

A Benchmark for Inpainting of Clothing Images with Irregular Holes

Furkan Kinli¹, Barış Özcan², and Furkan Kıracı³

Özyegin University, Istanbul, Turkey

{¹furkan.kinli, ³furkan.kirac}@ozyegin.edu.tr

²baris.ozcan.10097@ozu.edu.tr



Fig. 1: Masked & generated images of our model employing dilated partial convolutions for clothing image inpainting.

Abstract. Fashion image understanding is an active research field with a large number of practical applications for the industry. Despite its practical impacts on intelligent fashion analysis systems, clothing image inpainting has not been extensively examined yet. For that matter, we present an extensive benchmark of clothing image inpainting on well-known fashion datasets. Furthermore, we introduce the use of a dilated version of partial convolutions, which efficiently derive the mask update step, and empirically show that the proposed method reduces the required number of layers to form fully-transparent masks. Experiments show that dilated partial convolutions (DPCConv) improve the quantitative inpainting performance when compared to the other inpainting strategies, especially it performs better when the mask size is 20% or more of the image.

Keywords: image inpainting, fashion image understanding, dilated convolutions, partial convolutions

1 Introduction

Image inpainting is the task of filling the holes in a particular image with some missing regions in such a way that the generated image should be visually plausible and semantically coherent. There are numerous vision applications including object removal, regional editing, super-resolution, stitching and many others that can be practicable by employing different image inpainting strategies. In the literature, the most prominent studies [2, 41, 19, 57, 33, 58, 38] proposing different strategies for solving inpainting problems have focused on mostly natural scene understanding, street view understanding and face completion tasks.

Fashion image understanding is an active research field that has enormous potential of practical applications in the industry. With the achievements of deep learning-based solutions [16, 8, 15, 43] and increasing the number of datasets representing more real-world-like cases [44, 62, 36, 7], different solutions have been proposed for various image-related vision tasks in fashion domain such as clothing category classification [36, 7, 50, 20, 31], attribute recognition [36, 7, 17, 53, 59], clothing segmentation [32, 7, 37, 23, 29], clothing image retrieval [22, 5, 18, 51, 39, 27] and clothing generation [46, 17, 12, 30, 21, 61, 11, 55, 35, 13]. At this point, the recent breakthroughs in deep generative learning solutions [10, 25, 24] lead to arise some opportunities for the applications in fashion domain including designing new clothing items for recommendation systems [46, 17], virtual try-on systems [12, 30, 21] and fashion synthesis [61, 11, 35, 55, 13]. Despite these extensive studies on different tasks in fashion domain, image inpainting has not been extensively examined yet. Considering the practical impacts of achieving this task on real-world applications of fashion analysis, we present an extensive benchmark for clothing image inpainting by employing a generative approach that recovers the irregular missing regions in the set of clothing items.

To extend the practical advantages of this solution for fashion domain, we need further investigation of the drawbacks of the current fashion analysis systems working on real-world cases. Due to possible deformations and occlusions in real-world scenarios, understanding fashion images is a challenging task to make successful inferences. For instance, first, attribute recognition models trained with commercial images often fail to successfully infer the attributes of social media images (in other words, consumer images), due to the occlusions that appears on the parts representing the actual attribute of a particular clothing. Next, overlapping items in a single clothing image builds natural occlusion scenarios for both items when segmentation is applied to the image. Although such segmentation models [9, 52, 4] have the ability to infer the possible locations of the occluded region of an object, filling this region in a visually plausible and semantically correct way in order to analyze the segmented clothing items is yet to be achieved. Moreover, for a visual recommendation system working with real-world clothing images, any deformation on the extracted clothing items can change the possible combinations of it with the other items. Based on such drawbacks on fashion analysis systems, applying different inpainting strategies may be practical for understanding fashion images better. Therefore, we want



Fig. 2: Difficult scenarios for fashion image understanding due to the occlusions (natural, multiple persons, multiple clothing items)

to contribute to the fundamental research effort of solving inpainting problems by redirecting it to fashion domain.

In image inpainting literature, there are several strategies that use image statistics or deep learning-based solutions. PatchMatch [2] is one of the most prominent strategies relying on the image statistics, and it basically searches for the best fitting patch to fill the rectangular-formed missing parts in the images. Although it is possible to generate visually plausible results by using this inpainting strategy, it cannot achieve semantic coherence in hard scenarios since it only depends on the statistics of available parts in the image. On the other hand, deep learning-based solutions are more suitable for image inpainting since deep neural networks can inherently learn the hidden representations by preserving semantic priors. Earlier studies [41,19,57] try to solve the problem

of initialization of the pixels in the missing parts to condition the output by assigning a fixed value for these pixels. In these studies, the results have the visual artifacts, and need some additional post-processing methods (*e.g.* fast marching [48], Poisson image blending [42] and the refinement network [57]) to refine them. Partial convolutional mechanism [33] addresses this limitation by conditioning the only valid pixels. To achieve this, convolutional layers are masked and re-weighted, and then the mask is updated in such a way that progressively filling up the hole pixels. Recently, Nazeri *et al.* [38] considers inpainting problem as image completion by predicting the structure (*i.e.* edge maps) of the main content in the image. Apart from these studies, we focus on increasing the receptive field size of low-resolution feature maps of partial convolutions, and thus, we propose dilated partial convolutions whose the kernel of partial convolution [33] is dilated by a given parameter, and capable of gathering more information from near-to-hole regions in order to update the masks more efficiently without requiring to decrease their spatial dimensionality up to very low-resolution.

