Patch-wise Contrastive Style Learning for Instagram Filter Removal

Furkan Kınlı[1] Barış Özcan[2] Furkan Kıraç[3]
Video, Vision and Graphics Lab
Özyeğin University
{furkan.kinli[1], furkan.kirac[3]}@ozyegin.edu.tr, baris.ozcan.10097@ozu.edu.tr[2]

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.

1. Introduction

Social media filters (*e.g.* Instagram filters) transform an image into a different version by applying several transformations, and this modified version may have color-level or pixel-level corruptions and perturbations. These filters modify the original image by adjusting the contrast and brightness, or changing hue and saturation, or introducing different levels of blur and noise, or applying color curves or vignetting. Though these filters convert images to a more aesthetically pleasing appearance, they also make the content in those images more complicated to understand by learning-based algorithms. Therefore, removing the filters from social media images is a crucial preprocessing step completed for the visual analysis of social media contents.

Convolutional Neural Networks (CNNs) are one of the most common choices for solving different vision tasks, and there are several prominent studies that propose the fundamental solutions based on CNNs for these tasks such as classification [17,20,45,46], localization [33,40,41,50,56], segmentation [6, 16, 50, 56], tracking [2, 3, 51] and retrieval [19, 28]. However, the recent studies [30, 37, 53] argue that CNNs are not robust to the image-level corruptions and perturbations for the downstream tasks, and this leads to a significant decrease on the performance regardless of the task. At this point, the filters applied to the social media images can be considered as the natural example of the image-level corruptions and perturbations, which can be frequently encountered in several real-world vision applications. As exemplified in [27], CNNs do not give the exact segmentation outputs for the original image and its filtered versions, but also give intolerably inaccurate outputs for the filtered versions, due to the different levels of corruptions and different types of perturbations caused by filtering.

In the previous studies, there are two main approaches proposing a solution for better analyzing the filtered social media images: (1) the filter classification [5, 8, 9, 52], (2) learning the transformations applied by the filter [4,42]. The main drawback for both approaches is that they do not directly try to recover the original images, but only to learn the class, or to approximate the transformation matrix of the filters applied. Recently, a novel approach for recovering the original images from the filtered versions has been proposed by [27]. This approach mainly assumes that the filters applied inject the additional style information to the images, and thus considering the filter removal problem as a reverse style transfer problem. We combine adaptive feature normalization idea for filter removal as in [27] and the patch-wise contrastive learning mechanism [38], and improve them. In this study, we propose *Contrastive Instagram Filter Removal Network (CIFR)*, which employs novel patch sampling modules for contrastive semantic and style NCE losses leading to preserve the semantic information while removing the additional style information injected by the filters. This work has the following contributions:

