Abstract

Image-level corruptions and perturbations degrade the performance of CNNs on different downstream vision tasks. Social media filters are one of the most common resources of various corruptions and perturbations for real-world visual analysis applications. The negative effects of these distractive factors can be alleviated by recovering the original images with their pure style for the inference of the downstream vision tasks. Assuming these filters substantially inject a piece of additional style information to the social media images, we can formulate the problem of recovering the original versions as a reverse style transfer problem. We introduce Contrastive Instagram Filter Removal Network (CIFR), which enhances this idea for Instagram filter removal by employing a novel multi-layer patch-wise contrastive style learning mechanism. Experiments show our proposed strategy produces better qualitative and quantitative results than the previous studies. Moreover, we present the results of our additional experiments for proposed architecture within different settings. Finally, we present the inference outputs and quantitative comparison of filtered and recovered images on localization and segmentation tasks to encourage the main motivation for this problem.
Figure 1. Isolated patch sampling modules for distilling the content and style information. This figure shows the pipeline for only a single level features. The extracted feature maps by IFRNet Encoder ($E$) are first fed into the random sampling modules for the content ($RCS$) and style ($RSS$), separately. After encoding them by corresponding 2-layer MLP modules, ($H_c$) and ($H_s$), the content patches for the input and the output are sent to Content NCE module, and content NCE loss is calculated as proposed in [38]. Moreover, the style patches are extracted by $G$ for calculating the Gram matrix of the encoded features, and style NCE loss learns to select the patch with pure style over the filtered patch.

- We introduce **Contrastive Instagram Filter Removal Network (CIFR)**, which enhances the idea of reverse style transfer for recovering the original images proposed in [27] by adding patch-wise contrastive style learning mechanism to the objective functions.

- We compare the qualitative and quantitative results of CIFR with the benchmark presented in [27]. This benchmark contains the previous filter removal approaches [4, 27] and the fundamental [22, 55] and the related [11, 32, 38, 44] image-to-image translation studies.

- We present the additional results of our proposed architecture within the following settings: (1) using pre-trained weights of IFRNet [27]. (2) including only PatchNCE loss [38] to the objective functions in [27]. (3) excluding Identity Regularization [38] or (4) well-known consistency losses used in [27].

- We demonstrate the impact of removing the visual effects brought by Instagram filters on the performance of the downstream vision models like localization and segmentation.

2. Related works

**Instagram Filter Removal.** Removing Instagram filters is an emerging task in vision, and investigated by only limited number of studies in the literature. [4] is one of the prominent studies trying to remove the visual effects brought by Instagram filters, and it follows a strategy for adaptively learning the parametric local transformations for each filter by using CNNs. By using a similar idea, [42] proposes a CNN architecture for transferring the photographic effects of a filter among the images with different contents by predicting the coefficients of the transformations applied. [27] introduces an adversarial methodology that directly learns to remove Instagram filters by adaptively normalizing the style information injected by filters in the feature representation of the filtered images. Moreover, there are other studies that try to recognize the filters applied to the images, instead of directly removing their effects, by using the ancestor CNNs (e.g., AlexNet [29], LeNet [31]) [5], or more commonly-used CNNs (e.g. VGGs [45], ResNets [17]) [9] or Siamese CNNs [8].

**Reverse Style Transfer.** Recent studies in Style Transfer [12, 13, 21] demonstrate that the style information of a reference image can be transferred into a target image without losing the main context. This can be considered as many-to-many translation [22, 25, 38, 55] where any style information is captured from the feature representation of a reference image, and then fused into the feature representation of another image. Similarly, Reverse Style Trans-
Contrastive Learning. Contrastive learning is one of the most popular strategies in representation learning. Recent studies [7, 15, 18, 48, 49] show that a methodology of maximizing mutual information is capable of learning more effective representations without requiring any supervision or hand-crafted objective functions. Noise contrastive estimation (NCE) [14] has become the popular choice for this purpose, and [36] demonstrates that NCE can learn the feature representations in a better and efficient manner. NCE basically builds on the notion of semantic similarity among the associated signals where more similar ones are represented in more similar ways. These associated signals can be an image with itself [10, 15, 35, 43], an image with its feature representation [18], an image with its patches [23, 49], or multiple views [48] or its different transformed versions [7]. Moreover, [38] employs this mechanism for conditional image synthesis task in multi-layer and patch-wise manner. In this study, we adapt the patch-wise contrastive learning methodology proposed in [38] to reverse style transfer task, and introduce the idea of using isolated patch sampling modules for the content and style information for distilling the semantic and style similarities among the signals.