In summary, our main contributions are as follows:

- We introduce an extensive benchmark of clothing image inpainting on a variety of challenging datasets including FashionGen [44], FashionAI [62], DeepFashion [36] and DeepFashion2 [7], and attempt to redirect the fundamental research efforts on image inpainting problems to fashion domain.
- To the best of our knowledge, this is the first study to mention the practical usefulness of achieving visually plausible and semantically coherent inpainting of fashion images for industrial applications.
- We present the idea of using dilated version of partial convolutions for image inpainting tasks, where the dilated input window for partial convolutions derives more efficient mask update step.
- We empirically show that dilated partial convolutions reduce the number of layers that requires to lead to a mask without any holes, and thus, it makes possible to achieve better inpainting quality without reducing the spatial dimensionality of encoder output.

2 Related Work

Image inpainting is a challenging task, and it has been extensively studied for a decade in different vision-related research fields, especially on natural scene understanding and face recognition. Several traditional inpainting strategies such as [1,3,6,14,48] try to synthesize texture information in the holes by employing available image statistics. However, these methods often fail to achieve inpainting the images without any artifacts since using only available image statistics may mislead the results about the texture information between holes and non-hole regions when it varies. Barnes *et al.* [2] proposes a fast algorithm, namely *PatchMatch*, which iteratively searches for the best fitting patches to the missing parts of the image. Although this method can produce visually plausible results in a faster way, it still lacks of producing semantically coherent results, and far from being suitable for real-time image processing pipeline.

In deep learning-based solutions, earlier approaches use a constant value for initializing the holes before passing throughout the network. First, *Context Encoders* [41] employs an Auto-encoder architecture with adversarial training strategy in order to fill a central large hole in the images. Yang *et al.* [54] extends the idea in [41] with post-processing the output that considers only the available image statistics. Song *et al.* [47] proposes a robust training scheme in coarse-to-fine manner. Iizuka *et al.* [19] introduces a different adversarial training strategy to provide both local and global consistency in the generated images. Also, in [19], it is demonstrated that sufficiently larger receptive fields work well in image inpainting tasks, and dilated convolution is adopted for increasing the size of the receptive fields. Yu *et al.* [57] improves the idea of using dilated convolutions in inpainting task by adding contextual attention mechanism on top of low-resolution feature maps for explicitly matching and attending to relevant background patches. Liu *et al.* [33] presents partial convolutions where the convolution is masked and re-weighted to be conditioned on only valid pixels. Yu *et al.* [58] proposes gated convolutions that have the ability to learn the features dynamically at each spatial location across all layers, and this mechanism improves the color consistency and semantic coherence in the generated images. Liu *et al.* [34] employs coherent semantic attention and consistency loss to the refinement network to construct the correlation between the features of hole regions, even in deeper layers. Kinli *et al.* [28] observes the effect of collaborating with textual features extracted by image descriptions on inpainting performance. Nazeri *et al.* [38] proposes a novel method that predicts the image structure of the missing region in the form of edge maps, and these predicted edge maps are passed to the second stage to guide the inpainting.

3 Methodology

Our proposed model enhances the capability of partial convolutions [33], which alters this mechanism by adding a dilation factor to its convolutional filters, and employing self-attention to the decoder part of our model. In this section, we first explain the reasoning behind dilated partial convolutions and present its formulation. Then, we describe our architecture, and lastly discuss the loss functions of proposed model.

3.1 Dilated Partial Convolutions

Unlike in natural images where spatially-near pixels yield a larger correlation, for clothing image inpainting, the correlated pixels may be far apart in a particular image (*e.g.* a pattern on a shirt can cover a larger area of the image, and more importantly, this pattern may not be spatially continuous in such scenarios with occlusions or deformations). Even though both partial and dilated convolutional layers prove that they can produce visually plausible results in inpainting tasks, they have their own particular shortcomings. When partial convolutions are employed, the layers cannot gather the information from the correlation of far apart

pixels as in dilated convolutions, on the other hand, the network does not only focus on non-hole regions of the images by using only dilated convolutions [56]. We address both these shortcomings by introducing the use of dilated partial convolutions where the input window is dilated by a given parameter, and the mask of partial convolution is applied afterwards. This allows the network to focus on non-hole regions, and it has larger receptive fields to utilize correlations of far apart pixels. More importantly, as stated in [33], the consecutive layers of partial convolutions will eventually lead to a mask without any holes depending on the input mask, and we have empirically proven that dilated partial convolutions reduce the number of required layers to achieve this, and the input mask covers up a larger percentage of the input image in earlier part of the network. The discussion and detailed empirical results can be found in Discussion part. The formal definition of dilated partial convolutions can be formulated as follows:

$$\mathcal{M} = \sum_{m=-M}^M \sum_{n=-N}^N m(x - ln, y - lm) \quad (1)$$

$$Z = \frac{(2M + 1) \times (2N + 1)}{\mathcal{M}} \quad (2)$$

$$(f_m * g_l)(x, y) = \frac{\sum_{m=-M}^M \sum_{n=-N}^N f(x - ln, y - lm)g(n, m)m(x - ln, y - lm)}{Z} \quad (3)$$

where f , g and m represents the input, the convolutional kernel with a size of $(2M + 1) \times (2N + 1)$ and the corresponding mask, respectively. Eq. (2) has essentially the same scaling factor as in [33], but modified to be applicable to dilated partial convolutions. Note that when the dilation factor $l = 1$, Eq. (3) becomes the partial convolution without any dilation. Mask update is calculated in the same manner as in partial convolutions, and can be formulated as follows:

$$m' = \begin{cases} 1, & \text{if } \mathcal{M} > 1 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where the updated mask m' gets closer to become fully-transparent (*i.e.* all pixels become ones) after a certain number of layers. Dilated partial convolutions allows m' to become fully-transparent in earlier layers when compared to partial convolutions.