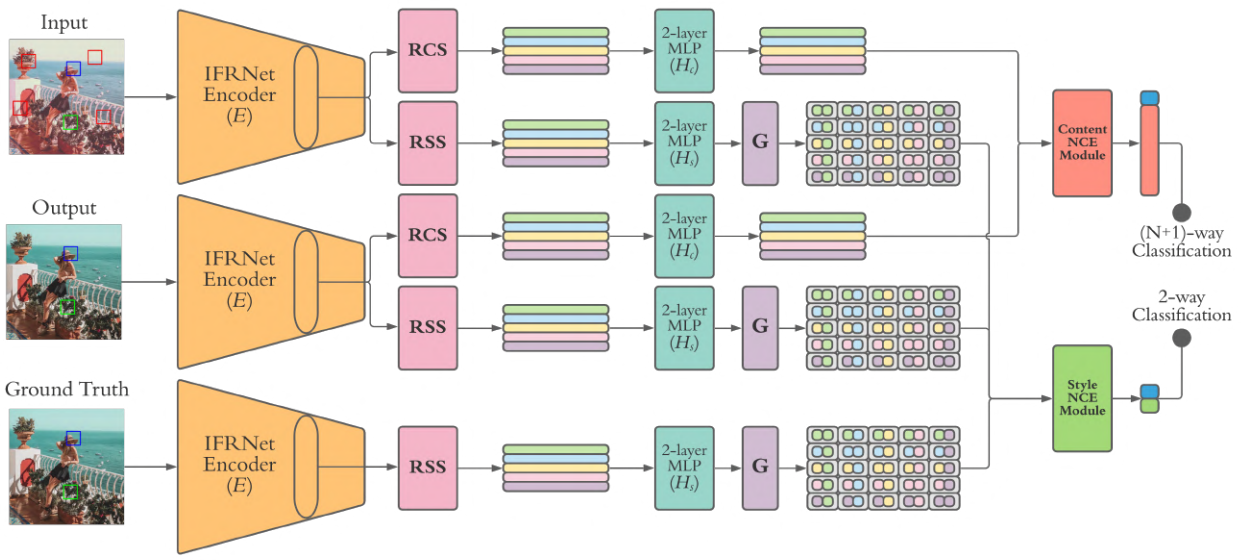

Figure 1. Isolated patch sampling modules for distilling the content and style information. This figure shows the pipeline for only a **single** level features. The extracted feature maps by IFRNet Encoder (E) are first fed into the random sampling modules for the content (**RCS**) and style (**RSS**), 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-

fer is described in [27] as *many-to-one* translation where the multiple styles injected into an image can be eliminated by adaptively normalizing its feature representations in different levels, so that we can reverse it into its pure style (*i.e*. without any additional style injected). In this study, we mainly follow the same idea for removing Instagram filters from the images. Assuming that the filters applied to the images interpolate a particular style information to the feature maps of these images, they can be swept away from the images during the extraction of feature representations.

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 $\tilde{\mathbf{X}} \in \mathbb{R}^{H \times W \times 3}$ including a style information of an arbitrary filter injected by some transformation functions $\mathbf{T}(\cdot)$ is turned into its original version $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$ (*i.e*. without any additional style information injected) by a style removal module $\mathbf{F}(\cdot)$. The main purpose in this task is to discover the best style removal module for given images with different non-linear transformations applied.

$$\mathbf{X} = \mathbf{F}(\tilde{\mathbf{X}}) \tag{1}$$

where $\tilde{\mathbf{X}} = \mathbf{T}(\mathbf{X})$ and $\mathbf{T}(\cdot)$ is a general transformation function representing one or more transformations applied to $\mathbf{X}$. Since finding $\mathbf{T}^{-1}(\cdot)$ for each single image is an ill-posed problem, we need to discover the best possible $\mathbf{F}(\cdot)$, which can be substituted with $\mathbf{T}^{-1}(\cdot)$ 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 $\mathbf{v} \in \mathbb{R}^K$, the positive $\mathbf{v}^+ \in \mathbb{R}^K$ and N negatives $\mathbf{v}^- \in \mathbb{R}^{N \times K}$ 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 to 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.

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

where N is the number of negative instances and τ 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:

$$\begin{aligned} \{\mathbf{z}^l\}_L &= \{H^l(E^l(\tilde{\mathbf{X}}))\}_L \\ \{\hat{\mathbf{z}}^l\}_L &= \{H^l(E^l(\hat{\mathbf{X}}))\}_L \end{aligned} \tag{3}$$

where $\mathbf{z}^l$ is the feature map of filtered image in l-th layer, $\hat{\mathbf{z}}^l$ 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, $\tilde{\mathbf{X}}$ and $\hat{\mathbf{X}}$ 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

