.5 trillion market[1] in the next five years. On the other hand, modern deep neural network architectures [23] and large-scale fashion databases [17] enable us to tackle these challenging tasks.

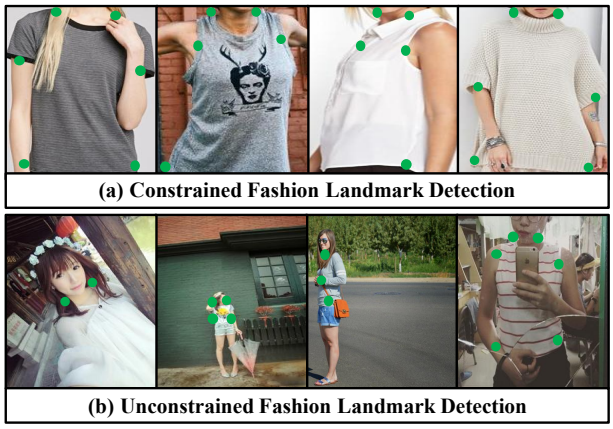

Figure 1: Comparison between input to (a) constrained fashion landmark detection, *e.g.* Deep Fashion Alignment (DFA) [18] and (b) unconstrained fashion landmark detection, *e.g.* Deep LAndmark Network (DLAN) in this work. DFA takes clothes bounding box as input while DLAN takes raw fashion images as input without any bounding box annotations.

Notably, a recent work Deep Fashion Alignment (DFA) [18] presented fashion landmarks, which are functional key points defined on clothes, such as corners of neckline, hemline, and cuff. Extracting features on the detected fashion landmarks significantly improve the performances of clothing image analysis. DFA assumed the clothing bounding boxes are given as prior information in both training and test. Using cropped fashion images as input, as shown in Fig. 1.(a), eliminates the scale variances and background clutters. However, obtaining additional bounding box annotations is both expensive and inapplicable in practice.

To overcome the above limitation, this work presents the problem of unconstrained fashion landmark detection, where clothing bounding boxes are not provided in both training and test as demonstrated in Fig. 1.(b). Unconstrained fashion landmark detection is confronted with two fundamental obstacles. First, clothing images are deformable objects in nature and thus subject to frequent style changes and occlusions that confuse existing systems. Second, depending on application scenarios, fashion items are often observed under different domains, such as selfies, street snapshots, and online shopping photos, which exhibit severe variations and also cross-domain distribution discrepancies.

[1]http://www.emarketer.com/

MM'17, October 23–27, 2017, Mountain View, CA, USA.

ISBN 978-1-4503-4906-2/17/10…$15.00
DOI: https://doi.org/10.1145/3123266.3123276

	# VGGs	# bbox anno.	end-to-end	# inference pass	speed (fps)	det. rate (%)
Sliding Window + DFA [18]	1	×	×	17	3.2	2.7
Clothes Proposal + DFA [18]	1	×	×	100	0.5	9.7
Clothes Detector + DFA [18]	2	16K	×	1	5.0	63.1
Joint RPN [20] + DFA [18]	2	16K	√	1	3.9	66.0
Deep LAndmark Network	1	×	√	1	**5.2**	**73.8**

Table 1: Summary of Deep LAndmark Network (DLAN) and other unconstrained landmark detection paradigms. From left to right: the number of convolutional neural networks (*i.e.* VGGs [23]) used, the number of bounding box annotations used during training, whether it enables end-to-end learning or not, the number of inference passes needed during testing, runtime speed ($frames\ per\ second$) and detection rate ($threshold = 35\ pixels$). Compared with the other alternative methods, DLAN achieves state-of-the-art performance and fast speed at the same time.

We solve these problems by proposing a novel *Deep LAndmark Network (DLAN)*[2], which consists of two important components, including a component of *Selective Dilated Convolution* for handling scale discrepancies, and a component of *Hierarchical Recurrent Spatial Transformer (HR-ST)* for handling background clutters. The first module employs dilated convolutions to capture the fine-grained fashion traits appeared in different scales. The second module incorporates attention mechanism [8] into DLAN, which makes the model recurrently search for and focus on the estimated clothes functional parts or regions. We also propose scale-regularized regression for robust learning.

This paper has three main **contributions**. First, we thoroughly investigate the problem of unconstrained fashion landmark detection for the first time. A large-scale dataset, *Unconstrained Landmark Database (ULD)*, is collected and contributed to this community, which we believe can facilitate future research. Second, we propose *Deep LAndmark Network (DLAN)* tailored for unconstrained landmark detection without bounding box annotations in both training and testing. It is capable to handle scale discrepancies and background clutters which are common in fashion images. Compared with the other alternative methods, DLAN achieves state-of-the-art performance and fast speed at the same time, as shown in Table 1. Third, since DLAN is an end-to-end learnable system, it can be easily transferred to a new domain without much adaptation. Extensive experiments demonstrate that DLAN exhibits excellent generalization across different categories and modalities.

2 RELATED WORK

Fashion Understanding in Computer Vision. Many human-centric applications depend on reliable fashion image understanding. And lots of efforts from the community have been devoted to pursue this goal. Recent advanced methods include predicting semantic attributes [1–3, 17, 26], clothes recognition and retrieval [6, 10, 11, 16, 17, 28], and fashion trends discovery [12, 21, 27]. To capture discriminative information on clothes, previous works have explored the representation of object proposals [11], bounding boxes [1, 3], parsing masks [14, 28–30], and fashion landmarks [18].

Among the above representations, detecting fashion landmarks is an effective and robust way for fashion recognition; but existing work Deep Fashion Alignment (DFA) [18] assumed the bounding box annotations are provided in both training and testing stages, which are impractical in real-world applications. Unlike previous method, we present DLAN, which can train and evaluate on full images without bounding box annotations.