3.2 Model Architecture

We have designed a similar *U-Net-like* [43] architecture as in [33], where all low resolution layers are replaced with dilated partial convolutions while the first four layers are left as partial convolutions. The binary mask in the first layer is defined

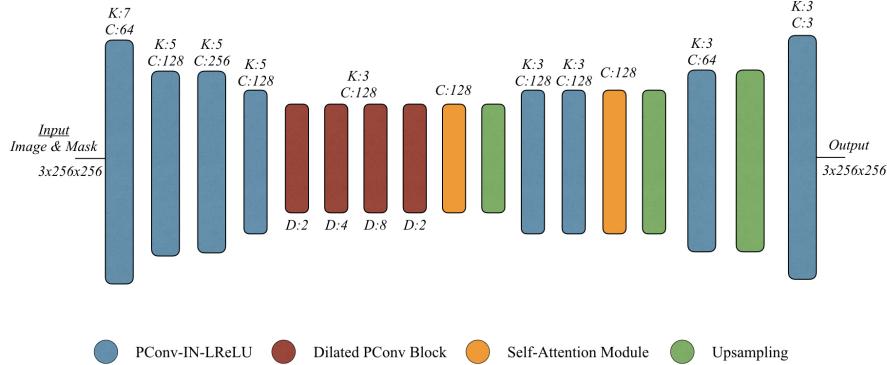


Fig. 3: Overview of our architecture design.

as the input corruption mask. Moreover, in the decoder stage, we have used self-attention module [49] to make use of spatially distant but related features. All of the skip connections are concatenated with the layer on corresponding level, instead of adding to them, and also dilated partial convolutions are residually-connected. The dilation rate is progressively increased up to the last dilated convolutional layer (*i.e.* multi-scale context aggregation), as in [56]. As working on the fashion datasets, the input sizes for all models are picked as 256×256 . Overall architecture design can be seen in Figure 3.

3.3 Loss Functions

We have followed a similar loss function scheme to [33], but with slight differences. The total loss function is introduced in Eq. (5), where \mathcal{L}_{pixel} is the pixel loss, \mathcal{L}_{style} is the style loss, \mathcal{L}_{adv} is the adversarial loss and \mathcal{L}_{tv} is the total variation (TV) loss. We found that the coefficients in Eq. (5) work better than the ones in [33] in our experimental settings. Moreover, the perceptual loss is substituted with an adversarial loss since we experimentally found that the benefits of the perceptual loss do not afford its computational cost, and also we want to take the advantage of adversarial training for our model.

$$\mathcal{L}_{total} = 10\mathcal{L}_{pixel} + 120\mathcal{L}_{style} + 10^{-3}\mathcal{L}_{adv} + 10^{-4}\mathcal{L}_{tv} \quad (5)$$

Pixel loss \mathcal{L}_{pixel} is the sum of ℓ_1 losses of the hole and non-hole regions between the output image \mathbf{I}_{out} and the ground truth image \mathbf{I}_{gt} , as given in (6), where $N_{\mathbf{I}_{gt}} = C \times H \times W$ as C, H and W are the input dimensions, and \hat{m} denotes the initial mask on the input.

$$\mathcal{L}_{pixel} = \frac{1}{N_{\mathbf{I}_{gt}}} (\left\| (1 - \hat{m}) \odot (\mathbf{I}_{out} - \mathbf{I}_{gt}) \right\|_1 + \left\| \hat{m} \odot (\mathbf{I}_{out} - \mathbf{I}_{gt}) \right\|_1) \quad (6)$$

The total style loss \mathcal{L}_{style} is calculated by summing the style losses for \mathbf{I}_{out} and \mathbf{I}_{comp} , denoted as $\mathcal{L}_{style_{out}}$ and $\mathcal{L}_{style_{comp}}$, respectively (Eq. (7)). To obtain the composite image \mathbf{I}_{comp} , all non-hole pixels in the output image are replaced with the ground truth pixels. Style loss requires to calculate the activation maps $\Psi^{\mathbf{I}_*}_p$ of pre-trained VGG-16 with respect to the input image \mathbf{I}_* , where p denotes the layer index. For a given layer index p , the normalization factor is defined as $K_p = \frac{1}{C_p H_p W_p}$.

$$\begin{aligned} \mathcal{L}_{style} = & \sum_{p=0}^{P-1} \frac{1}{C_p^2} \left\| K_p ((\Psi^{\mathbf{I}_{out}}_p)^T (\Psi^{\mathbf{I}_{out}}_p) - (\Psi^{\mathbf{I}_{gt}}_p)^T (\Psi^{\mathbf{I}_{gt}}_p)) \right\|_1 + \\ & \sum_{p=0}^{P-1} \frac{1}{C_p^2} \left\| K_p ((\Psi^{\mathbf{I}_{comp}}_p)^T (\Psi^{\mathbf{I}_{comp}}_p) - (\Psi^{\mathbf{I}_{gt}}_p)^T (\Psi^{\mathbf{I}_{gt}}_p)) \right\|_1 \end{aligned} \quad (7)$$

The adversarial loss given in Eq. (8) is adopted from [10], where \mathcal{G} is the generator network, \mathcal{D} is the discriminator network, and \mathbf{I}^M is the masked input image. The discriminator \mathcal{D} network is a simple CNN with a depth of 5, and trained to optimize the loss function given in Eq. (9). To train the discriminator better, we flipped the labels with the probability of 0.1 to make the labels noisy for the discriminator, and also we applied label smoothing where, for each instance, if it is real, then replace its label with a random number between 0.7 and 1.2, otherwise, replace it with 0.0 and 0.3. Lastly, we tried to apply random dropout for the decoder part of our model.