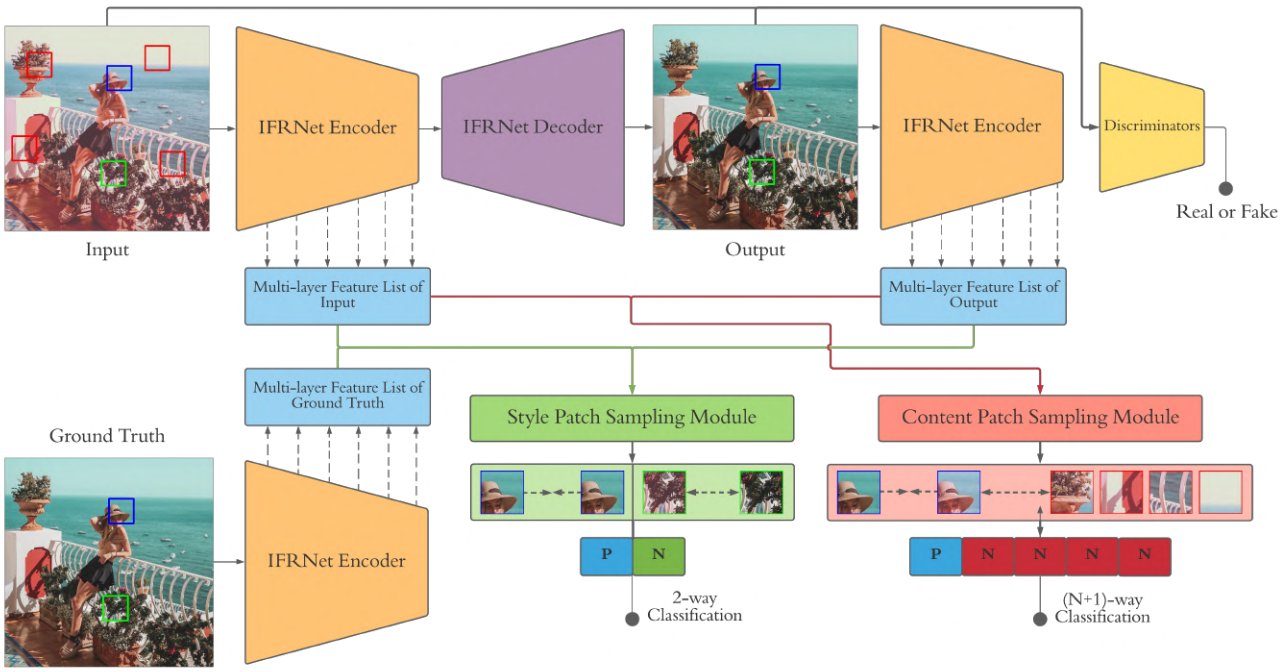

Figure 2. Overall architecture of Contrastive Instagram Removal Network (CIFR). The stacked features extracted by IFRNet encoder [27] are fed into our proposed isolated patch sampling modules. For each patch level, content NCE loss and style NCE loss are disjointly calculated for distilling the content and style information.

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:

$$\begin{aligned}\{\tilde{\mathbf{z}}_c^l, \tilde{\mathbf{z}}_s^l\}_L &= \{H_c^l(E^l(\tilde{\mathbf{X}})), \mathbf{G}^l(H_s^l(E^l(\tilde{\mathbf{X}})))\}_L \\ \{\hat{\mathbf{z}}_c^l, \hat{\mathbf{z}}_s^l\}_L &= \{H_c^l(E^l(\hat{\mathbf{X}})), \mathbf{G}^l(H_s^l(E^l(\hat{\mathbf{X}})))\}_L \\ \{\mathbf{z}_c^l, \mathbf{z}_s^l\}_L &= \{H_c^l(E^l(\mathbf{X})), \mathbf{G}^l(H_s^l(E^l(\mathbf{X})))\}_L \end{aligned} \quad (4)$$

where $\mathbf{G}^l \in \mathbb{R}^{K\times K}$ is the Gram matrix, the inner product between the features mapped by H_s^l in l-th layer, $\tilde{\mathbf{z}}_c, \hat{\mathbf{z}}_c, \mathbf{z}_c$ stand for the content feature maps and $\tilde{\mathbf{z}}_s, \hat{\mathbf{z}}_s, \mathbf{z}_s$ 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 $\mathcal{L}_C$, and style NCE $\mathcal{L}_S$, for extracting the content and style information separately. Our extended version of PatchNCE loss is shown in Equation 5.