Joint Localization and Landmark Detection. Previous methods have studied joint localization and landmark detection in the context of face alignment [32] and human body pose estimation [24]. For example, Zhu *et. al.* modeled facial landmarks as different parts and learned a tree-structured model to find optimal configuration of these parts. Its performance was limited by the expressive power of hand-crafted features. Tompson *et. al.* adopted multi-scale Fully Convolutional Network (FCN) to mimic the traditional 'sliding window + image pyramid inference' paradigm, where the FCN is evaluated on exhaustive multi-scale patches cropped from the input. However, the resulting model has heavy computational burden and also is sensitive to background clutters. Our approach proposes a *Hierarchical Recurrent Spatial Transformer* module to localize fashion items in one feed-forward pass. Furthermore, scale variations are addressed by *Selective Dilation Convolutions*.

Table 1 compares DLAN with existing unconstrained landmark detection paradigms, including 'Sliding Window + DFA', 'Clothes Proposal + DFA', 'Clothes Detector + DFA' and 'Joint Region Proposal Network (RPN) [20] + DFA'. For 'Sliding Window + DFA', we apply DFA on sliding-window-extracted patches (which are also augmented by multi-scale pyramids) and get the final results by voting. For 'Clothes Proposal + DFA', DFA is applied to the top-100 object proposals generated by EdgeBox [33]. The final results are also obtained by voting. For 'Clothes Detector + DFA', we first obtain clothes bounding box by Fast R-CNN [7] and then apply DFA on these bounding boxes. These two models are trained and inferenced sequentially. For 'Joint RPN + DFA', we replace the fully-connected layers beyond RoI pooling of RPN [20] with the fully-connected layers for landmark detection in DFA [18]. Then we train these two sub-networks end-to-end for joint inference. Compared with the other alternative methods, DLAN achieves state-of-the-art performance and fast speed at the same time.

3 DATABASE

Fashion landmark detection has been mostly studies in the context of online fashion shop images, whose characteristics are well demonstrated in the Fashion Landmark Detection dataset (FLD) collected by Liu et al. [18]. Here, to thoroughly investigate the problem of unconstrained landmark detection, we contribute *Unconstrained Landmark Database (ULD)*, which comprises 30K images with comprehensive fashion landmark annotations. The images in

[2]The code and models are available at https://github.com/yysijie/DLAN/.

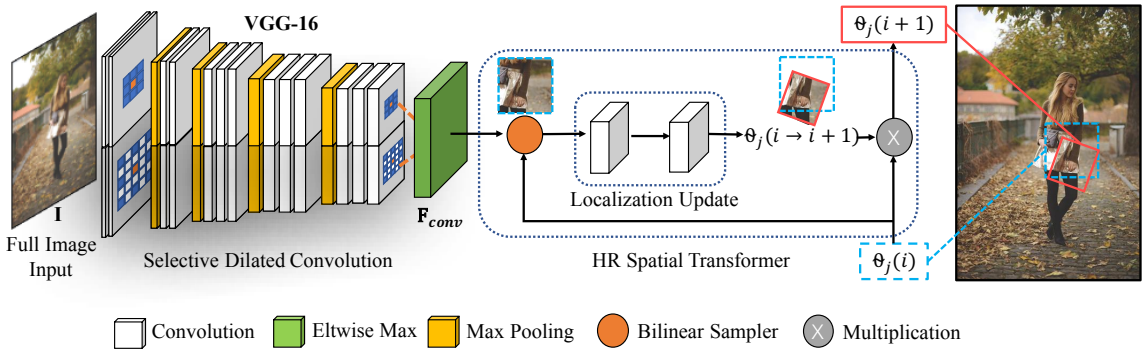

Figure 2: Pipeline of Deep LAndmark Network (DLAN) for unconstrained fashion landmark detection, where clothing bounding boxes are not provided in both training and test. DLAN contains two dedicated modules, including a *Selective Dilated Convolution* for handling scale discrepancies, and a *Hierarchical Recurrent Spatial Transformer* for handling background clutters.

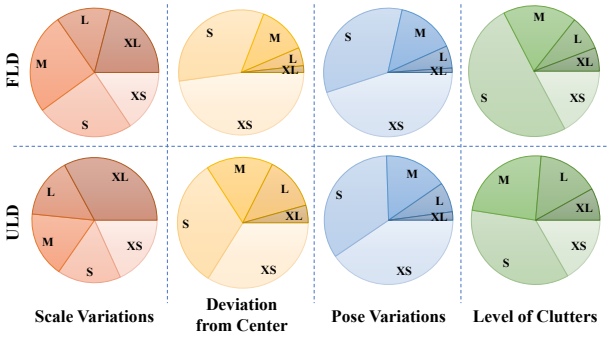

Figure 3: Dataset statistics of Fashion Landmark Detection Benchmark (FLD) [18] and Unconstrained Landmark Database (ULD). We compare these two datasets *w.r.t* scale variations, deviation from center, pose variations and level of clutters. 'XS' represents extreme small variation, 'S' represents small variation, 'M' represents medium variation, 'L' represents large variation and 'XL' represents extreme large variation.

ULD are collected from fashion blogs, forums and the consumer-to-shop retrieval benchmark of DeepFashion [17]. Most of images in ULD are taken and submitted by common customers.

ULD *v.s.* FLD. ULD differs from FLD in two major aspects. First, unlike e-commerce shop images, the images in ULD haven't gone through standardization process. Thus, its clothes items often locate among cluttered background and deviate from image center. Second, unlike professional fashion models, common customers tend to take cloth snapshots from a much wider spectrum of poses and scales. A successful approach should be both discriminative to true landmark traits and robust to all the other variations.