$$\mathcal{L}_{adv_G} := \mathbb{E} \left[(\mathcal{D}(\mathcal{G}(\mathbf{I}^M)) - 1)^2 \right] \quad (8)$$

$$\mathcal{L}_{adv_D} := \mathbb{E} \left[\mathcal{D}(\hat{\mathbf{I}})^2 \right] + \mathbb{E} \left[(\mathcal{D}(\mathbf{I}) - 1)^2 \right] \quad (9)$$

Lastly, total variation loss L_{tv} , as given in (10), enforces the spatial continuity on the generated images, where R is the region after applying a 1-pixel dilation to a hole region in the input image.

$$\mathcal{L}_{tv} = \sum_{(i,j) \in R, (i,j+1) \in R} \frac{\|\mathbf{I}_{comp}^{i,j+1} - \mathbf{I}_{comp}^{i,j}\|_1}{N_{\mathbf{I}_{comp}}} + \sum_{(i,j) \in R, (i+1,j) \in R} \frac{\|\mathbf{I}_{comp}^{i+1,j} - \mathbf{I}_{comp}^{i,j}\|_1}{N_{\mathbf{I}_{comp}}} \quad (10)$$

4 Experiments

In this study, we introduce an extensive benchmark on fashion image inpainting, and also propose enhanced version of partial convolutions, namely *dilated partial convolutions*. We have conducted our experiments on four different well-known fashion datasets, which are FashionGen [44], FashionAI [62], DeepFashion [36] and DeepFashion2 [7].

4.1 Experimental Setup

Training details In our experiments, we used Adam optimizer [26] with $\beta_1 = 0.5$ and $\beta_2 = 0.9$ for both generator and discriminator networks. The initial learning rate for generator network is 2×10^{-4} , and for discriminator network, is 10^{-4} . We trained our model for ~ 120.000 steps for each dataset with batch size of 64. We only applied horizontal flipping to the input images, no other data augmentation technique is applied. For DeepFashion and DeepFashion2 datasets, the size of images varies, so we use random cropping (central cropping for testing) and resizing to obtain the size of 256×256 for training images. We implemented our framework with PyTorch library [40], and used 2x NVIDIA Tesla V100 GPUs for our training. Source code: <https://github.com/birdortyedi/fashion-image-inpainting>

Datasets To create an extensive benchmark for fashion image inpainting, we picked four common fashion datasets to use in our experiments. Before training, we prepared the datasets to the image inpainting setup by applying inpainting masks¹ to the images, which erase some parts randomly. To make sure that the images have some holes on clothing parts, we generated the masks with a heuristic, where at least a small portion of the clothing parts has been erased by the mask. Then, we run our experiments for each method on each dataset separately.

FashionGen (FG) [44] contains several different fashion products collected from an online platform selling the luxury goods from independent designers. Each product is represented by an image, the description, its attributes, and relational information defined by professional designers. We extract the instances belonging to these categories from the dataset. At this point, training set has 200K clothing images from 9 different categories (*top, sweater, pant, jean, shirt, dress, short, skirt, coat*), and validation set has 30K clothing images.

FashionAI (FAI) [62] is introduced in FashionAI Global Challenge 2018 by The Vision & Beauty Team of Alibaba Group and Institute of Textile & Clothing of The Hong Kong Polytechnic University. For this dataset, there are 180K consumer images of different clothing categories for training, and 10K for testing.

DeepFashion (DF) [36] is a dataset containing $\sim 800K$ diverse fashion images with their rich annotations (*i.e.* 46 categories, 1.000 descriptive attributes, bounding boxes and landmark information) ranging from well-posed product images to real-world-like consumer photos. In our experiments on this dataset, we follow the same procedure as in [36] to split training and testing sets, so we have 207K clothing images for training, and 40K for testing.

DeepFashion2 (DF2) [7] is one of the largest open-source fashion database that contains 491K high-resolution clothing images with 13 different categories and numerous attributes. Similar to DeepFashion, we again follow the procedure in the paper of dataset [7] to split the published version of the data.

¹ <https://github.com/karfly/qd-imd>

Table 1: Quantitative comparison between our proposed model and the state-of-the-art methods on four well-known fashion datasets.

Mask Ratios	[0.0:0.2]				[0.2:0.4]				[0.4:]			
Datasets	FG	FAI	DF	DF2	FG	FAI	DF	DF2	FG	FAI	DF	DF2
ℓ_1 (PM) %	0.66	1.16	1.34	1.32	1.34	2.11	2.03	1.96	3.70	6.12	5.74	5.43
ℓ_1 (PConv) %	0.73	0.92	0.91	0.94	1.18	1.48	1.34	1.47	2.82	4.86	3.76	3.70
ℓ_1 (GConv) %	0.76	0.95	0.93	0.95	1.22	1.58	1.23	1.36	2.99	5.20	3.65	3.62
ℓ_1 (DPCConv) %	0.70	0.91	0.87	0.92	1.14	1.39	1.20	1.33	2.61	4.03	3.36	3.38
PSNR (PM)	16.14	12.21	12.30	12.55	14.71	11.23	11.37	11.49	13.36	9.74	9.96	10.09
PSNR (PConv)	16.04	12.33	12.43	12.73	15.04	11.49	11.66	11.94	13.91	11.11	11.47	11.89
PSNR (GConv)	15.98	12.20	12.34	12.66	14.86	11.42	11.61	11.84	13.79	11.02	11.39	11.77
PSNR (DPCConv)	16.23	12.27	12.56	12.89	15.15	11.81	11.88	12.12	14.16	11.63	11.84	12.28
SSIM (PM)	0.891	0.773	0.784	0.804	0.816	0.737	0.744	0.769	0.738	0.622	0.671	0.677
SSIM (PConv)	0.851	0.784	0.777	0.804	0.770	0.735	0.729	0.733	0.721	0.649	0.689	0.688
SSIM (GConv)	0.845	0.751	0.768	0.796	0.783	0.723	0.731	0.741	0.717	0.624	0.670	0.681
SSIM (DPCConv)	0.855	0.782	0.794	0.811	0.824	0.739	0.750	0.777	0.766	0.683	0.717	0.719
FID (PM)	2.99	5.44	4.36	5.78	6.21	11.98	11.65	12.30	25.92	28.91	27.94	29.74
FID (PConv)	3.76	6.04	5.93	6.22	5.56	11.05	10.36	11.28	23.47	27.42	24.39	27.52
FID (GConv)	3.57	5.86	5.63	6.18	5.28	10.84	10.12	11.01	23.18	27.19	24.21	26.95
FID (DPCConv)	3.18	5.24	4.66	5.48	5.20	10.08	9.26	10.21	22.82	25.55	23.94	25.62