$$\begin{aligned}\mathcal{L}_C(E,H,\mathbf{X},\tilde{\mathbf{X}}) &= \mathbb{E}_{x\sim\mathbf{X},\tilde{x}\sim\tilde{\mathbf{X}}} \sum_{l=1}^{L}\sum_{t=1}^{T^l} \ell(\hat{\mathbf{z}}_c^{l,t}, \tilde{\mathbf{z}}_c^{l,t}, \tilde{\mathbf{z}}_c^{l,T\setminus t}) \\ \mathcal{L}_S(E,H,\mathbf{X},\tilde{\mathbf{X}}) &= \mathbb{E}_{x\sim\mathbf{X},\tilde{x}\sim\tilde{\mathbf{X}}} \sum_{l=1}^{L}\sum_{t=1}^{T^l} \ell(\hat{\mathbf{z}}_s^{l,t}, \mathbf{z}_s^{l,t}, \tilde{\mathbf{z}}_s^{l,t'}) \\ \mathcal{L}_{PatchNCE} &= \gamma_c\mathcal{L}_C + \gamma_s\mathcal{L}_S \end{aligned} \quad (5)$$

where T^l is the list of different spatial locations for patches at l-th layer, t' represents a single arbitrary spatial location different than t, γ_c and γ_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.

$$\mathcal{L}^{D}_{WGAN-GP} = -\mathbb{E}_{x\sim\mathbf{X}}[D(x)] + \mathbb{E}_{\hat{x}\sim\hat{\mathbf{X}}}[D(\hat{x})] + \lambda_{gp}\mathcal{L}_{GP}$$
$$\mathcal{L}^{G}_{WGAN-GP} = -\mathbb{E}_{\hat{x}\sim\hat{\mathbf{X}}}[D(\hat{x})] \quad (6)$$

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

$$\mathcal{L}_G = \lambda_p\mathcal{L}_{PatchNCE} + \lambda_c\mathcal{L}_{Cons} + \lambda_a\mathcal{L}^{G}_{WGAN-GP} \quad (7)$$

where λ_p, λ_c and λ_a are the coefficients for the patch-wise NCE loss, consistency loss and adversarial loss, and set to 5×10^{-1}, 10^{-3} 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 2×10^{-4}, 10^{-4} and 10^{-5}, respectively. We did not use any scheduling for the learning rates during training. The temperature parameter τ 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