Data Statistics. Fig. 1 presents some visual examples illustrating the difference between FLD and ULD. Compared with FLD, we can see that images in ULD exhibit heavy background clutters, large pose variations and scale variations. We further quantitatively analyze and contrast the dataset statistics of ULD to that of FLD, as illustrated in Fig. 3. The databases are scrutinized along four axes: scale variations, deviation from center, pose variations and background clutters. The 'level of background clutters' is measure by the number of object proposals needed to localize the fashion item, following [4]. The 'pose variations' is defined by the residual energy of landmark configurations after removing principal component while the 'deviation from center' is defined as the normalized distance between clothes center and image center. It can be observed that ULD contains substantial foreground scatters and background clutters, as well as dramatic geometric deformations (especially zoom-in). The unconstrained landmark database (ULD) exhibits real-world variations, and is challenging for existing clothes recognition methods, which we believe can facilitate future research.

4 APPROACH

Framework Overview. Fig. 2 shows the pipeline of our proposed Deep LAndmark Network (DLAN). Given a raw fashion image I and assume there are overall J fashion landmarks, our goal is to predict landmark locations $l_j, j = 1, 2, \ldots, J$ without bounding box annotations during both training and testing. It is a typical case for online fashion applications, where the demo images and posts are subject to constant change, thus collecting bounding box annotations is impractical. DLAN goes beyond the traditional constrained landmark detection and provides a principle framework for accurate unconstrained landmark detection. DLAN contains two dedicated modules, including a *Selective Dilated Convolution* for handling scale discrepancies, and a *Hierarchical Recurrent Spatial Transformer* for handling background clutters.

4.1 Deep LAndmark Network (DLAN)

In this section, we elaborate the detailed components of our approach. Similar to Deep Fashion Alignment (DFA) [18], we adopt VGG-16 [23] as our backbone network.

Converting into FCN. First, to enable full-image inference, we convert DFA into Fully Convolutional Network (FCN), which we call *fully convolutional DFA*. Specifically, following [19], two fully-connected ('fc') layers in DFA are transformed to two convolutional layers in DLAN, respectively. The first 'fc' layer learns $7 \times 7 \times 512 \times 4096$ parameters, which can be altered to 4096 filters, each of which is $7 \times 7 \times 512$. The second 'fc' layer learns a 4096×4096 weight matrix, corresponding to 4096 filters. Each filter is $1 \times 1 \times 4096$.

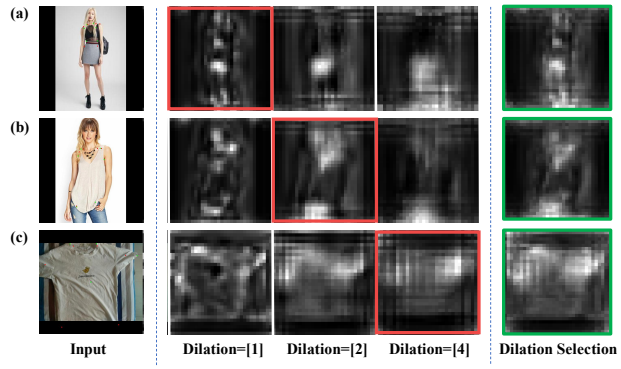

Figure 4: Illustration of *Selective Dilated Convolutions*. From left to right: input image, convolutional feature map F_{conv} of 'Dilation = [1]', 'Dilation = [2]', 'Dilation = [4]' and 'Selective Convolution'.

Selective Dilated Convolutions. Real-life fashion images often exhibit substantial scale variations, *e.g.* zoom-in and zoom-out. To cope with these scale discrepancies, we further augment fully-convolutional DFA with *Selective Dilated Convolutions*, which is denoted as $Conv_{SD}$. We further denote the feature maps in layer i as F_i. Specifically, for convolutional filters $k_i, i = conv1, \ldots, conv5$, besides the response $F_{i-1} * k_i$ obtained from their original receptive fields, we further collect responses $F_{i-1\star 2^s} * k_i$ from exponentially expanded receptive fields [31], where $\star$ represents expanded sampling and $*$ represents convolution operation. Assume there are overall S expanded receptive fields. We call each set of scale-specific responses as a scale tower for $s = 1, \ldots, S$. All scale towers share weights for all subsequent processes. Each scale tower captures the convolutional responses for that scale s. The final convolutional response F_{conv} is obtained by selecting the element-wise maximum *conv5* response among all scale towers:

$$F_{conv} = Conv_{SD}(I) = \max_s F_{conv4\star 2^s} * k_{conv5} \quad (1)$$

The selected scale is denoted as $Dilation = [2^{s_{max}}]$. Sec. 5.1 empirically supports that selection by picking element-wise maximum response is superior to other alternatives, such as average fusion. Our *selective dilated convolution* effectively adapts fine-grained single-scale filters to more flexible inputs. For zoom-out image (Fig. 4 (a)), the response of low-scale filters ($Dilation = [1]$) is selected; while for zoom-in image (Fig. 4 (c)), the response of high-scale filters ($Dilation = [4]$) is selected.

4.2 Hierarchical Recurrent Spatial Transformer (HR-ST)

Besides scale variations, web-style fashion inputs also see lots of global deviation from center and local geometric deformation. It is desirable to remove background clutters and transform target regions into canonical form. Spatial transformer [8] has been recently introduced to learn to roughly align feature maps for subsequent tasks. Specifically, given feature maps F_{conv} of the original input image, the spatial transformer $\mathcal{T}_\Theta$ seeks to find a geometric transform Θ that will produce "aligned" feature maps F_{trans} without explicit supervision[3]:

$$F_{trans} = \mathcal{T}_\Theta(F_{conv}) \quad (2)$$

Note that the geometric transform Θ is also estimated from its input F_{trans}. Then we can perform landmark regression for all landmarks $j = 1, 2, \ldots, J$ on these "aligned" feature maps F_{trans}:

$$\hat{l}'_j = \mathcal{R}(F_{trans}),\ j = 1, 2, \ldots, J \quad (3)$$

where $\mathcal{R}(\cdot)$ is the regression function which takes the form of fully-connected layers here. However, direct regressing original landmark locations *w.r.t* this transformed feature map F_{trans} is hard since it entangles the original image coordinates and also the learned geometric transformation Θ. Here, we advocate to first predict the *relative landmark coordinates* $\hat{l}'_j$ and then transform them back to the *original coordinates* $\hat{l}_j$:

$$\hat{l}_j = \Theta \cdot \hat{l}'_j \quad (4)$$

It should be noted that the original coordinates $\hat{l}_j = (\hat{x}_j, \hat{y}_j)$ have been normalized. Thus $\hat{x}_j$, $\hat{y}_j$ range from -1 to 1 for landmarks within the image. In this way, all the regression can be performed under a consistent coordinate system, thus easing the prediction difficulties. The whole pipeline is illustrated in Fig. 5 (a).

Another challenge is that unconstrained fashion images usually undergo drastic global and local deformations. A single geometric transform Θ is neither expressive enough for all the possible variations nor easy to estimate in a single pass. In addition, traditional spatial transformer [8] is known to be sensitive to background clutters and only captures local information. To overcome these barriers, we propose *Hierarchical Recurrent Spatial Transformer (HR-ST)*, which recurrently learns a group of geometric transforms[4] $\Theta_j(i)$ for fashion landmark $j = 1, 2, \ldots, J$ under recurrent step $i = 1, 2, \ldots, M$. Fig. 5 (b) depicts the pipeline of our hierarchical recurrent spatial transformer, whose details are elaborated below.

Recurrent Update. In the context of unconstrained fashion landmark detection, each transformation $\Theta_j(i)$ can be quite complex. To ease the estimation difficulty, we further make $\Theta_j(i)$ recurrently updated:

$$\Theta_j(i) \leftarrow \Theta_j(i-1) \cdot \Theta_j(i-1 \rightarrow i) \quad (5)$$

In each recurrent step $i = 2, 3 \cdots, M$, our hierarchical recurrent spatial transformer module only has to predict a refinement transformation $\Theta_j(i-1 \rightarrow i)$ instead of a direct transformation. It fully exploits the decision dependencies within input samples and thus is easy to estimate. In our experiment, we take $M = 3$ steps for HR Spatial Transformer.

Hierarchical Modeling. Since clothing item naturally forms a tree-like structure with root (clothes/human body) and leaves (local landmark patches), we decompose $\Theta_j(M)$ into global base transformation $\Theta(1)$ and local refinement transformation $\Theta_j(i-1 \rightarrow i), i > 1$ for each landmark j:

$$\Theta_j(M) = \Theta_{global} \cdot \Theta_{local} = \Theta(1) \cdot \prod_{i=2}^{M} \Theta_j(i-1 \rightarrow i) \quad (6)$$

[3] It is also known as differentiable image sampling with a bilinear kernel, which allows end-to-end training via stochastic gradient descent (SGD) [13].

[4] Similar to [8], $\Theta_j(i)$ takes the form of a 2D affine transform, which has 6 parameters to estimate, representing x, y translation, rotation and scaling respectively.

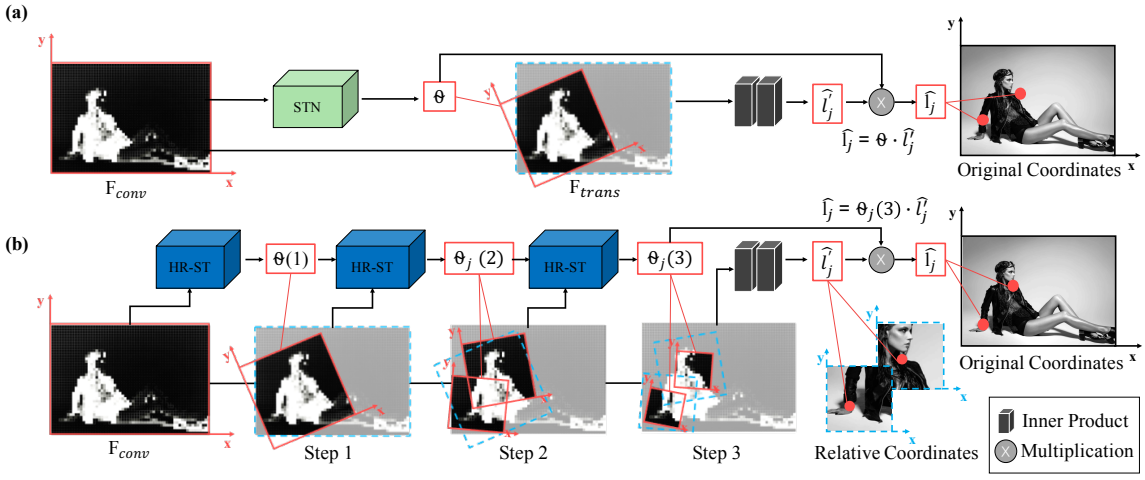

Figure 5: Pipeline of (a) single spatial transformer network (STN) with coordinate transform and (b) our hierarchical recurrent spatial transformer (HR-ST). *HR Spatial Transformer* recurrently learns a group of geometric transforms $\Theta_j(i)$ which progressively attend to landmark regions. By combining the strength of hierarchy and recurrent update, *HR Spatial Transformer* is able to jointly localize clothes and detect landmarks by predicting a sequence of global and local geometric transformations without explicit supervision.

where Θ_{global} essentially localizes clothes items while Θ_{local} detects local landmark patch j, as shown in Fig. 5 (b). Hierarchical decomposition of $\Theta_j(M)$ enables powerful modeling of both global and local deformations of clothing items.

By combining the strength of hierarchy and recurrent update, *HR Spatial Transformer* is able to jointly localize clothes and detect landmarks by predicting a sequence of global and local geometric transformations without explicit supervision.

