information

MDPI

Article

Attentional Colorization Networks with Adaptive Group-Instance Normalization

Yuzhen Gao [1,2,*], Youdong Ding [1,2], Fei Wang [1,2] and Huan Liang [1,2]

[1] Shanghai Film Academy, Shanghai University, Shanghai 200444, China; ydding@shu.edu.cn (Y.D.); s852989587@shu.edu.cn (F.W.); Lianghuan@shu.edu.cn (H.L.)
[2] Shanghai Engineering Research Center for Motion Picture Special Effects, Shanghai 200444, China
* Correspondence: gyz18723239@shu.edu.cn

Received: 2 September 2020; Accepted: 9 October 2020; Published: 13 October 2020

Abstract: We propose a novel end-to-end image colorization framework which integrates attention mechanism and a learnable adaptive normalization function. In contrast to previous colorization methods that directly generate the whole image, we believe that the color of the significant area determines the quality of the colorized image. The attention mechanism uses the attention map which is obtained by the auxiliary classifier to guide our framework to produce more subtle content and visually pleasing color in salient visual regions. Furthermore, we apply Adaptive Group Instance Normalization (AGIN) function to promote our framework to generate vivid colorized images flexibly, under the circumstance that we consider colorization as a particular style transfer task. Experiments show that our model is superior to previous the state-of-the-art models in coloring foreground objects.

Keywords: colorization; attention mechanism; normalization

1. Introduction

Colorization is a method of propagating color to a grayscale image, and the colorized image should be reasonable in content and visually comfortable. This problem is highly ill-posed and dramatically ambiguous. Under normal circumstances, we can easily draw simple conclusions from the semantics of the scenes and the texture of the objects: the sky and the ocean are blue, and the grass and forests are green. However, for intricacy artifacts, it is difficult to reproduce their true color. Moreover, the huge workload of pure hand-painting has discouraged dedicated artists, not to mention the ordinary users. To solve these problems, an increasing number of researchers have begun to develop automatic coloring methods.

In this paper, we propose a novel end-to-end image colorization framework which integrates attention mechanism and an adaptive normalization function. Previous learning-based colorization methods generate the entire image directly, ignoring the attention mechanism in human perception. Our framework colors image from grayscale domain with the guidance of the attention map which is obtained by the encoder feature map and importance weights acquired from the auxiliary classifier. Both generator and discriminator are affiliated with attention maps to focus on the importance salient region. The attention map facilitates the color propagation in the generator, while optimizes the discriminator in detail by distinguishing the difference between colorized image and ground-truth images from color domain.

We consider colorization as a particular style transfer task where color information rather than a certain style is transferred to the image from grayscale domain. Thus, multiple normalization functions play a significant role in producing vivid colorized images. Inspired by the Adaptive Layer-Instance Normalization (AdaLIN) [1], we present the Adaptive Group-Instance Normalization (AGIN). The AGIN function promotes attention mechanism to guide our model to produce visually

appealing color flexibly and freely. Specifically, the parameters in AGIN is trained to learn the appropriate weights of Group Normalization (GN) [2] and Instance Normalization (IN) [3], where they perform well in the small batch size work and the individual picture work, respectively.

Our main contributions in this paper are as follows:

- We proposed a novel end-to-end framework for colorization with attention mechanism and AGIN which is a learnable normalization function.
- Our framework is guided by attention maps produced by the auxiliary classifier to know where the salient area is and to give more delicate color.
- AGIN is a learnable normalization function which helps our framework generate reasonable color flexibly and freely without transforming the network.

2. Related Works

2.1. Networks

The early color networks were very simple. For example, Koleini et al. [4] trained Artificial Neural Networks(ANN) by matching the pixels of gray image and color image, and Cheng et al. [5] first applied Convolutional Neural Networks (CNN) to the colorization of grayscale images. Putri et al. [6] inverted sketches into photos by predicting colors based on Deep CNN. In addition, Vitynskyi et al. [7] proposed a promising approach based on the neural-like structure of the Successive Geometric Transformations Model(SGTM), which improved the accuracy of image classification and regression methods. However, this kind of network has not been applied in the field of image translation.

Generative Adversarial Networks(GAN) can be an excellent solution for many ill-posed image processing problem and already have multiple remarkable achievements, such as colorization, image inpainting, super-resolution, style transfer, and so on. In pursuit of higher quality images, various novel GAN are beginning to prevail. DCGAN [8] uses CNN to implement generator and discriminator and replaces the pooling layer with strided convolutions. Although CNN and GAN are successfully combined, GAN is less robust. Isola et al. [9] applied Conditional GAN in image-to-image translation task and achieved wonderful results even with highly complex structure. Although GAN are growing rapidly, the inside of the generator is still as confusing as the black box. StyleGAN [10] does a good job in this respect by passing the latent code through non-linear mapping and affine transformation, and then through adaptive instance normalization in each convolutional layer control generator. CycleGAN [11] differs from the three GAN mentioned above in that it can handle unpaired data, which means that it pays more attention to the migration of features between images and images. Moreover, cycle consistency loss can avoid the contradiction between generators.