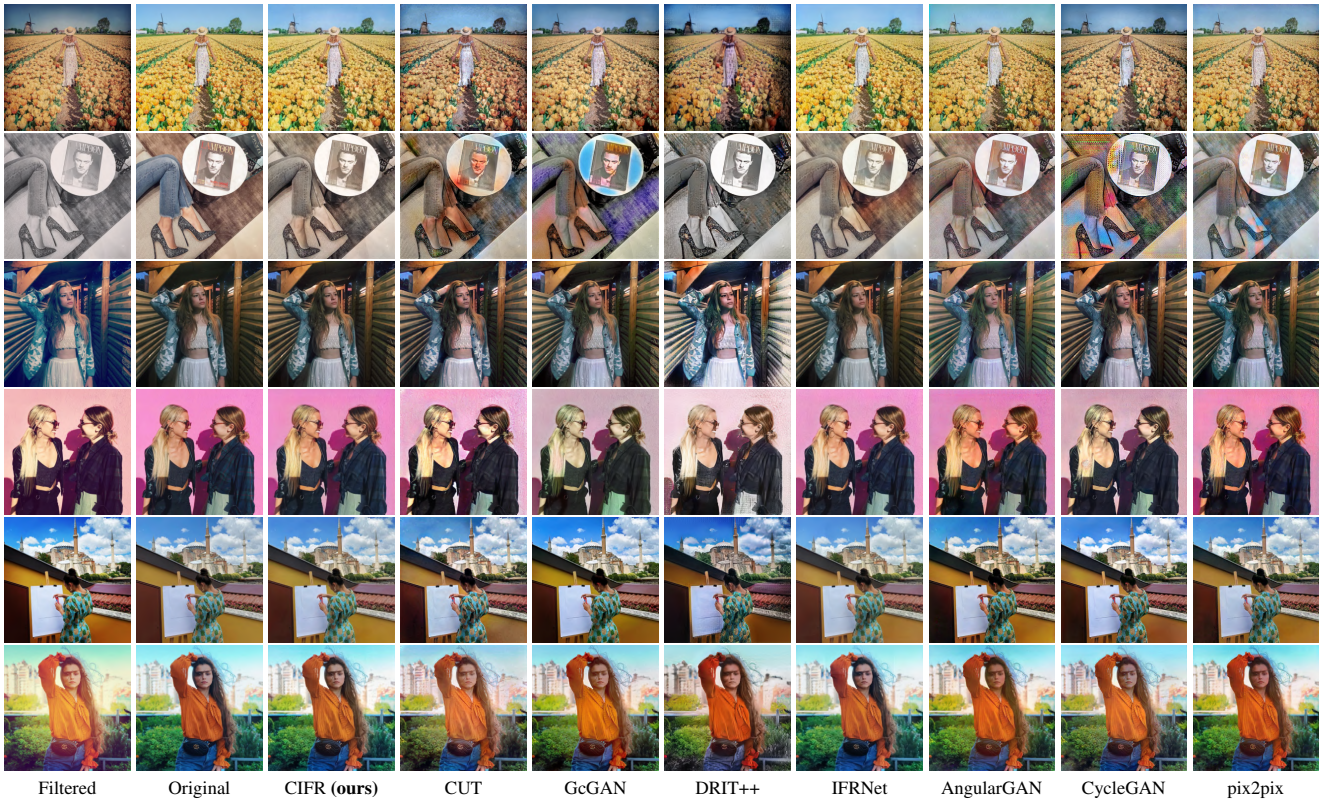

Figure 3. Comparison of the qualitative results of Instagram filter removal on IFFI dataset. Filters applied (top to bottom): *Sutro*, *Willow*, *Nashville*, *Amaro*, *Lo-Fi*, *Toaster*.

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 (CIE-ΔE) [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 CIE-ΔE 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.

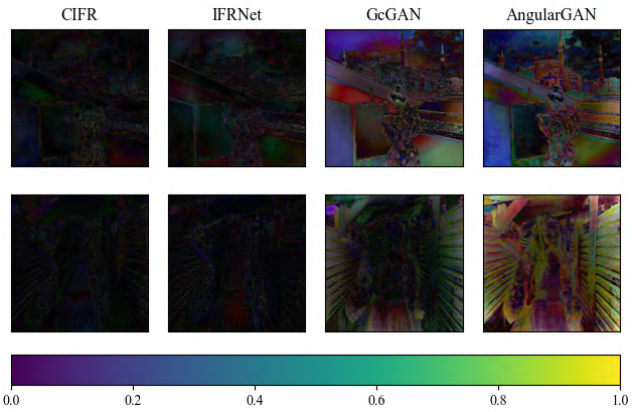

Figure 4. The residuals extracted by the absolute difference between the recovered images and their original version.

The changes in our settings can be listed as follows: (1) we start our training from scratch, not using pre-trained weights of IFRNet. (2) we only include content PatchNCE loss to the objective functions as in [38]. (3) we exclude Identity Regularization [38] from the extended version of our PatchNCE loss. (4) we leave out the semantic and texture