4.3 Learning

For our DLAN training, we employ the l_2 Euclidean distance to constitute the landmark regression loss[5]. For simplicity, the recurrent step i is omitted for all following notations.

$$L_{regression} = \sum_{1}^{J} \frac{1}{2} \|l_j - \hat{l}_j\|_2^2 = \sum_{1}^{J} \frac{1}{2} \|l_j - \Theta_j \hat{l}'_j\|_2^2, \tag{7}$$

where l_j is the ground truth landmark location and $\hat{l}_j$ is our prediction under the original image coordinates. Θ_j is the estimated geometric transformation while $\hat{l}'_j$ is the predicted relative landmark location by DLAN.

Scale Regularization. Since the geometric transformation Θ_j can result in arbitrary-size output, we would like to regularize the output scale of the estimated transformation Θ_j. Θ_j transform a square which overlaps boundary of input to a quadrangle. It can be proved that the area of this quadrangle is $4 \det \Theta_j$, four times determinant of Θ_j . We can supervise this area for regularizing output scale:

$$L_{scale} = \sum_{1}^{J} \frac{1}{2} (\lambda C_{scale} - 4 \det \Theta_j)^2. \tag{8}$$

[5]Following [18], we marked out the occluded and invisible landmarks in our final loss, i.e. the prediction error gradients of these landmarks are not propagated back to the network.

where C_{scale} represents the area of the convex hull of ground truth landmarks and the scaling factor λ controls the spatial extent of our attended landmark patches.

Our final loss function is the combination of these two:

$$L = L_{regression} + L_{scale}. \tag{9}$$

Back Propagation. To make our DLAN a fully differentiable system, we define the gradients with respect to both the estimated geometric transformation Θ_j and the predicted relative landmark location $\hat{l}'_j$. Then the partial derivatives of our loss function L are provided as follows:

$$\frac{\partial L}{\partial \hat{l}'_j} = -\sum_{1}^{J} \Theta_j^T (l_j - \Theta_j \hat{l}'_j) \tag{10}$$

$$\frac{\partial L_{regression}}{\partial \Theta_j} = -\sum_{1}^{J} (l_j - \Theta_j \hat{l}'_j) \hat{l}'^T_j \tag{11}$$

$$\frac{\partial L_{scale}}{\partial \Theta_j} = -\sum_{1}^{J} 4(\lambda C_{scale} - 4 \det \Theta_j) \det \Theta_j (\Theta_j^{-1})^T \tag{12}$$

where Θ_j^T is the transpose of Θ_j and Θ_j^{-1} is the inverse of Θ_j. Rather than only relying on indirect propagated errors to estimate Θ as in [8], our gradients impose more direct and strong supervisions on both Θ_j and $\hat{l}'_j$, enabling stable convergence. Our *Deep LAndmark Network (DLAN)* is an end-to-end trainable system that can jointly optimize clothes localization and landmark detection.

5 EXPERIMENTS

This section presents evaluation and analytical results of Deep LAndmark Network (DLAN), as well as showing it enables excellent generalization across different clothing categories and modalities.

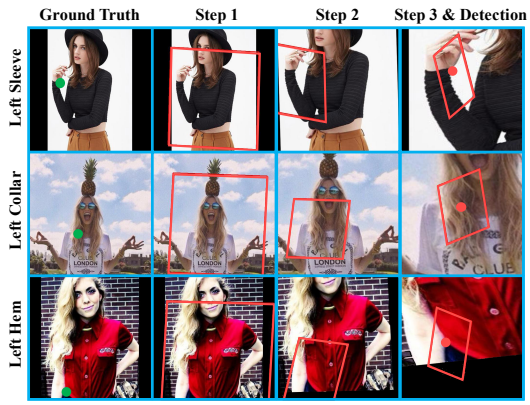

Figure 6: Illustration of Hierarchical Recurrent Spatial Transformer (HR-ST). From left to right: input image, the spatially transformed output of HR-ST in step 1, 2 and 3.

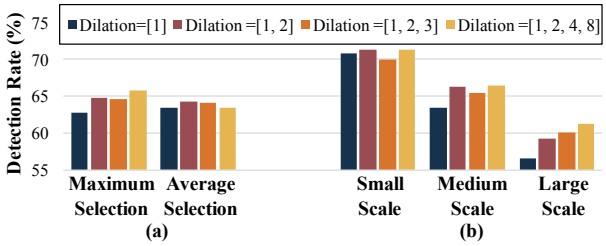

Figure 7: Comparative study on selective dilated convolutions: (a) maximum selection v.s. average selection, (b) performance on different scale variations.

Experiment Settings. We initialize our *Deep LAndmark Network (DLAN)* with the released model of Deep Fashion Alignment (DFA)[6]. Then, DLAN is trained on the full images (size of 512 × 512) of *Unconstrained Landmark Detection (ULD)*, where 16K images are used for training, 8K are used for validation and 6K are used for testing. All the results are reported on ULD test set.

Evaluation Metrics. Following existing works [25] on human pose estimation, we employ percentage of detected landmarks (PDL) to evaluate fashion landmark detection. PDL is calculated as the percentage of detected landmarks under certain overlapping criterion. Typically, higher values of PDL indicate better results. We fix the distance threshold to 35 pixels throughout all the experiments.

Competing Methods. DLAN is benchmarked against several best performing methods for unconstrained landmark detection, which are roughly classified into three groups: (1) *sliding window based*: 'Sliding Window + DFA'. Note that DFA is not re-trained here. (2) *sequential localization and landmark detection*: 'Clothes Proposal + DFA' and 'Clothes Detector + DFA' all fall into this category. For the latter, these two networks are trained separately. (3) *joint localization and landmark detection*: 'RPN + DFA'. Note that these two sub-networks are jointly trained.