2.2. Colorization

The scribble-based colorization methods diffuse the user's color hints (such as color points, strokes, and blocks) to the entire grayscale image, while the color propagation is based on low-scale features. Levin et al. [12] and Zhang et al. [13] proposed the scribble-based methods diffusing the color of the strokes prompted by the user to the entire image. Sangkloy's method [14] allows the user to control the generation of color images through sketches and sparse strokes. Levin et al. [12] first proposed that adjacent pixels with similar luminance also have similar colors. According to this theory, boundary information is not even required from a user's sparse simple stroke to a complex full color image. More advanced work extend from the luminance information to the textures, and solve color bleeding with edge guidance. The system developed by Zhang et al. [13] can fuse user's sparse low-level stroke information and high-level semantic information to color the grayscale image. Although all the above methods relieve the user's burden to a certain extent, they still need abundance or less manual intervention. Moreover, the rationality of image colors depends to a large extent on the user's strokes, which means that the image quality is constrained by the user's professionalism and rich experience.

Therefore, exemplar-based colorization which is one of fully automatic methods is prevailing to reduce the burden on users.

The exemplar-based colorization methods provide a similar color reference map to the target grayscale image for more direct coloring. Welsh et al. [15] focus on the global information of the image, which colors the target image by matching the brightness information between the reference image and each pixel of the target image. Tai et al. [16] paid attention to the local information of the image and segmented the image with soft boundaries to achieve color transfer and propagation. However, for substantial content regions, it is difficult to obtain low-level features by these methods. The system designed by He et al. [17,18] can recommend appropriate references based on luminance and semantic information, reducing the steps of manually screening, then achieving full-automatic coloring. Yoo et al. [19] use small-scale data to produce high quality images with a colorization model which has memory components. The above methods all suffered from the same problem that the reference does not exactly match all brightness information in source domain. Then, how to select an applicable reference becomes a challenge. Hence, the learning-based method learns color transfer pattern from large-scale data and applies different loss functions to restrain the quality of the generated color images.

The learning-based colorization methods obtain networks by training on large-scale data, and networks can automatically generate various results without user intervention. Almost the same period, Larsson et al. [20]; Iizuka et al. [21]; Zhang et al. [22] proposed similar methods with different loss functions based on CNN. Larsson and Zhang applied classification loss and Iizuka applied L_2 regression loss. Isola et al. [9] believe that L_1 loss can reduce image blur, so the combination of L_1 loss and GAN loss is applied. To produce more diverse colorization results, Messaoud et al. [23] established a conditional random field, and Cao et al. [24] developed a fully convolutional generator with multi-layer noise. Zhao et al. [25] exploit pixel-level semantic information to guide the generator.

2.3. Class Activation Mapping

Zhou et al. [26] pioneered a class activation mapping, which uses global average pooling to obtain the weights of each convolutional layer and multiply the weighted sum by each feature map. We can input an image of any size, as long as it is simply upsampled to the source image size, salient area of the image will be showed. Grad-CAM [27] is an improvement based on CAM which uses the global average of the gradient to calculate weights, and no need to change the network structure.

2.4. Normalization

We consider colorization as the transfer of color features to the target grayscale image, which is the same as the style transfer. At the same time, the normalization function used in style transfer can also be used in colorization through adjustment. Although Batch Normalization (BN) [28] has made good achievements, a large number of improvement methods are emerging, such as: Layer Normalization [29], GN [2], etc. BN shows robustness in a wide range of batch sizes, even when it is small. In [3], because its results depend on a certain image instance, they achieve remarkable results in style transfer. To have a better image effect, a composite method such as Batch-Instance Normalization(BIN) [30], Adaptive Instance Normalization (AdaIN) [31], and Conditional Instance Normalization (CIN) [32] is often used instead of using IN alone. CIN improves the layer affine parameters of IN. By using the same network parameters, different style effects can be obtained. BIN can selectively normalize the style. AdaIN generates an image of any given type by using adaptive affine parameters, which functions as an exchange style.