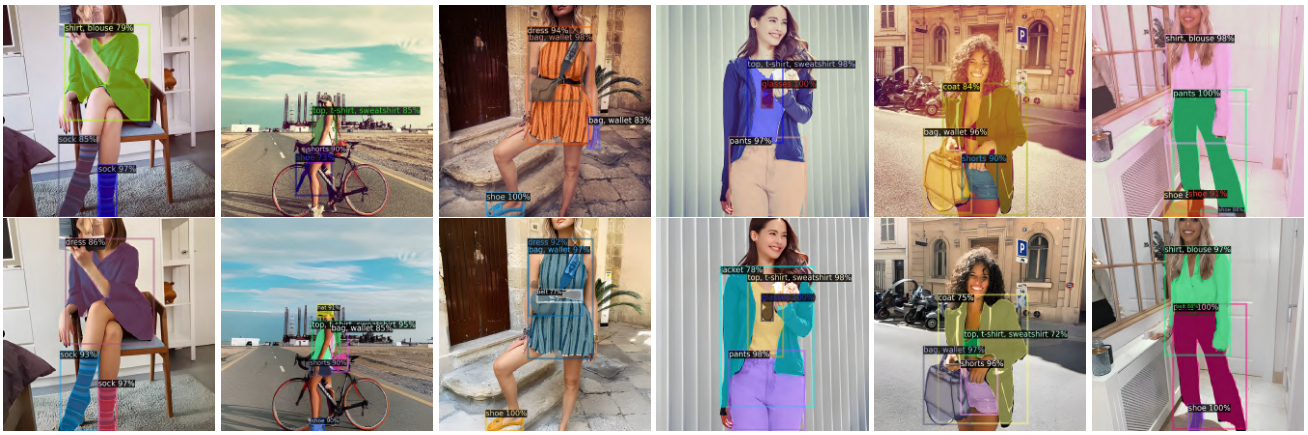

Figure 5. Demonstrating the impact of Instagram filter removal on downstream vision tasks like detection and segmentation. Examples are obtained from IFFI dataset [27], and predicted by Attr-Mask-RCNN trained on Fashionpedia dataset, which is introduced in [24]. Rows: (1) the results of the images filtered by different Instagram filters, (2) the results of the images unfiltered by CIFR. Zoom in for better view.

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.

Method	SSIM ↑	PSNR ↑	LPIPS ↓	CIE-ΔE ↓
PE [4]	0.748	23.41	0.069	39.55
pix2pix [22]	0.825	26.35	0.048	30.32
CycleGAN [55]	0.819	22.94	0.065	36.59
AngularGAN [44]	0.846	26.30	0.048	31.11
IFRNet [27]	0.864	**30.46**	0.025	20.72
DRIT++ [32]	0.626	16.23	0.162	47.95
GcGAN [11]	0.838	21.75	0.060	38.54
FastCUT [38]	0.763	20.08	0.083	39.86
CUT [38]	0.744	20.96	0.081	38.64
CIFR-no-pre-training	**0.888**	29.24	0.02441	20.65
CIFR-no-style-nce	0.859	28.13	0.03426	23.01
CIFR-no-id-reg	0.879	29.40	0.02528	19.82
CIFR-no-consistency	0.874	29.42	0.02708	21.23
CIFR	0.880	30.02	**0.02321**	**19.05**

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.