[6]We use the released VGG-16 model, which is public available at https://github.com/liuziwei7/fashion-landmarks .

					DLAN
Fully Convolutional DFA?	✓	✓	✓	✓	✓
Spatial Transformer?		✓	✓	✓	✓
Selective Dilated Convolutions?			✓	✓	✓
HR Spatial Transformer?				✓	✓
Scale Regularization?					✓
detection rate (%)	56.9	62.8	64.8	71.2	**73.8**

Table 2: Ablation study on different components of *Deep LAndmark Network (DLAN)*. 'Selective Dilated Convolutions', 'Hierarchical Recurrent Spatial Transformer' and 'Scale Regularization' are effective for unconstrained fashion landmark detection.

# Step	det. rate (%)	time (ms)
1	64.8	177.7
2	70.0	+ 5.7
3	71.2	+ 3.1
4	71.4	+ 3.0
5	**71.5**	+ 2.9

(a)

	det. rate (%)
w/o S.R.	71.2
$\lambda = 0.8$	73.3
$\lambda = 0.4$	**73.8**
$\lambda = 0.2$	73.0
$\lambda = 0.1$	72.8

(b)

Table 3: Comparative study on hierarchical recurrent spatial transformer and scale regularization: (a) step-wise performance and speed analysis of HR spatial transformer, (b) performance *w.r.t* the scaling factor λ. 'det. rate' represents detection rate and 'w/o S.R.' represents without scale regularization.

5.1 Ablation Study

In this section, we perform an in-depth study of each component in Deep LAndmark Network (DLAN).

Component-wise Investigation. Table 2 presents the component-wise investigation of Deep LAndmark Network (DLAN). First, we convert DFA to fully-convolutional DFA as described in Sec. 4.1. Then, we gradually add 'selective dilated convolutions', 'hierarchical recurrent spatial transformer' and 'scale regularization' to this base model, which lead to 7.9%, 6.4% and 2.6% gains respectively. Our final model achieves the detection rate of 73.8%, demonstrating its superiority over base model on the problem of unconstrained fashion landmark detection.

Effectiveness of Selective Dilated Convolutions. Here we explore the hyper-parameter choices within selective dilated convolutions, as presented in Fig. 7. Selective dilated convolution is intended to learn to aggregate multi-scale information. *Dilation* = [1, 2, 4, 8] represents the scale towers 1, 2, 4 and 8 are used. First, we examine different aggregation techniques, namely 'maximum selection' which picks the maximum response across all scale towers, and 'average selection', which takes the average response of all scale towers. Comparative study shows that 'average selection' is inferior to 'maximum selection' since it fuses information from different scales together without discrimination. To demonstrate the merit of each scale tower, we further compare the performance of selective dilated convolutions on three subsets: images with small scale clothes, medium scale clothes and large scale clothes. For example, it can be observed that including the information from scale tower 4 and 8 (*i.e. Dilation* = [1, 2, 4, 8]), which possess expanded receptive fields, can indeed boost the performance on large scale clothes.

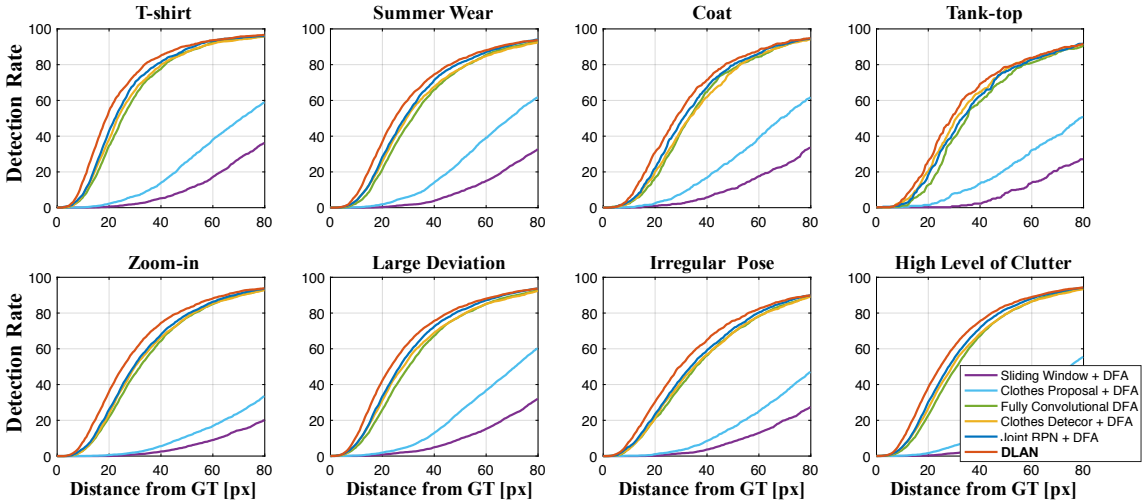

Figure 8: Performance comparisons of unconstrained fashion landmark detection on different clothing categories (the first row) and different clothing variation types (the second row). [px] represents pixels.

	L. Collar	R. Collar	L. Sleeve	R. Sleeve	L. Hem	R. Hem	Mean
Fully Convolutional DFA	75.4%	75.7%	52.1%	52.7%	61.2%	61.6%	60.8%
Clothes Detector + DFA	76.3%	76.1%	56.3%	57.6%	61.7%	61.1%	63.1%
Joint RPN + DFA	79.5%	79.8%	55.0%	57.7%	65.4%	66.6%	66.0%
DLAN	**83.3%**	**83.7%**	**64.6%**	**66.7%**	**71.7%**	**72.4%**	**73.8%**

Table 4: Performance comparisons of unconstrained fashion landmark detection on different fashion landmarks. 'L. Collar' represents left collar while 'R. Collar' represents right collar.