3. Network

For the existing grayscale image domain X_g and color image domain X_c, our purpose is to train a mapping $G_{g \to c}$ that can generate X_c domain images from the X_g domain images. Our framework comprises two generators (G_g, G_c) and two discriminators (D_g, D_c), which both incorporate attention

mechanism. The framework structure is based on CycleGAN, so only G_g and D_c in the forward cycle are explained here (see Figure 1). The reverse cycle is consistent with its principle. To distinguish the input of the forward cycle is represented by x and the input of the reverse cycle is represented by y.

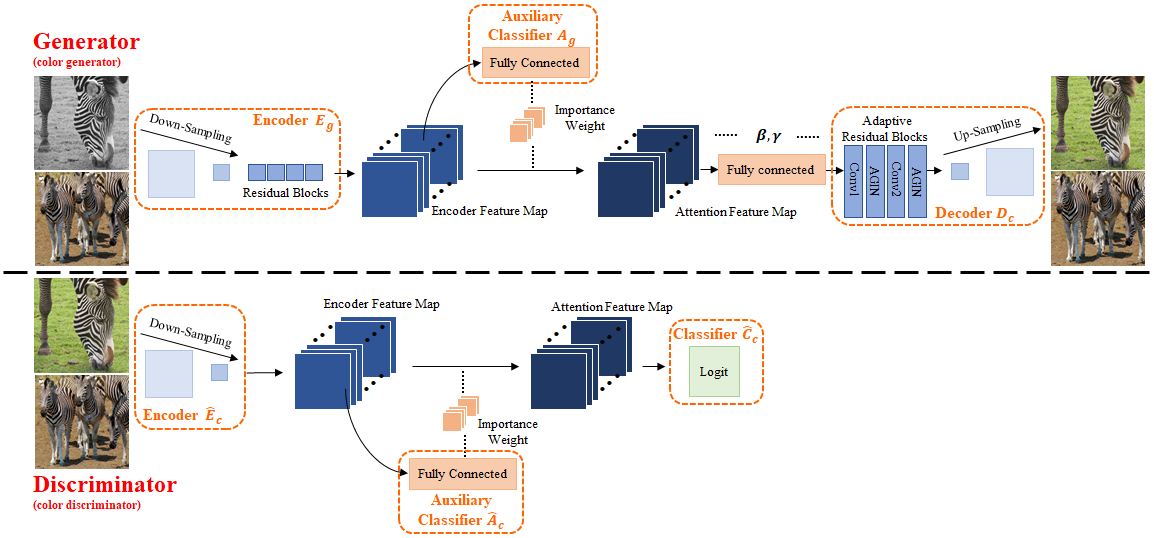

Figure 1. The architecture of our framework, and the details are covered in Section 3.1.

3.1. Model

3.1.1. Generator

Let $x \in \{X_g, X_c\}$, and $G_g(x)$ represents the output of an image from the gray image domain to the color image domain. G_g is generator which consists of an encoder E_g, a decoder D_c and an auxiliary classifier A_g, where $E_g(x)$ is the activation map of encoder, $E_g^i(x)$ is the i-th activation map, and $E_g^{i_{(a,b)}}(x)$ is the value at (a, b). $A_g(x)$ represents the degree of correspondence between x and the image in the X_g domain. The auxiliary classifier of CAM [26] uses global average pooling to learn the importance weights of the i-th activation map. We exploit the combination of Global Average Pooling (GAP) and Median Pooling (MP) to learn the edge feature better, and the importance weights of the i-th activation map is W_g^i. Therefore, $A_g(x) = \sum_{i=1}^{n} w_g^i \sum_{a,b} E_g^{(a,b)}(x)$. A set of X_g domain-specific attention feature map $M_g(x)$ can be generated by the importance weights and the previous convolutional layers, where $M_g(x) = \{W_g^i(x) * E_g^i(x) | 1 \leq i \leq n\}$, and n is the amount of encoder activation maps. Inspired by AdaLIN [1], We integrate the residual blocks with AGIN which is a fusion of GN [2] and IN [3].

$$AGIN(M, \beta, \gamma) = \gamma \cdot (\rho \cdot m_G + (1 - \rho) \cdot m_I) + \beta, \tag{1}$$

$$m_G = \gamma \left(\frac{M - \mu_G}{\sigma_G + \epsilon} \right) + \beta, m_I = \gamma \left(\frac{M - \mu_I}{\sigma_I + \epsilon} \right) + \beta, \tag{2}$$

$$\rho \leftarrow clip_{[0,1]}(\rho - \tau \Delta \rho) \tag{3}$$

where μ_G, μ_I and σ_G, σ_I are the mean of x on group scale, channel scale and standard deviation respectively. β and γ are affine transformation parameters with predictions generated by fully connected layer. τ is learning rate. $\rho \in [0, 1]$, where ρ is restricted by $\Delta\rho$ which is a dynamically computed parameter vector(e.g., the gradient). The value of ρ represents the choice of normalization method. If the value approaches 0, it means that this task is more suitable for IN, and if the value approaches 1, it means that GN is more important for this task.