3. Methodology

3.1. Patch-wise Contrastive Style Learning

Following the same assumption for the definition of reverse style transferring idea in [27], Instagram filter removal task can be explained in such a way that a given image \( \hat{X} \in \mathbb{R}^{H \times W \times 3} \) including a style information of an arbitrary filter injected by some transformation functions \( T(\cdot) \) is turned into its original version \( X \in \mathbb{R}^{H \times W \times 3} \) (i.e. without any additional style information injected) by a style removal module \( F(\cdot) \). The main purpose in this task is to discover the best style removal module for given images with different non-linear transformations applied.

\[
X = F(\hat{X}) \tag{1}
\]

where \( \hat{X} = T(X) \) and \( T(\cdot) \) is a general transformation function representing one or more transformations applied to \( X \). Since finding \( T^{-1}(\cdot) \) for each single image is an ill-posed problem, we need to discover the best possible \( F(\cdot) \), which can be substituted with \( T^{-1}(\cdot) \) with the minimum amount of reconstruction error.

Contrastive learning can be described as building the representations of the instances on the notion of semantic similarity among their associated signals. These signals can be represented as a query with its corresponding example instance (i.e. positive) and some non-corresponding example instances (i.e. negatives). The query \( v \in \mathbb{R}^{K} \), the positive \( v^+ \in \mathbb{R}^{K} \) and \( N \) negatives \( v^- \in \mathbb{R}^{NK} \) are mapped into \( K \)-dimensional vectors by an encoding structure. Note that these vectors are required to be unit vectors to avoid collapsing and exploding in their space, and thus should be normalized. The problem is formulated as an \((N+1)\)-way classification problem to maximize mutual information, where the query and the positive instances are closer to each other, while the query is located far away from the negatives in the vector space. The objective function for this problem stands for learning to select the positive instance over the negatives for a particular query instance, and it is defined in Equation 2.

\[
\ell(v, v^+, v^-) = -log \left[ \frac{\exp(v \cdot v^+ / \tau)}{\exp(v \cdot v^+ / \tau) + \sum_{n=1}^{N} \exp(v \cdot v^-_n / \tau)} \right] \tag{2}
\]

where \( N \) is the number of negative instances and \( \tau \) is the temperature parameter for scaling.

In this study, we employ the multi-layer and patch-based contrastive learning objective [38] to eliminate the visual effects brought by Instagram filters. A particular patch of an image filtered by any arbitrary Instagram filter should associate with the patch of its original version at the exact location. The other patches typically do not associate with this patch. The patch at each spatial location can be represented by the feature maps computed by an encoder in a different scale in each layer. Note that the feature maps in deeper layers correspond to larger patches. We can demonstrate this patch sampling module as follows:

\[
\{z^l\}_L = \{H^l(E^l(\hat{X}))\}_L \nonumber \\
\{\hat{z}^l\}_L = \{H^l(E^l(\hat{X}))\}_L \tag{3}
\]

where \( z^l \) is the feature map of filtered image in \( l \)-th layer, \( \hat{z}^l \) is the feature map of unfiltered image, \( H^l \) is the mapper network for patch sampling (i.e. a two-layer MLP network as in [7]), \( E^l \) is the encoder network, \( L \) is the number of layers in \( E \). \( \hat{X} \) and \( \hat{X} \) represent the filtered and unfiltered images, respectively.

We extend this idea in [38] by using isolated patch sampling modules for the content and style information, where the single level pipeline can be seen in Figure 1. The main motivation behind this practice is to distill the learning process of the semantic and style similarities among the associated patches. At this point, we can leverage the original version of the images to capture the pure style (i.e. without
any additional style injected). To achieve this, we extract the Gram matrices of the feature maps to express the style information via the feature correlations, and employ them to our contrastive learning pipeline. We can formulate the isolated patch sampling modules as follows:

\[
\begin{align*}
\{\tilde{z}_c^l, \tilde{z}_s^l\}_L &= \{H_c^l(E_c^l(\tilde{X})), G_c^l(H_c^l(E_c^l(\tilde{X})))\}_L \\
\{\hat{z}_c^l, \hat{z}_s^l\}_L &= \{H_c^l(E_c^l(\check{X})), G_c^l(H_c^l(E_c^l(\check{X})))\}_L \\
\{z_c^l, z_s^l\}_L &= \{H_c^l(E_c^l(X)), G_c^l(H_c^l(E_c^l(X)))\}_L
\end{align*}
\]

where \(G^l \in \mathbb{R}^{K \times K}\) is the Gram matrix, the inner product between the features mapped by \(H_c^l\) in \(l\)-th layer, \(\tilde{z}_c, \check{z}_c, z_c\) stand for the content feature maps and \(\tilde{z}_s, \check{z}_s, z_s\) for the style feature maps of the filtered, unfiltered and original images, respectively. \(H_c\) represents the content patch sampling module, while \(H_s\) is the style patch sampling module.

For the content matching, we try to match the corresponding filtered and unfiltered patches at the same location, while exploiting the other patches within the filtered image as negatives. We also try to emulate the pure style in the patches of the original image for the unfiltered patches. Note that, within the computational constraint, the most affordable way of capturing the pure style in such a contrastive setup is to build the strategy as 2-way classification where the negative instance has the exact same semantic information, but with different style injected. At this point, we combine two contrastive learning objectives, namely content NCE \(L_C\), and style NCE \(L_S\), for extracting the content and style information separately. Our extended version of PatchNCE loss is shown in Equation 5.

\[
\begin{align*}
L_C(E, H, X, \check{X}) &= \mathbb{E}_{x \sim X, \tilde{x} \sim \tilde{X}} \sum_{l=1}^{L} \sum_{t=1}^{T_l} \ell(\hat{z}_c^{lt}, \tilde{z}_c^{lt}, \tilde{z}_c^{lT_l}) \\
L_S(E, H, X, \check{X}) &= \mathbb{E}_{x \sim X, \tilde{x} \sim \tilde{X}} \sum_{l=1}^{L} \sum_{t=1}^{T_l} \ell(\hat{z}_s^{lt}, z_s^{lt}, \tilde{z}_s^{lt}) \\
\mathcal{L}_{\text{PatchNCE}} = \gamma_c L_C + \gamma_s L_S
\end{align*}
\]

where \(T_l\) is the list of different spatial locations for patches at \(l\)-th layer, \(l'\) represents a single arbitrary spatial location different than \(l\), \(\gamma_c\) and \(\gamma_s\) are the coefficients of the content and style NCE losses. In our experiments, both coefficients are set to 0.5.

### 3.2. Architecture

In our architecture design, we mostly follow the design of IFRNet proposed in [27]. IFRNet has an encoder-decoder structure with a style extractor module for applying adaptive feature normalization to all layers in the encoder. Style extractor module learns to adapt the affine parameters for the feature representations encoded by a pre-trained
VGG network by using different fully-connected heads for each layer. The affine parameters are sent into adaptive instance normalization (AdaIN) [21] layers in each encoder level to eliminate the external style information. Note that any related information about the original style is retained by including skip connections to the normalized feature maps [27]. At the end, we stack the features in all levels in the encoder, and feed this multi-layer feature list to our proposed isolated patch sampling modules for the content and style information. We introduce the technical details about isolated patch sampling modules in Section 3.1.

As distinguished from IFRNet, we do not have an auxiliary classifier for classifying the filter type in our design. In [27], the auxiliary classifier has been used for satisfying a naive way of maximizing mutual information between corresponding input and output instances. However, we do not prefer to include this kind of a classifier since we can achieve this practice in a more elegant way via Noise Contrastive Estimation [14].

We feed the latent representations with no external style information to the decoder of IFRNet. The decoder contains six consecutive upsampling and residual convolutional blocks, and learns to generate the recovered image with adversarial training. Discriminators are designed as in [22,27] to penalize the global image and local patches at different scales. Our proposed architecture, namely Contrastive Instagram Filter Removal Network (CIFR) is shown in Figure 2.