4.2 Quantitative Results

As is the case with the previous studies [57,33,58,34], we evaluate the performance of our model with four different metrics, which are ℓ_1 error percentage, SSIM [60], PSNR and FID [45]. Then, we compare the quantitative performance of DPCConv with PatchMatch (PM) [2] as a non-learning method, Partial Convolutions (PConv) [33] and Gated Convolutions (GConv) [58], as learning-based methods. We used the third-party implementations for PM² and GConv³, but reimplemented PConv according to the layer implementation⁴ and the architecture details referred in their paper [33]. The quantitative results of both our model and the state-of-the-art methods/models are shown in Table 1. The observations are as follows: (1) Dilated partial convolutions are more robust to the changes in the mask ratios. (2) According to SSIM metric, although PM performs well in the cases where the mask ratio is smaller, the performances on all measurements significantly decrease for more complex cases. (3) The performances of all methods are very similar on less complex dataset (*i.e.* FashionGen), whereas the impact of our proposed model can be clearly observed as mask ratio increases in the settings of more complex datasets (*i.e.* DeepFashion and DeepFashion2). (4) Using dilated version of partial convolutions in image inpainting improves the overall performance without reducing the spatial dimensionality of the encoder output. (5) Overall, our model, namely *DPCConv*, outperforms the current inpainting strategies on four well-known fashion datasets.

² <https://github.com/MingtaoGuo/PatchMatch>

³ https://github.com/avalonstrel/GatedConvolution_pytorch

⁴ <https://github.com/NVIDIA/partialconv>

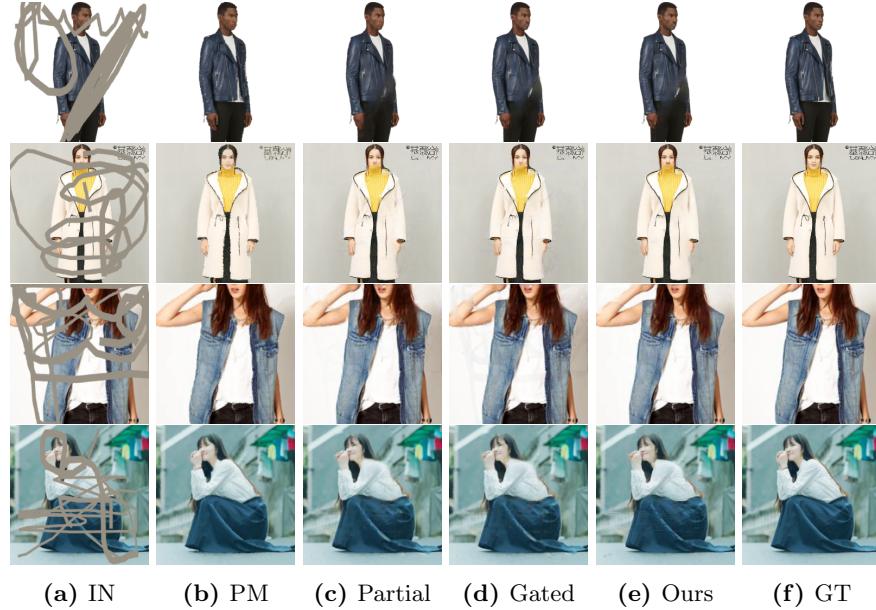


Fig. 4: Comparison the results of our proposed model and the state-of-the-art methods.

4.3 Qualitative Results

Next, we compare the visual compatibility of the results of our model with the other state-of-the-art methods [2,33,58]. Figure 4 shows the results on four well-known fashion datasets [44,62,36,7]. We apply the same settings with ours to the training of all methods, and do not perform any post-processing for the outputs. The results demonstrate that all inpainting strategies can produce visually plausible and semantically coherent clothing images, and the visual outputs of these inpainting strategies can be utilized in order to increase the effectiveness of fashion image understanding solutions (*e.g.* removing the disrupted regions and inpainting them). However, DPCConv accomplish it by employing a shallower network architecture, and taking advantage of the efficient mask update step of dilated partial convolutions. Furthermore, to compare the strategies, we can say that (1) DPCConv and PConv seem to produce very similar outputs, but when looking into the details, DPCConv preserves the visual coherence better (*e.g.* In Figure 4, the collar of the coat in row 2 & the head of woman in row 4). (2) PM cannot produce smooth outputs in contrast to the other methods. (3) GConv cannot fill the holes by preserving the semantic coherence, and residue of the input mask can be still seen in GConv outputs.

Ablation Study: We conduct additional experiments to observe the effect of using dilated convolutions at certain stages of the inpainting networks. The first

Table 2: The evaluation of the usage of dilation in certain stages of the networks for different inpainting strategies.