3.1.2. Discriminator

$G_g(X_g)$ is a domain which contains generated fake color images. Let $x \in \{X_g, G_g(X_g)\}$ represent x from the color domain and the fake color domain. Our discriminator $\hat{D}_c$ is comprised of an encoder $\hat{E}_c$, a classifier $\hat{C}_c$ and an auxiliary classifier $\hat{A}_c$. With a input x, we acquire from the encoder with feature maps $\hat{E}_c(x)$ which can be used to obtain the importance weights $\hat{W}_c$. The attention feature maps is calculated using $\hat{M}_c(x) = \{\hat{W}_c^i(x)\hat{E}_c^i(x)|1 \leq i \leq n\}$, and exploited by $\hat{D}_c$. $\hat{E}_c$ is trained by $\hat{A}_c$ which along with $\hat{D}_c$ are trained to discriminate where x belongs to, X_g or $G_g(X_g)$.

3.2. Loss

The full loss function of our framework is composed of four parts.

3.2.1. Adversarial Loss

Both forward mapping G_g and reverse mapping G_c apply adversarial losses:

$$\begin{aligned}\mathcal{L}_{adv,g}(G, \hat{D}_c, X, Y) = \mathbb{E}_{y \sim X_c}[(\hat{D}_c(y))^2] \\ + \mathbb{E}_{x \sim X_g}[((1 - \hat{D}_c(G_g(x)))^2]\end{aligned} \tag{4}$$

where G_g aims to generate fake images $G_g(x)$ to fool $\hat{D}_c$, and $\hat{D}_c$ tries to distinguish whether the generated images are from domain X_g or X_c. Concisely, function $\min_{G_g} \max_{\hat{D}_c} \mathcal{L}_{\text{adv,G}_\text{g}}(G_g, \hat{D}_c, X, Y)$ represents G_g tries to minimize this function, on the contrary, $\hat{D}_c$ needs to maximize it. Similarly, the function of revers mapping apply the $\min_{G_c} \max_{\hat{D}_g} \mathcal{L}_{\text{adv,G}_\text{c}}(G_c, \hat{D}_g, Y, X)$.

3.2.2. Cycle Consistency Loss

To avoid the color of all generated images from X_g tending to one image in the X_c and each image generated by G_g can also be restored from G_c to x, We introduce cycle consistency loss:

$$\begin{aligned}\mathcal{L}_{cycle}(G_g, G_c) = \mathbb{E}_{x \sim X_g(x)}[\|G_c(G_g(x)) - x\|_1] \\ + \mathbb{E}_{y \sim X_c(y)}[\|G_g(G_c(y)) - y\|_1]\end{aligned} \tag{5}$$

3.2.3. Content Loss

With the purpose of ensuring the input and output are similar in content, we apply content loss to restrain generators.

$$\mathcal{L}_{Content}(G_g) = \mathbb{E}_{x \sim X_c}[|x - G_g(x)|_1] \tag{6}$$

3.2.4. CAM Loss

For $x \in \{X_g, X_c\}$ we trained G_g and D_c with the parameters inferred from auxiliary classifiers A_g and $\hat{A}_c$. With CAM losses:

$$\begin{aligned}\mathcal{L}_{cam,G_g}(A_g) = -(\mathbb{E}_{x \sim X_g}[\log(A_g(x)) \\ + \mathbb{E}_{x \sim X_c}[\log(1 - A_g(x)]\end{aligned} \tag{7}$$

G_g can be aware of where improvement is needed to generate images more similar to the images in X_c.

$$\begin{aligned}\mathcal{L}_{cam,\hat{D}_c}(\hat{A}_c) = \mathbb{E}_{x \sim X_c}[(\hat{A}_c(x))^2] \\ + \mathbb{E}_{x \sim X_g}[(1 - \hat{A}_c(G_g(x)))^2]\end{aligned} \tag{8}$$

D_c gets to know where to identify details that can distinguish the difference between two domain images.

3.2.5. Full Function

Our overall loss function is:

$$\min_{G_g,G_c,A_g,A_c} \max_{D_g,D_c,\hat{A}_g,\hat{A}_g} \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{cycle} + \lambda_3 \mathcal{L}_{content} + \lambda_4 \mathcal{L}_{cam} \tag{9}$$

where $\lambda_1 = 1, \lambda_2 = 9, \lambda_3 = 9, \lambda_4 = 999$.