3.3. Objective Function

In our study, we combine three different objective functions for our pipeline: (1) multi-layer patch-wise contrastive style loss, (2) consistency loss, (3) adversarial loss. The first one is introduced in Equation 5, and it is responsible for distilling the learning process of the patch-level semantic and style similarities. Next, we include the consistency loss used in [27] to our final objective function in order to ensure the semantic and texture consistency of the output. We also employ a common adversarial training strategy \( i.e. WGAN-GP \) [1] in order to enhance the realism of the recovered images, and it is demonstrated in Equation 6. Our final objective function for the generator \( \mathcal{L}_G \) can be seen in Equation 7.

\[
\mathcal{L}^G_{WGAN-GP} = -\mathbb{E}_{x \sim p(x)}[D(x)] + \mathbb{E}_{\tilde{x} \sim p(\tilde{x})}[D(\tilde{x})] + \lambda_{gp} \mathcal{L}_{GP} \\
\mathcal{L}^D_{WGAN-GP} = -\mathbb{E}_{\tilde{x} \sim p(\tilde{x})}[D(\tilde{x})]
\]

where \( D \) stands for the discriminator network, \( \mathcal{L}_{GP} \) is the gradient penalty term.

\[
\mathcal{L}_G = \lambda_p \mathcal{L}_{PatchNCE} + \lambda_c \mathcal{L}_{Cons} + \lambda_a \mathcal{L}^G_{WGAN-GP}
\]

where \( \lambda_p, \lambda_c \) and \( \lambda_a \) are the coefficients for the patch-wise NCE loss, consistency loss and adversarial loss, and set to \( 5 \times 10^{-1}, 10^{-2} \) and \( 10^{-3} \), respectively.

4. Results

In this study, we investigate the performance of multi-layer patch-wise contrastive learning approach on Instagram filter removal task, which can be described as a reverse style transfer problem. We compare the performance of our proposed architecture, namely CIFR, against the previous filter removal approaches [4,27], the fundamental [22,55] and the related [11,32,38,44] image-to-image translation studies, and its own variants with different training settings. Note that we have obtained the available results in [27], and re-trained the rest of compared methods on IFFI dataset with their default hyper-parameters settings. Lastly, we show the qualitative and quantitative impact of removing Instagram filters from the images on the performance of downstream vision tasks \( i.e. \) localization, segmentation.

4.1. Experimental Setup

We tested our methodology on IFFI dataset, which is introduced by [27], and contains 9,600 high-resolution and aesthetically pleasing images along with their filtered versions by 16 different Instagram filters. There are 8,000 training and 1,600 test images in this dataset. In our experiments, we resized the images to the resolution of 256, and only applied random horizontal flipping before feeding them to our proposed model. We used the pre-trained weights of IFRNet [27] in our default settings. We have picked Adam optimizer [26] with \( \beta_1 = 0.5 \) and \( \beta_2 = 0.999 \) for all modules, and the learning rates for the generator, discriminator and patch sampling modules are set to \( 2 \times 10^{-4}, 10^{-4} \) and \( 10^{-5} \), respectively. We did not use any scheduling for the learning rates during training. The temperature parameter \( \tau \) is set to 0.07. We have conducted our experiments on 2x NVIDIA RTX 2080Ti GPU with batch size of 8. We have trained proposed architecture for 40K steps for pre-trained settings, and 120K for training from scratch. We have implemented the code in PyTorch [39]. The source code can be found at https://github.com/birdortyedi/cifr-pytorch.

4.2. Qualitative Comparison

We present the qualitative results of our proposed architecture and the other compared methods on Instagram filter removal in Figure 3. When compared to the previous studies, the results show that CIFR improves the quality of recovered images by composing adaptive feature normalization idea and multi-layer patch-wise contrastive style learning for filter removal. CIFR minimizes the inconsistency on the background and foreground color tones, and it leads to have less artifacts \( i.e. \) checkerboard, false color, filter residuals on different parts of the outputs. At this point, we present the comparison of the residual images of the most successful four methodologies in our benchmark in Figure 4. This figure verifies that the residuals are typically formed
by the visual effects brought by the corresponding filter, and CIFR performs best on effectively removing these effects. The residuals are calculated by the scaled absolute error between the output and the original version.

4.3. Quantitative Analysis

We have followed the same procedure in [27] for evaluating the quantitative performance of CIFR where four common image similarity metrics are employed in the experiments. These metrics are SSIM, PSNR, Learned Perceptual Image Patch Similarity (LPIPS) [54] and CIE 2000 Color Difference ($\text{CIE-}\Delta E$) [34]. We show the results in Table 1 for our proposed architecture and the other compared methods obtained by training on IFFI dataset [27]. Our method has generally better quantitative performance than the other methods in the benchmark. Particularly, CIFR surpasses the previous studies in SSIM and $\text{CIE-}\Delta E$ metrics. For PSNR measurements, although CIFR outperforms the fundamental [22, 55] and the related [11, 32, 38, 44] image-to-image translation studies and the prominent method [4], it falls behind IFRNet [27] by 0.02%.