Admittedly, adopting the Selective Dilated Convolution is a performance/speed tradeoff. By removing Selective Dilated Convolution, the detection rate (%) drops from 73.80 to 70.22, while the executing speed (FPS, frames per second) increases from 5.2 to 7.8. *Selective dilated convolution* is a flexible yet effective technique for handling scale variations.

Effectiveness of HR Spatial Transformer. Next, we provide a step-wise analysis of hierarchical recurrent spatial transformer, which is listed in Table 3a. Hierarchical recurrent spatial transformer has the advantage of iteratively removing background clutters. By recurrent updating, the landmark detection performance improves steadily while the overhead computational time is negligible. From step-1 to step-3, our final detection rate increases by 6.4% while the incurring running time is only 8.8 ms. Diminishing returns are observed when further extending step-3 to step-5. To strike a balance between performance and speed, the recurrent step is set to 3 in our following experiments. The visual results of each step are illustrated in Fig. 6. *Hierarchical recurrent spatial transformer* progressively attends to the true landmark regions regardless of various background clutters.

Effectiveness of Scale Regularization. We inspect the effect of *scale regularization* by varying the scaling factor λ, which controls the desired landmark areas that our hierarchical recurrent spatial transformer will attend to. From Table 3b, we can observe that adding scale regularization during training always leads to 2% $\sim$ 3% performance gains. Specifically, best performance is achieved at $\lambda = 0.4$, *i.e.* the landmark areas are encouraged to be approximately 20% of the input coordinates. Therefore, we set λ to 0.4 in the following experiments.

5.2 Benchmarking

To illustrate the advantage of DLAN, we compare it with state-of-the-art unconstrained landmark detection methods like 'Fully-convolutional DFA', 'Clothes Detector + DFA' and 'Joint RPN + DFA'. We also analyze the strengths and weaknesses of each method on unconstrained fashion landmark detection.

Per-category Analysis. Fig. 8 (the first row) shows the percentage of detection rates on different clothing categories, where we have three observations. First, our DLAN announces the most advantages when the distance threshold is small (*i.e* the left side of the curves). DLAN is capable of accurately detecting fashion landmarks thanks to the power of hierarchical recurrent spatial transformer. Second, 'T-shirt' is the easiest clothing category to detect fashion landmarks while 'Tank-top' is the hardest. By inspecting closer, we find that 'T-shirt' generally have much more textures and local distinguishable traits. Third, DLAN consistently outperforms both 'Clothes Detector + DFA' and 'Joint RPN + DFA' on all clothing categories, showing that our dedicated modules and end-to-end trainable system are effective for unconstrained fashion landmark detection.

Per-variation Analysis. Fig. 8 (the second row) shows the percentage of detection rates on different clothing variation types. Again,

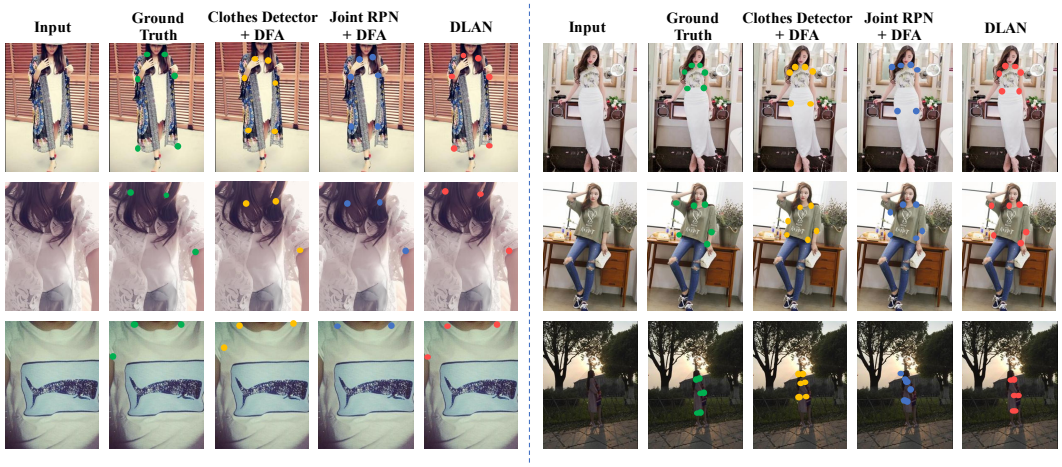

Figure 9: Visual comparisons of unconstrained fashion landmark detection by different methods. From left to right: input image, ground truth fashion landmark locations, landmark detection results by 'Clothes Detector + DFA', 'Joint RPN + DFA' and our DLAN.

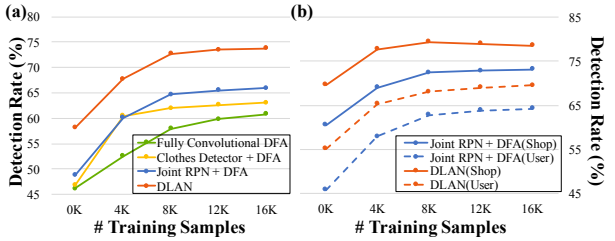

Figure 10: Generalization of DLAN *w.r.t* (a) the number of training samples, (b) input modalities (*i.e.* shop images v.s. user images).

DLAN outperforms all the other methods at all distance thresholds. We have two additional observations. First, DLAN exhibits excellent performance on images with zoom-in variations, when compared with other alternatives. Selective dilated convolutions enable our model to aggregate information from different input scale. Second, images with irregular poses create great obstacle for all the methods. Incorporating pairwise relationship between landmarks might partially alleviate this barrier and is a future direction to explore.

Per-landmark Analysis. Table 4 demonstrates the percentage of detection rates on different fashion landmarks, with the distance threshold fixed at 35 pixels. We can observe that our approach achieves the best performance and has substantial advantages over all the fashion landmarks. Another observation is that detecting 'sleeve' is more challenging than detecting 'collars' and 'hemlines' since the positions of sleeves are much more flexible due to various clothing styles and human poses in real life.