4. Implementation

4.1. Architecture

Our generator is composed of an encoder, a decoder, and an auxiliary classifier. The encoder consists of two convolutional layers of down-sampling with the stride size of two and four residual blocks. The decoder consists of two up-sampling convolutional layers with the stride size of one and four adaptive residual blocks which is equipped with AGIN, unlike in the decoder where only instance normalization is used. We use two scales of PatchGAN [9] in the discriminator network for identification, in which the size of the local patch size is 70×70 and the size of the global patch size is 286×286. In discriminator, we use spectral normalization. The ReLUs used in the generator are not leaky, while ReLUs in the discriminator are leaky, with a slope of 0.2.

4.2. Training

To expand the training data, we first resized input images with the size of 256×256 to 286×286, and then randomly cropped back to the size of 256×256. The batch size in experiment is set to one. We applied Adam [33] in training with a learning rate of 0.0002 and momentum parameters $\beta_1 = 0.5, \beta_2 = 0.999$.

5. Experiments

5.1. Dataset

We train networks on COCO [34] and VisualGenome [35]. For the training, all images are resized to 256×256. In addition, all grayscale images are obtained by grayscale conversion of color images.

5.2. Comparisons with State-of-the-Art

We first get the results of the proposed model (see Figure 2), and also conduct ablation experiments on attention mechanism to prove its validity. We compare our model with the colorization state-of-the-art(Zhang et al. [22]; Larsson et al. [20]; Iizuka et al. [21]). The colorization results are shown in Figure 3. From the overall chrominance, the results of our model are more realistic and convincing (row 1, 3). Our model is also superior in terms of detail coloring (row 2, 4). Also, Our model can color foreground objects correctly while carefully handling edge problems (5, 6). In addition, we compare the performance of our model with that of other outstanding image translation models (see Figure 4). Furthermore, the qualitative and quantitative evaluations are also used to evaluate the quality of generated images.

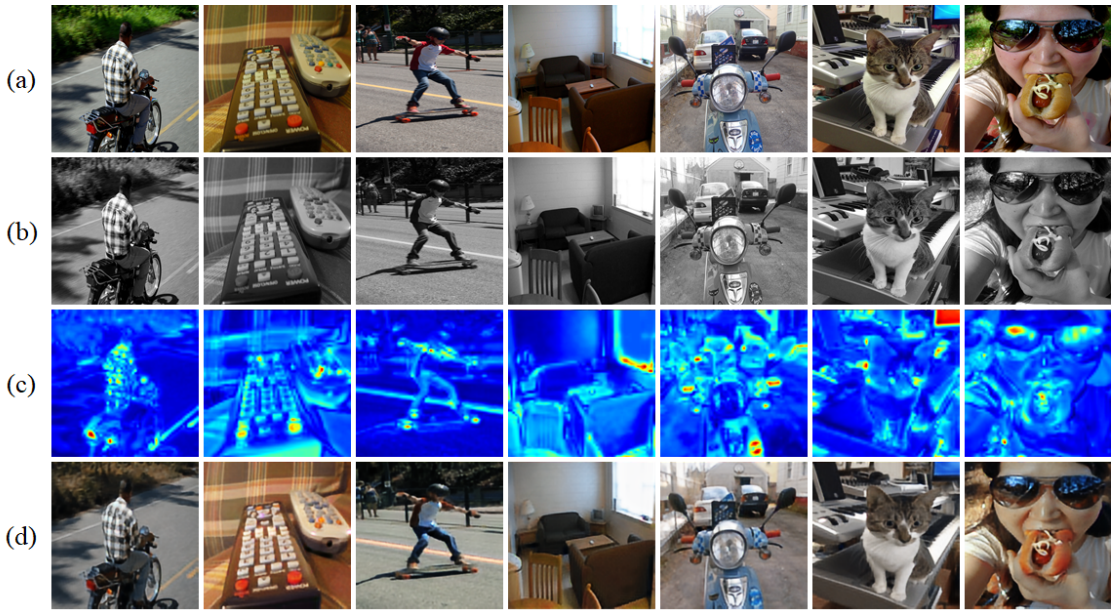

Figure 2. Colorized results and their visualization of the attention maps: (**a**) Ground truth, (**b**) Targets, (**c**) Attention maps, (**d**) Our results.