Mask Ratios	[0.0:0.2]				[0.2:0.4]				[0.4:]			
Datasets	FG	FAI	DF	DF2	FG	FAI	DF	DF2	FG	FAI	DF	DF2
ℓ_1 (DGConv) %	0.72	0.92	0.90	0.92	1.17	1.43	1.21	1.33	2.68	4.21	3.41	3.47
ℓ_1 (DPCConv †) %	0.62	1.02	0.94	1.20	1.21	1.72	1.86	2.04	3.18	5.16	5.66	5.39
ℓ_1 (DPCConv) %	0.70	0.91	0.87	0.92	1.14	1.39	1.20	1.33	2.61	4.03	3.36	3.38
PSNR (DGConv)	16.11	12.31	12.54	12.91	15.09	11.76	11.92	12.10	14.12	11.56	11.70	11.92
PSNR (DPCConv †)	15.96	12.14	12.32	12.61	14.75	11.27	11.39	11.61	13.40	10.97	11.04	11.19
PSNR (DPCConv)	16.23	12.27	12.56	12.89	15.15	11.81	11.88	12.12	14.16	11.63	11.84	12.28
SSIM (DGConv)	0.858	0.779	0.788	0.802	0.818	0.736	0.754	0.774	0.750	0.688	0.712	0.704
SSIM (DPCConv †)	0.869	0.773	0.762	0.788	0.791	0.717	0.728	0.735	0.719	0.626	0.664	0.679
SSIM (DPCConv)	0.855	0.782	0.794	0.811	0.824	0.739	0.750	0.777	0.766	0.683	0.717	0.719
FID (DGConv)	3.38	5.49	4.90	5.89	5.06	10.36	9.53	10.46	22.87	25.71	24.02	25.88
FID (DPCConv †)	3.96	6.01	6.73	5.82	6.16	11.70	11.99	11.38	24.02	26.89	24.54	26.24
FID (DPCConv)	3.18	5.24	4.66	5.48	5.20	10.08	9.26	10.21	22.82	25.55	23.94	25.62

one is called $DPCConv^\dagger$ whose layers before dilated partial convolutional block also have dilation of 2 in their kernels. The latter, namely $DGConv$, is the same architecture with our model, but employs gated convolutions, instead of partial convolutions. Table 2 demonstrates the quantitative evaluation of these models and ours. The observations are concluded as follows: (1) Using dilation in every stage of the network without applying multi-scale context aggregation strategy has negative effect on the performance of our model. (2) Due to the computational burden, applying multi-scale context aggregation to each stage is not feasible for this task. (3) Gated convolutions with dilation on the kernels leading to the lower-level feature maps shows a similar impact on the qualitative results, which DPCConv does it on PConv. (4) However, DPCConv still mostly outperforms DGConv on different mask ratios on four well-known fashion datasets.

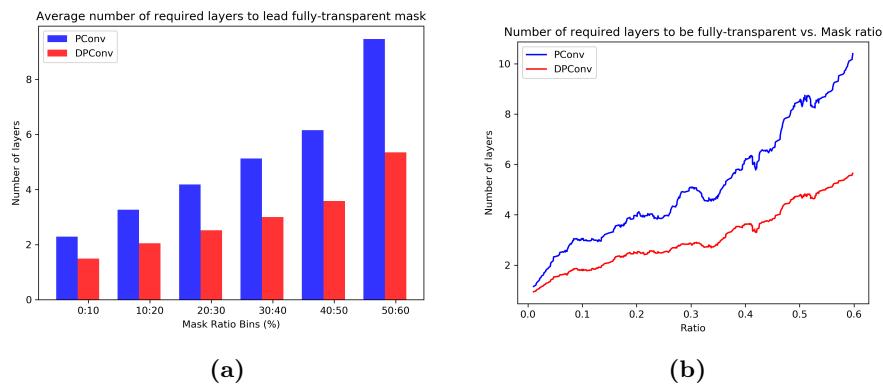


Fig. 5: Experimental results of analyzing the behavior of partial convolutions and dilated partial convolutions given random masks.

4.4 The Effect of Dilation in Partial Convolutions

Dilated partial convolutions complete the masks to become fully-transparent (*i.e.* masks without any zeros) throughout the network with less number of consecutive layers, when compared to partial convolutions. To empirically prove this, we designed an experiment to analyze the behaviour of both layers given different input masks. In this experiment, we used 12.000 random masks with ~ 2.000 masks for each 10% range of mask ratios from 0% up to 60%. For each range of mask ratios, the average number of layers required to obtain a fully-transparent mask is calculated, and the maximum number of layers that can be stacked is limited to 20. As illustrated in Figure 5a, dilated partial convolutions can reach fully-transparent masks 2.1 layers earlier on average (*i.e.* $\sim 15\%$ less layers). Figure 5b emphasizes the exponential growth of the required number of layers with respect to all masks in the case of using partial convolutions or dilated partial convolutions. The result of this experiment shows that the dilated version of partial convolutions has a practical impact of **leading fully-transparent masks in higher resolution without requiring to go deeper throughout the network**, and thus it leads to a more efficient mask update step and faster learning process. Note that PCConv decreases the feature map to 2×2 in its decoder part while DPConv starts to upsampling at the feature maps sized 32×32 .