Ablations. In this study, we build our proposed architecture in different settings to observe its performance in detail.
consistency losses used in [27]. Note that the exact same architecture, as shown in Figure 2, and hyper-parameters are used for all these settings. Table 1 also presents the results of additional experiments for proposed architecture. Our first observation is that we can achieve on-par performance by training our architecture from scratch, and it still performs better quantitative performance than the previous studies. Secondly, distilling the learning process of semantic and style similarities for the patch-wise contrastive learning strategy significantly improves the results of Instagram filter removal on all experiments. This demonstrates that capturing the pure style of the original images is one of the key aspects of removing the filters. Next, Identity Regularization has limited impact on the overall performance, and it can be omitted for reducing the training workload. Finally, excluding the consistency losses used in [27] from the final objective function leads to a decrease on the performance. However, these losses are quiet expensive functions for the training workload. Therefore, we believe that there is a still room for leaving out these expensive loss terms while improving the performance.

Impact on Vision Tasks. As pointed out in [27], due to different distractive factors, CNNs may not perform well for the real-world applications as much as in the standard benchmark studies. Noise or blurring in real-world scenarios or different transformations applied to the images can be the examples of these distractive factors. Likewise, Instagram filters transform the images into a different version whose feature maps change substantially. These changes arguably lead to the performance degradation on visual understanding tasks. At this point, we demonstrate the impact of removing Instagram filters from the images before feeding them to the downstream vision models. To achieve this, we make inferences of filtered and recovered images for localization and segmentation tasks. Filtered test images of IFFI dataset and their unfiltered versions by our proposed architecture are predicted by Attr-Mask-RCNN model, which is proposed in [24], and trained on Fashionpedia dataset [24]. Note that we made annotated the localization and segmentation ground truth of IFFI dataset by human annotators. Figure 5 shows that Instagram filters may degrade the performance of the downstream vision tasks (e.g. misclassification, missing detections, wrong detection location, problematic segmentation maps, etc.). This can be mitigated by wiping off the visual effects brought by these filters, and recover the images back to their original versions. In addition to this, we extend the evaluation of using filter removal strategy as a pre-processing step for the downstream

<table>
<thead>
<tr>
<th>Method</th>
<th>SSIM</th>
<th>PSNR</th>
<th>LPIPS</th>
<th>CIE-ΔE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE [4]</td>
<td>0.748</td>
<td>23.41</td>
<td>0.069</td>
<td>39.55</td>
</tr>
<tr>
<td>pix2pix [22]</td>
<td>0.825</td>
<td>26.35</td>
<td>0.048</td>
<td>30.32</td>
</tr>
<tr>
<td>CycleGAN [55]</td>
<td>0.819</td>
<td>22.94</td>
<td>0.065</td>
<td>36.59</td>
</tr>
<tr>
<td>AngularGAN [44]</td>
<td>0.846</td>
<td>26.30</td>
<td>0.048</td>
<td>31.11</td>
</tr>
<tr>
<td>IFRNet [27]</td>
<td>0.864</td>
<td>30.46</td>
<td>0.025</td>
<td>20.72</td>
</tr>
<tr>
<td>DRIT++ [32]</td>
<td>0.626</td>
<td>16.23</td>
<td>0.162</td>
<td>47.95</td>
</tr>
<tr>
<td>GeGAN [11]</td>
<td>0.838</td>
<td>21.75</td>
<td>0.060</td>
<td>38.54</td>
</tr>
<tr>
<td>FastCUT [38]</td>
<td>0.763</td>
<td>20.08</td>
<td>0.083</td>
<td>39.86</td>
</tr>
<tr>
<td>CUT [58]</td>
<td>0.744</td>
<td>20.96</td>
<td>0.081</td>
<td>38.64</td>
</tr>
<tr>
<td>CIFR-no-pre-training</td>
<td>0.888</td>
<td>29.24</td>
<td>0.024</td>
<td>20.65</td>
</tr>
<tr>
<td>CIFR-no-style-nce</td>
<td>0.859</td>
<td>28.13</td>
<td>0.034</td>
<td>23.01</td>
</tr>
<tr>
<td>CIFR-no-id-reg</td>
<td>0.879</td>
<td>29.40</td>
<td>0.025</td>
<td>19.82</td>
</tr>
<tr>
<td>CIFR-no-consistency</td>
<td>0.874</td>
<td>29.42</td>
<td>0.027</td>
<td>21.23</td>
</tr>
<tr>
<td>CIFR</td>
<td>0.880</td>
<td>30.02</td>
<td>0.023</td>
<td>19.05</td>
</tr>
</tbody>
</table>