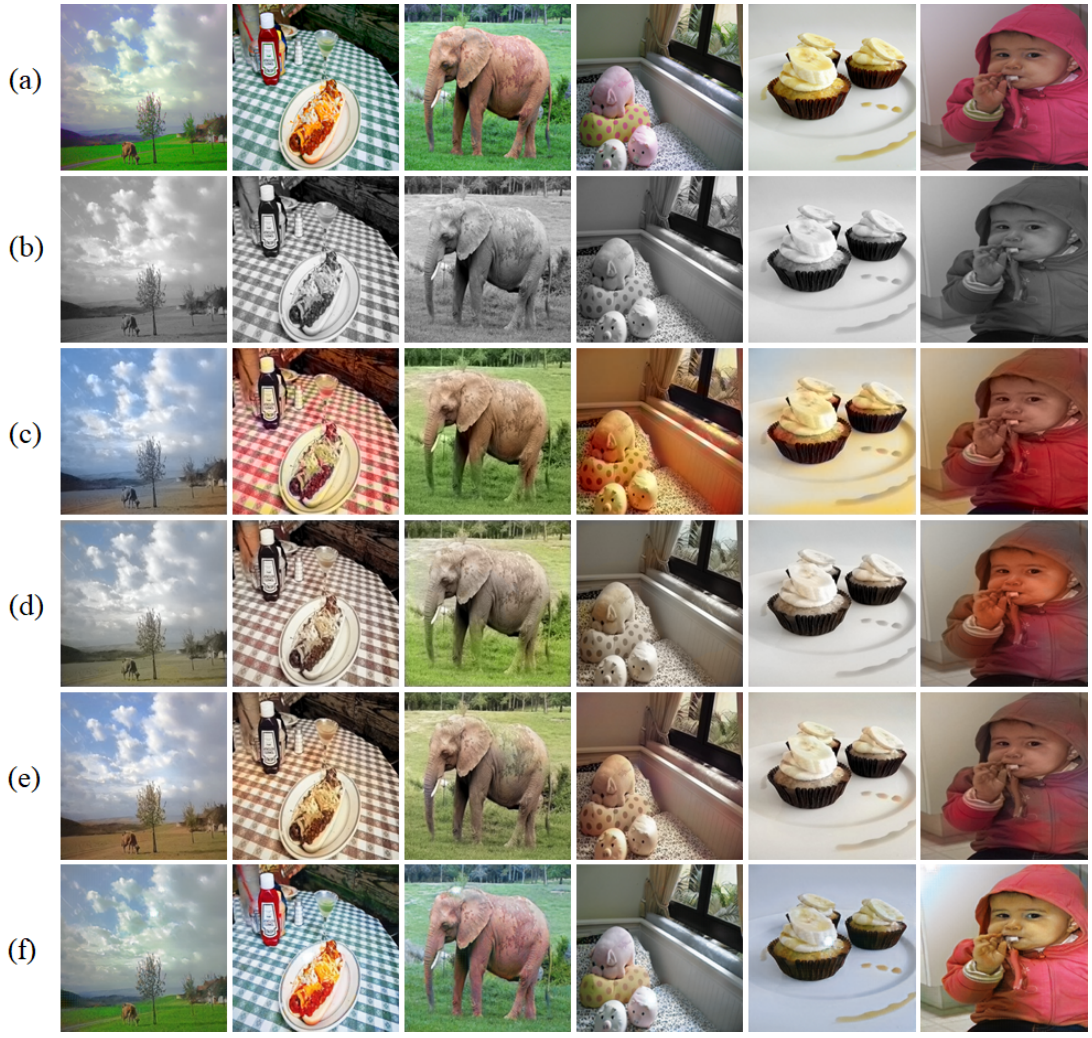

Figure 3. Comparison on different colorization methods: (**a**) Ground truth, (**b**) Targets, (**c**) Zhang et al. [22], (**d**) Larsson et al. [20], (**e**) Iizuka et al. [21], (**f**) Ours.