We also train and evaluate DLAN on Fashion Landmark Detection Benchmark (FLD) [18] under the same protocol as [18]. Without bounding box annotations during testing, our DUN achieves an average detection rate of 55% on FLD test set. According to Fig. 6 in [18], this performance surpasses one stage DeepPose (34%) and DFA (53%), which have access to the ground truth bounding boxes. Visual comparisons of unconstrained fashion landmark detection by different methods are given in Fig. 9. DLAN produces accurate and robust predictions.

5.3 Generalization of DLAN

To demonstrate the generalization ability of Deep LAndmark Network (DLAN), we further investigate how DLAN performs *w.r.t* the number of training samples as well as the input modalities.

 # Training Samples. We first investigate an important question in real-world applications: how many training samples are needed for DLAN to perform well? From Fig. 10 (a), we can see that DLAN is able to achieve competitive results even when only small-scale training data is available. For example, when there are only hundreds of training samples used, DLAN can still hit nearly 60% detection rate and attain more than 10% advantage over the other methods. The recurrent update mechanism in DLAN makes it fully exploit the decision dependencies within input samples, thus less data-hungry.

Input Modalities. Then we inspect how input modalities (*i.e.* shop images v.s. user images) affect the unconstrained fashion landmark detection performance. Fig. 10 (b) shows that while user images are generally more challenging, DLAN outperforms the existing best-performing method by large margins on both modalities. Selective dilated convolutions and hierarchical recurrent spatial transformer enable DLAN to eliminate various variations/clutters and quickly attend to target regions.

6 CONCLUSION

In this work, we study the problem of unconstrained fashion landmark detection, where no bounding box annotations are provided during both training and testing. To address this challenging task, we propose Deep LAndmark Network (DLAN) with two dedicated modules: *Selective Dilated Convolutions* for handling scale discrepancies and *Hierarchical Recurrent Spatial Transformer* for handling background clutters. In addition, a large-scale *unconstrained landmark detection database (ULD)* is contributed to the community, which reflects real-life difficulties and thus serves as a suitable touchstone for existing vision-for-fashion systems. Extensive experiments demonstrate the effectiveness of DLAN over other state-of-the-art methods. DLAN also exhibits excellent generalization across different clothes categories and modalities, making it extreme suitable for real-world fashion analysis.

REFERENCES

[1] Lukas Bossard, Matthias Dantone, Christian Leistner, Christian Wengert, Till Quack, and Luc Van Gool. 2012. Apparel classification with style. In *ACCV*.

[2] Huizhong Chen, Andrew Gallagher, and Bernd Girod. 2012. Describing clothing by semantic attributes. In *ECCV*.

[3] Qiang Chen, Junshi Huang, Rogerio Feris, Lisa M Brown, Jian Dong, and Shuicheng Yan. 2015. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *CVPR*.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

[5] Wei Di, Catherine Wah, Arpit Bhardwaj, Robinson Piramuthu, and Neel Sundaresan. 2013. Style finder: fine-grained clothing style detection and retrieval. In *CVPR Workshops*.

[6] Jianlong Fu, Jinqiao Wang, Zechao Li, Min Xu, and Hanqing Lu. 2012. Efficient clothing retrieval with semantic-preserving visual phrases. In *ACCV*.

[7] Ross Girshick. 2015. Fast r-cnn. In *ICCV*.

[8] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and others. 2015. Spatial transformer networks. In *NIPS*.

[9] Yushi Jing, David Liu, Dmitry Kislyuk, Andrew Zhai, Jiajing Xu, Jeff Donahue, and Sarah Tavel. 2015. Visual search at pinterest. In *KDD*.

[10] Yannis Kalantidis, Lyndon Kennedy, and Li-Jia Li. 2013. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *ICMR*.

[11] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. 2015. Where to buy it: matching street clothing photos in online shops. In *ICCV*.

[12] M Hadi Kiapour, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. 2014. Hipster wars: discovering elements of fashion styles. In *ECCV*.

[13] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1, 4 (1989).

[14] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. 2015. Human parsing with contextualized convolutional neural network. In *ICCV*.

[15] Kevin Lin, Huei-Fang Yang, Jen-Hao Hsiao, and Chu-Song Chen. 2015. Deep learning of binary hash codes for fast image retrieval. In *CVPR Workshop*.

[16] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. 2012. Street-to-shop: cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*.

[17] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*.

[18] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2016. Fashion landmark detection in the wild. In *ECCV*.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

[21] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. 2015. Neuroaesthetics in fashion: modeling the perception of beauty. In *CVPR*.

[22] Edgar Simo-Serra and Hiroshi Ishikawa. 2016. Fashion style in 128 floats: joint ranking and classification using weak data for feature extraction. In *CVPR*.

[23] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[24] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*.

[25] Alexander Toshev and Christian Szegedy. 2014. Deeppose: human pose estimation via deep neural networks. In *CVPR*.

[26] Xianwang Wang and Tong Zhang. 2011. Clothes search in consumer photos via color matching and attribute learning. In *ACM MM*.

[27] Kota Yamaguchi, Tamara L Berg, and Luis E Ortiz. 2014. Chic or Social: visual popularity analysis in online fashion networks. In *ACM MM*.

[28] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. 2013. Paper doll parsing: retrieving similar styles to parse clothing items. In *ICCV*.

[29] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. 2012. Parsing clothing in fashion photographs. In *CVPR*.

[30] Wei Yang, Ping Luo, and Liang Lin. 2014. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*.

[31] Fisher Yu and Vladlen Koltun. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* (2015).

[32] Xiangxin Zhu and Deva Ramanan. 2012. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*.

[33] C Lawrence Zitnick and Piotr Dollár. 2014. Edge boxes: Locating object proposals from edges. In *ECCV*.