4.5 Practical Usage in Fashion Domain

Image inpainting can improve the performance of fashion image understanding solutions when applied as pre-process or post-process. To give a brief idea about

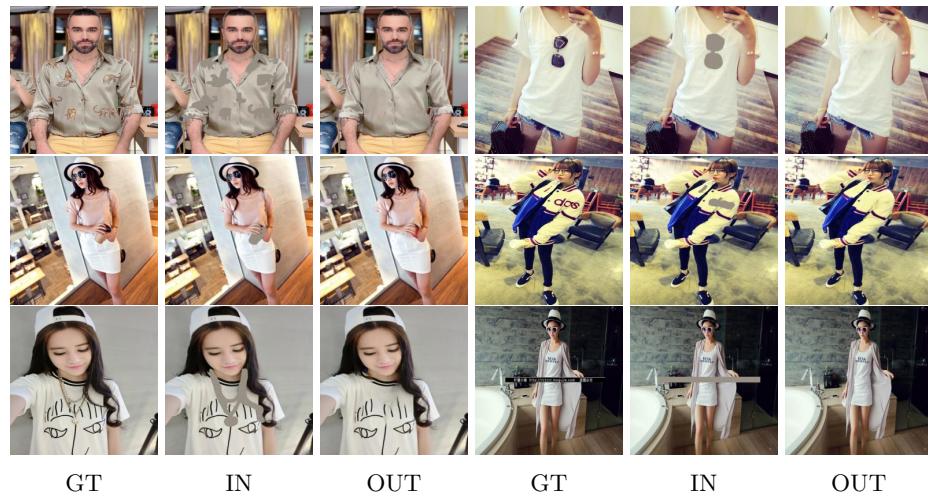


Fig. 6: Example inpainting results of the images containing some overlapping or disruptive parts on clothing items.



Fig. 7: More inpainting results of DPCConv on four different datasets. Zoom in for better view.

how it can be useful for such systems, we picked a number of images that contain different items overlapping the clothing items or that have some disruptive parts (*e.g.* banner, logo). We have manually created the masks that remove these items/parts from the images. Figure 6 demonstrates the inpainting results of picked images where DPCConv trained on DeepFashion2 dataset is employed for testing. At this point, we showed that clothing image inpainting can be useful for fashion editing (*e.g.* removing accessories like eye-glasses and necklace, logos, banners or non-clothing items, even changing the design of clothing), and it makes the main-stream fashion image understanding solutions work better.

5 Conclusion

In this study, we present an extensive benchmark for clothing image inpainting, which may be practical for industrial applications in fashion domain. Qualitative and quantitative comparisons demonstrate that proposed method improves image inpainting performance when compared to the previous state-of-the-art methods, and produce visually plausible and semantically coherent results for clothing images. Overall performances of inpainting strategies proves that AI-based fashion image understanding solutions can employ inpainting to their pipeline in order to improve the general performance.

References

1. Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels. *IEEE Transactions on Image Processing* **10**(8), 1200–1211 (Aug 2001). <https://doi.org/10.1109/83.935036>
2. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* **28**(3), 24 (2009)
3. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. p. 417424. SIGGRAPH 00, ACM Press/Addison-Wesley Publishing Co., USA (2000). <https://doi.org/10.1145/344779.344972>
4. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
5. Di, W., Wah, C., Bhardwaj, A., Piramuthu, R., Sundaresan, N.: Style finder: Fine-grained clothing style detection and retrieval. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (June 2013)
6. Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. p. 341346. SIGGRAPH 01, Association for Computing Machinery, New York, NY, USA (2001). <https://doi.org/10.1145/383259.383296>
7. Ge, Y., Zhang, R., Wu, L., Wang, X., Tang, X., Luo, P.: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images (2019)
8. Girshick, R.: Fast r-cnn. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. pp. 1440–1448 (Dec 2015). <https://doi.org/10.1109/ICCV.2015.169>
9. Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., He, K.: Detectron. <https://github.com/facebookresearch/detectron> (2018)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc. (2014), <http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf>
11. Gunel, M., Erdem, E., Erdem, A.: Language guided fashion image manipulation with feature-wise transformations. In: *First Workshop on Computer Vision in Art, Fashion and Design* – in conjunction with ECCV 2018 (2018)
12. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7543–7552 (June 2018). <https://doi.org/10.1109/CVPR.2018.00787>
13. Han, X., Wu, Z., Huang, W., Scott, M.R., Davis, L.S.: Finet: Compatible and diverse fashion image inpainting. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 4481–4491 (2019)
14. Hays, J., Efros, A.A.: Scene completion using millions of photographs. *Commun. ACM* **51**(10), 8794 (Oct 2008). <https://doi.org/10.1145/1400181.1400202>
15. He, K., Gkioxari, G., Dollr, P., Girshick, R.: Mask r-cnn. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. pp. 2980–2988 (Oct 2017). <https://doi.org/10.1109/ICCV.2017.322>