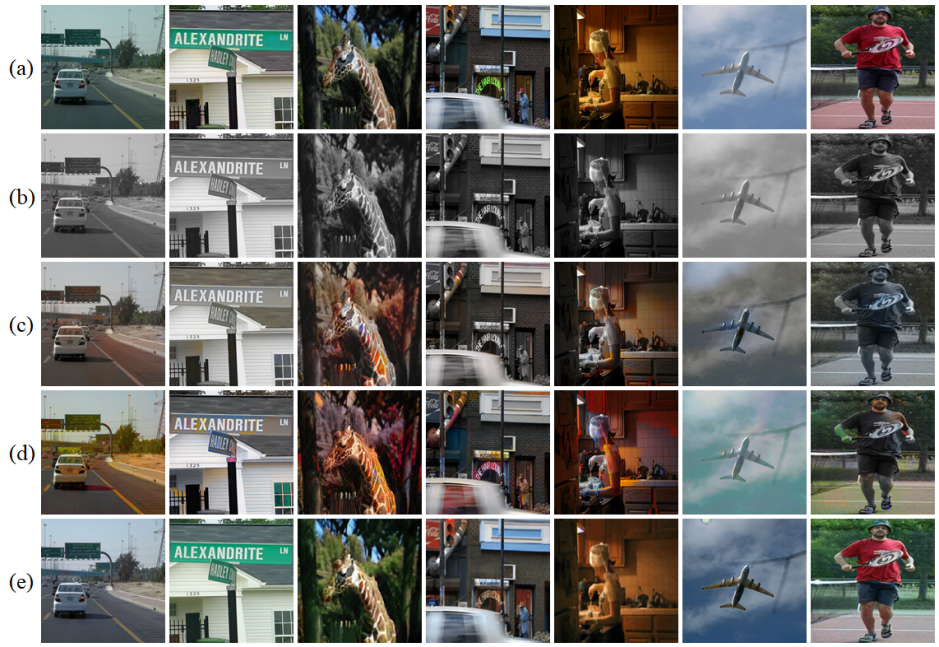

Figure 4. Comparison on different images translation methods: (**a**) Ground truth, (**b**) Targets, (**c**) CycleGAN, (**d**) Pix2Pix, (**e**) Ours.

5.3. CAM Ablation Experiment

To prove the effectiveness of attention mechanism, we conducted CAM ablation experiment. We can find from the CAM ablation experiment results (see Figure 5) that color bleeding problem (row d, columns 1, 3, 4, 5, 6) and color failure caused by blurring boundaries (row d, columns 2, 7) are common in the results without CAM. The addition of CAM can make the model pay more attention to the key areas when coloring and deal with the boundary more carefully. The colorized results with CAM is shown in Figure 5e which confirms that attention mechanism plays a positive role in the color bleeding problem of colorization.

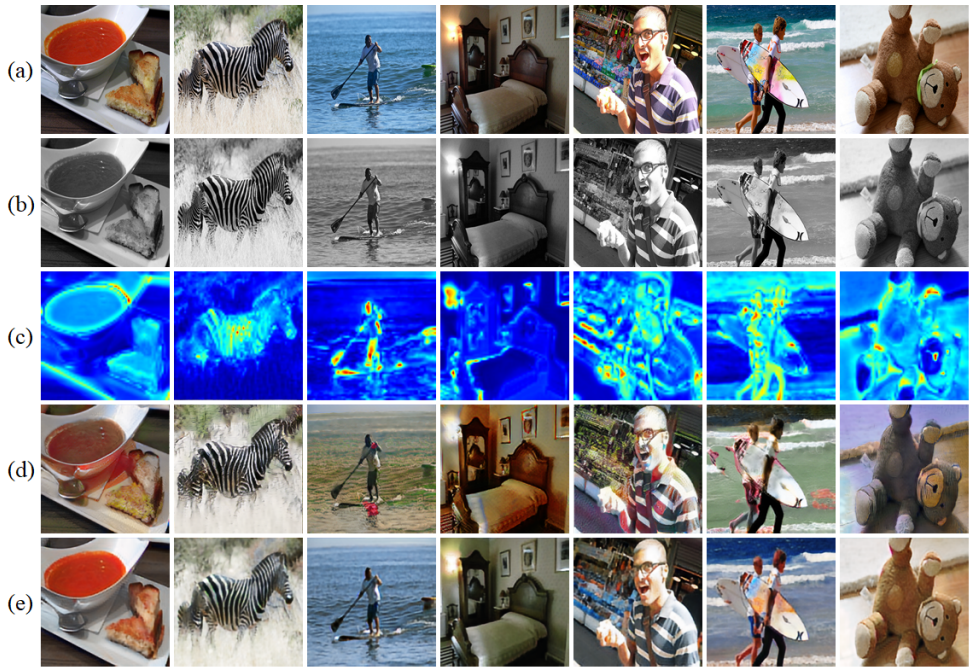

Figure 5. Colorized results on the CAM ablation experiment: (**a**) Ground truth, (**b**) Targets, (**c**) Attention maps, (**d**) Our results without CAM, (**e**) Our results with CAM.

5.4. AGIN Ablation Experiment

We consider colorization as a particular style transfer task, i.e., transferring color information rather than a specific style. We use AGIN to combine the advantages of GN and IN for better color transferring. As we introduced in Section 3.1, the value of ρ is learnable. When the value of ρ is learn to approach 0, it means that the normalization layers tend to adopt IN. When the value of ρ is learn to approach 1, it means that the normalization layers tend to adopt GN. Hence, we conducted AGIN ablation experiment to confirm that the AGIN used in generator is beneficial to produce vivid color. GN computes the group-wise features, so unreasonable colors may appear in the generated image (see Figure 6, row c). IN calculates the channel-wise features, too many of them are retained, so that the overall chrominance of the colorized image is dark and the contrast is not enough (see Figure 6, row d). Therefore, we believe that AGIN can combine the advantages of GN and IN to allocate the weight adaptively, which can make colorized images more visually pleasing.