Table 1. Quantitative comparison of proposed architecture, its own variants and the compared methods on IFFI dataset. Obtained the available results from [27], and re-trained the rest from scratch.
In this experiment, we measure the performance of Attr-Mask-RCNN on IFFI dataset by using per-category mean average precision (mAP) of bounding boxes and segmentation masks. Table 2 presents the quantitative results of both filtered and recovered images on localization and segmentation tasks. Note that we pick 6 most frequently appeared clothing categories of IFFI dataset for evaluating per-category mAP. The results support the main motivation behind the idea of directly removing any externally applied filters from the images to improve the overall performance of the downstream vision tasks. Moreover, it also verifies that different levels of perturbations have a negative impact on understanding the image contents, and thus hindering the further analysis of them in particular applications.

## 5. Conclusions

In this study, we propose a novel strategy for removing Instagram filters, which is a patch-wise contrastive learning mechanism for distilling the learning process of the semantic and style similarities. In addition to matching filtered and unfiltered patches at the same location, we also try to imitate the pure style of an original image by enabling a single negative instance for contrastive style learning, which has the same semantic information with the query instance, but with different style. Experiments show that our proposed architecture for this strategy mostly outperforms the performance of the previous studies on IFFI dataset and better to prevent to reduce the overall performance of the downstream vision models.

### Discussion of Limitations

We believe that there is still some room for improvements on the limitations of our strategy and the task itself. First, the volume and the diversity of the dataset used for this task (i.e. the number of instances, filters and annotations) are limited when compared to the datasets of the other vision tasks, and it can be increased to be able to conduct more extensive experiments on this task. Next, the consistency loss shown as the part of Equation 7 be able to conduct more extensive experiments on this task. The hyper-parameters mainly defining the filtering workload. The hyper-parameters mainly defining the filtering workload. Therefore, a more elegant objective function can be designed for replacing with the downstream vision models.

<table>
<thead>
<tr>
<th>Filters</th>
<th>Top Shirt Pants Dress Shoe Glasses</th>
<th>Segmentation (mAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Shirt Pants Dress Shoe Glasses</td>
<td>15.189 17.639</td>
<td>15.200 11.634</td>
</tr>
<tr>
<td>Top Shirt Pants Dress Shoe Glasses</td>
<td>15.175 7.525 21.051</td>
<td>7.525 10.179</td>
</tr>
<tr>
<td>Top Shirt Pants Dress Shoe Glasses</td>
<td>R-CIFR</td>
<td>15.077 9.571</td>
</tr>
<tr>
<td>Top Shirt Pants Dress Shoe Glasses</td>
<td>15.175 7.525 21.051</td>
<td>7.525 10.179</td>
</tr>
<tr>
<td>Top Shirt Pants Dress Shoe Glasses</td>
<td>R-CIFR</td>
<td>15.077 9.571</td>
</tr>
<tr>
<td>Top Shirt Pants Dress Shoe Glasses</td>
<td>15.175 7.525 21.051</td>
<td>7.525 10.179</td>
</tr>
</tbody>
</table>

Table 2. Quantitative comparison for employing filter removal strategy to the data before feeding it to the downstream vision tasks. We present per-category mean average precision (mAP) scores of both filtered and recovered images on localization and segmentation tasks. We use our proposed architecture, namely CIFR and the prior work [27], for Instagram filter removal, and Attr-Mask-RCNN architecture proposed in [24] for clothing localization and segmentation, which is trained on Fashionpedia dataset [24] and tested on IFFI dataset [27].
References


[23] Phillip Isola, Daniel Zoran, Dilip Krishnan, and E. Adelson. Learning visual groups from co-occurrences in space and time. ArXiv, abs/1511.06811, 2015. 3