16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (June 2016). <https://doi.org/10.1109/CVPR.2016.90>
17. Hsiao, W.L., Grauman, K.: Creating capsule wardrobes from fashion images. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (Jun 2018). <https://doi.org/10.1109/cvpr.2018.00748>
18. Huang, J., Feris, R.S., Chen, Q., Yan, S.: Cross-domain image retrieval with a dual attribute-aware ranking network. In: Proceedings of the IEEE international conference on computer vision. pp. 1062–1070 (2015)
19. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Trans. Graph. **36**(4) (Jul 2017). <https://doi.org/10.1145/3072959.3073659>
20. Inoue, N., Simo-Serra, E., Yamasaki, T., Ishikawa, H.: Multi-label fashion image classification with minimal human supervision. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (Oct 2017)
21. Jae Lee, H., Lee, R., Kang, M., Cho, M., Park, G.: La-viton: A network for looking-attractive virtual try-on. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (Oct 2019)
22. Jagadeesh, V., Piramuthu, R., Bhardwaj, A., Di, W., Sundaresan, N.: Large scale visual recommendations from street fashion images. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 19251934. KDD 14, Association for Computing Machinery, New York, NY, USA (2014). <https://doi.org/10.1145/2623330.2623332>
23. Ji, W., Li, X., Zhuang, Y., Bourahla, O.E.F., Ji, Y., Li, S., Cui, J.: Semantic locality-aware deformable network for clothing segmentation. In: IJCAI. pp. 764–770 (2018)
24. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018), <https://openreview.net/forum?id=Hk99zCeAb>
25. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2019). <https://doi.org/10.1109/cvpr.2019.00453>
26. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (12 2014)
27. Kinli, F., Özcan, B., Kirac, F.: Fashion image retrieval with capsule networks. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
28. Kinli, F., Özcan, B., Kirac, F.: Description-aware fashion image inpainting with convolutional neural networks in coarse-to-fine manner. In: Proceedings of the 2020 6th International Conference on Computer and Technology Applications. p. 7479. ICCTA 20 (2020). <https://doi.org/10.1145/3397125.3397155>
29. Korneliussen, M., Martinsson, J., Mogren, O.: Generative modelling of semantic segmentation data in the fashion domain. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (Oct 2019)
30. Kubo, S., Iwasawa, Y., Suzuki, M., Matsuo, Y.: Uvton: Uv mapping to consider the 3d structure of a human in image-based virtual try-on network. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (Oct 2019)
31. Knl, F., Kra, F.: Fashioncapsnet: Clothing classification with capsule networks. Journal of Informatics Technologies **13**, 87 – 96 (2020). <https://doi.org/10.17671/gazibtd.580222>

32. Liang, X., Lin, L., Yang, W., Luo, P., Huang, J., Yan, S.: Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval. *IEEE Transactions on Multimedia* **18**(6), 1175–1186 (2016)
33. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: The European Conference on Computer Vision (ECCV) (2018)
34. Liu, H., Jiang, B., Xiao, Y., Yang, C.: Coherent semantic attention for image inpainting. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
35. Liu, L., Zhang, H., Ji, Y., Wu, Q.M.J.: Toward ai fashion design: An attribute-gan model for clothing match. *Neurocomputing* **341**, 156–167 (2019)
36. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
37. Martinsson, J., Mogren, O.: Semantic segmentation of fashion images using feature pyramid networks. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
38. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: Structure guided image inpainting using edge prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (Oct 2019)
39. Opitz, M., Waltner, G., Possegger, H., Bischof, H.: Bier boosting independent embeddings robustly. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 5199–5208 (Oct 2017). <https://doi.org/10.1109/ICCV.2017.555>
40. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
41. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.: Context encoders: Feature learning by inpainting. In: Computer Vision and Pattern Recognition (CVPR) (2016)
42. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. *ACM Trans. Graph.* **22**(3), 313318 (Jul 2003). <https://doi.org/10.1145/882262.882269>
43. Ronneberger, O., P.Fischer, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). LNCS, vol. 9351, pp. 234–241. Springer (2015), <http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a>, (available on arXiv:1505.04597 [cs.CV])
44. Rostamzadeh, N., Hosseini, S., Boquet, T., Stokowiec, W., Zhang, Y., Jauvin, C., Pal, C.: Fashion-Gen: The Generative Fashion Dataset and Challenge. ArXiv e-prints (Jun 2018)
45. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in neural information processing systems. pp. 2234–2242 (2016)
46. Sbai, O., Elhoseiny, M., Bordes, A., LeCun, Y., Couprie, C.: Design: Design inspiration from generative networks. In: The European Conference on Computer Vision (ECCV) Workshops (September 2018)

47. Song, Y., Yang, C., Lin, Z., Liu, X., Li, H., Huang, Q.: Contextual Based Image Inpainting: Infer, Match and Translate. In: Proceedings of the 15th European Conference on Computer Vision. Computer Vision Foundation, Munich, Germany (Sep 2018)
48. Telea, A.: An image inpainting technique based on the fast marching method. *J. Graphics, GPU, & Game Tools* **9**, 23–34 (2004)
49. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc. (2017), <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>
50. Wang, W., Xu, Y., Shen, J., Zhu, S.C.: Attentive fashion grammar network for fashion landmark detection and clothing category classification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
51. Wang, Z., Gu, Y., Zhang, Y., Zhou, J., Gu, X.: Clothing retrieval with visual attention model. 2017 IEEE Visual Communications and Image Processing (VCIP) (Dec 2017). <https://doi.org/10.1109/vcip.2017.8305144>
52. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. <https://github.com/facebookresearch/detectron2> (2019)
53. Yamaguchi, K., Okatani, T., Sudo, K., Murasaki, K., Taniguchi, Y.: Mix and match: Joint model for clothing and attribute recognition. In: BMVC. vol. 1, p. 4 (2015)
54. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4076–4084 (July 2017). <https://doi.org/10.1109/CVPR.2017.434>
55. Yildirim, G., Jetchev, N., Vollgraf, R., Bergmann, U.: Generating high-resolution fashion model images wearing custom outfits. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (Oct 2019)
56. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: ICLR (2016)
57. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (Jun 2018). <https://doi.org/10.1109/cvpr.2018.00577>
58. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
59. Zhang, S., Song, Z., Cao, X., Zhang, H., Zhou, J.: Task-aware attention model for clothing attribute prediction. *IEEE Transactions on Circuits and Systems for Video Technology* (2019)
60. Zhou Wang, Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612 (April 2004). <https://doi.org/10.1109/TIP.2003.819861>
61. Zhu, S., Fidler, S., Urtasun, R., Lin, D., Loy, C.C.: Be your own prada: Fashion synthesis with structural coherence. 2017 IEEE International Conference on Computer Vision (ICCV) (Oct 2017). <https://doi.org/10.1109/iccv.2017.186>
62. Zou, X., Kong, X., Wong, W., Wang, C., Liu, Y., Cao, Y.: Fashionai: A hierarchical dataset for fashion understanding. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)