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

Filters		Localization (mAP)						Segmentation (mAP)					
		Top	**Shirt**	**Pants**	**Dress**	**Shoe**	**Glasses**	**Top**	**Shirt**	**Pants**	**Dress**	**Shoe**	**Glasses**
1977	Filtered	7.976	0.000	11.348	5.406	16.084	9.505	9.773	0.000	10.228	6.713	13.424	7.657
	R-IFRNet	12.653	6.931	13.871	11.042	24.318	**13.175**	11.521	7.178	12.815	11.978	**19.716**	9.769
	R-CIFR	**13.115**	**10.891**	**15.175**	**11.314**	**24.332**	10.297	**14.088**	**10.561**	**13.307**	**12.866**	19.004	**9.901**
Amaro	Filtered	11.269	2.970	14.132	7.525	21.051	7.525	10.414	3.960	13.508	10.179	15.323	7.525
	R-IFRNet	**13.035**	6.188	13.890	10.144	26.027	10.594	**11.658**	7.426	14.001	11.560	20.232	9.208
	R-CIFR	12.673	**8.168**	**16.006**	**11.083**	**28.626**	**10.693**	11.057	**8.911**	**14.598**	**11.644**	**22.275**	**9.901**
Brannan	Filtered	10.790	2.475	13.228	6.943	19.572	10.990	11.607	3.960	11.484	7.017	15.271	7.168
	R-IFRNet	13.673	6.931	12.895	**10.562**	**26.027**	8.911	13.359	6.436	12.665	**11.453**	**21.615**	8.020
	R-CIFR	**14.999**	**9.901**	**13.516**	9.537	25.709	**11.221**	**15.200**	**11.634**	**13.264**	10.911	20.977	**8.581**
Hudson	Filtered	13.294	5.941	13.512	9.285	24.554	13.861	13.818	5.941	12.437	11.243	18.329	10.693
	R-IFRNet	**15.093**	6.188	13.964	10.664	27.041	13.812	15.558	7.426	**14.420**	**11.863**	21.283	**11.023**
	R-CIFR	14.322	**10.297**	**14.844**	**10.673**	**29.872**	**11.287**	**15.815**	**11.337**	14.308	11.241	**21.654**	10.297
Nashville	Filtered	12.322	6.931	12.110	10.326	21.806	**11.089**	11.432	6.436	11.305	10.927	16.387	8.079
	R-IFRNet	13.707	6.931	14.645	**10.686**	24.811	7.525	14.546	7.426	12.643	10.485	19.994	6.733
	R-CIFR	**15.077**	**9.571**	**15.705**	9.884	**28.064**	10.108	**14.712**	**9.901**	**13.452**	**11.193**	**22.437**	**8.515**
Perpetua	Filtered	14.494	5.941	14.238	7.475	21.133	**14.072**	14.628	5.941	12.202	8.376	17.263	**12.208**
	R-IFRNet	15.407	6.188	13.768	**11.634**	26.154	12.541	15.879	6.931	12.932	**11.997**	19.264	10.693
	R-CIFR	**16.903**	**8.168**	**15.880**	11.541	**28.047**	13.333	**16.939**	**9.406**	**13.186**	11.861	**22.065**	10.033
Valencia	Filtered	12.481	6.188	12.105	9.010	23.083	9.901	12.558	7.426	10.671	9.036	18.131	7.683
	R-IFRNet	**14.490**	6.436	**15.904**	10.771	27.347	10.337	14.315	7.178	14.138	**12.291**	20.624	9.743
	R-CIFR	14.467	**9.901**	14.809	**11.735**	**30.238**	**10.693**	**14.932**	**9.653**	**14.464**	12.263	**23.358**	**10.198**
X-Pro II	Filtered	13.604	6.188	12.555	8.465	21.637	**12.752**	12.389	6.188	11.111	8.540	16.369	9.795
	R-IFRNet	**15.252**	8.168	13.746	10.815	25.722	11.116	15.751	8.911	12.605	12.546	19.757	9.003
	R-CIFR	15.189	**9.818**	**14.538**	**12.397**	**27.538**	12.488	**16.253**	**9.571**	**13.360**	**13.385**	**21.394**	**10.078**
Original	-	**17.639**	9.941	**17.425**	**13.223**	**30.868**	**17.471**	**18.758**	10.178	**16.432**	**15.509**	**24.937**	**15.278**

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].

vision tasks. 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 is an expensive and restrictive objective function for training workload. The hyper-parameters mainly defining the computational burden of training (*i.e*. batch size and input size) mostly depend on this function. Therefore, a more elegant objective function can be designed for replacing with it. This architecture is implemented as a pre-processor for the downstream vision tasks, however it can be also designed as a single model where the filter removal is done just before the feature extraction, as in [47] for resizing.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017. 5

[2] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional siamese networks for object tracking. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 850–865, 2016. 1

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016. 1

[4] Simone Bianco, Claudio Cusano, Flavio Piccoli, and Raimondo Schettini. Artistic photo filter removal using convolutional neural networks. *Journal of Electronic Imaging*, 27:1, 12 2017. 1, 2, 5, 6, 7

[5] Simone Bianco, Claudio Cusano, and Raimondo Schettini. Artistic photo filtering recognition using cnns. In *International Workshop on Computational Color Imaging*, volume 10213 of *Lecture Notes in Computer Science*, pages 249–258. Springer, 2017. 1, 2

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 3

[8] Yu-Hsiu Chen, Ting-Hsuan Chao, Sheng-Yi Bai, Yen-Liang Lin, Wen-Chin Chen, and Winston H. Hsu. Filter-invariant image classification on social media photos. MM '15, page 855–858, New York, NY, USA, 2015. Association for Computing Machinery. 1, 2

[9] Wei-Ta Chu and Yu-Tzu Fan. Photo filter classification and filter recommendation without much manual labeling. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2019. 1, 2

[10] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin A. Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(9):1734–1747, 2016. 3

[11] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-Consistent Generative Adversarial Networks for One-Sided Unsupervised Domain Mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 5, 6, 7

[12] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv*, Aug 2015. 2

[13] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. 2017. 2

[14] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304. PMLR, 13–15 May 2010. 3, 5

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 1, 2

[18] Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR 2019*. ICLR, April 2019. 3

[19] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In Aasa Feragen, Marcello Pelillo, and Marco Loog, editors, *Similarity-Based Pattern Recognition*, pages 84–92, Cham, 2015. Springer International Publishing. 1

[20] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2, 5

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 2, 5, 6, 7

[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

[24] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Bharath Hariharan, Claire Cardie, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *ECCV*, 2020. 7, 8

[25] L. Karacan, Z. Akata, A. Erdem, and E. Erdem. Manipulating attributes of natural scenes via hallucination. *ACM Transactions on Graphics*, 39(1), Nov. 2019. 2

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[27] Furkan Kinli, Baris Ozcan, and Furkan Kirac. Instagram filter removal on fashionable images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 736–745, June 2021. 1, 2, 3, 4, 5, 6, 7, 8

[28] Gregory R. Koch. Siamese neural networks for one-shot image recognition. 2015. 1

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105. Curran Associates, Inc., 2012. 2

[30] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017. 1

[31] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998. 2

[32] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Kumar Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation viadisentangled representations. *International Journal of Computer Vision*, pages 1–16, 2020. 2, 5, 6, 7

[33] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[34] Ming Luo, Guihua Cui, and B. Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application*, 26:340 – 350, 10 2001. 6

[35] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 3

[36] Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 3

[37] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015. 1

[38] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 2020. 1, 2, 3, 5, 6, 7

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[40] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 1

[42] Mrinmoy Sen and Prasenjit Chakraborty. A deep convolutional neural network based approach to extract and apply photographic transformations. In Neeta Nain, Santosh Kumar Vipparthi, and Balasubramanian Raman, editors, *Computer Vision and Image Processing*, pages 155–162, Singapore, 2020. Springer Singapore. 1, 2

[43] Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Data-driven visual similarity for cross-domain image matching. *ACM Transaction of Graphics (TOG) (Proceedings of ACM SIGGRAPH ASIA)*, 30(6), 2011. 3

[44] Oleksii Sidorov. Conditional gans for multi-illuminant color constancy: Revolution or yet another approach? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 5, 6, 7

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 1, 2

[46] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[47] Hossein Talebi and Peyman Milanfar. Learning to resize images for computer vision tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 497–506, October 2021. 8

[48] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *Lecture Notes in Computer Science*, page 776–794, 2020. 3

[49] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. 3

[50] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition, 2020. 1

[51] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE, 2017. 1

[52] Zhe Wu, Zuxuan Wu, Bharat Singh, and Larry Davis. Recognizing instagram filtered images with feature de-stylization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12418–12425, Apr. 2020. 1

[53] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 2, 5, 6, 7

[56] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D. Cubuk, and Quoc V. Le. Rethinking pre-training and self-training. *CoRR*, abs/2006.06882, 2020. 1