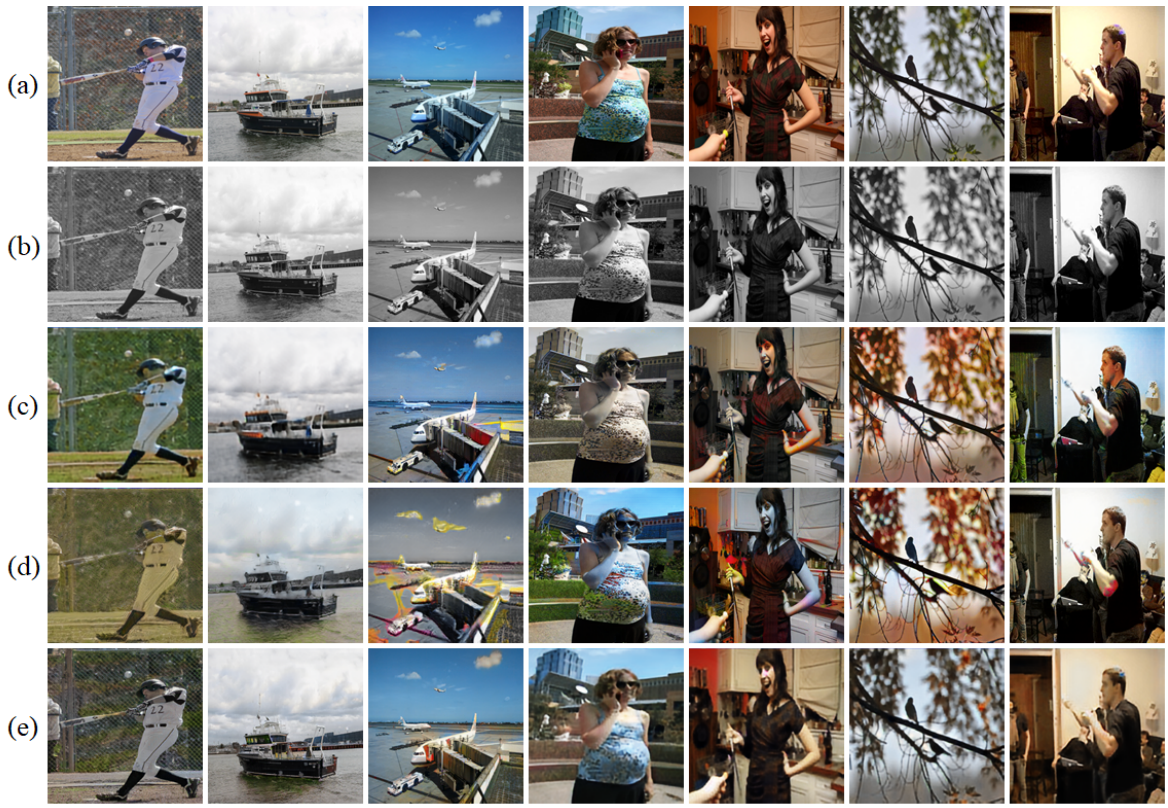

Figure 6. Colorized results on the CAM ablation experiment: (**a**) Ground truth, (**b**) Targets, (**c**) Results using GN only, (**d**) Results using IN only, (**e**) Results using AGIN.

5.5. Qualitative and Quantitative Evaluations

To evaluate the quality of the colorized images, we conducted a preference study. 197 observers(including researchers and people without any colorization knowledge) are asked to select the best colorized image from images generated by different methods. As can be seen from Table 1, the results of our method were approved by the majority of users. Table 2 also shows that our method can produce higher quality images than other methods.

To be visually pleasing, we also evaluated the naturalness of the colorized images. We compare our model with the colorization state-of-the-art (Zhang et al. [22]; Larsson et al. [20]; Iizuka et al. [21]). 15 observers are randomly shown 500 images (ground truth images, and colorized images generated by our method and the state-of-the-art methods, 100 images each) one at a time, and asked to judge the image is natural to themselves or not. We let observers intuitively determine whether the image is

natural. Similarly, we compare the naturalness of our method with that of image translation. Table 3 shows that 93.21% of colorized images generated by our method are considered as natural, which bears out our model is able to generate natural and visually pleasing color images.

Table 1. Qualitative score on colorized results.

Method \ Dataset	COCO	VisualGenome	Landscape
Zhang et al. [22]	15.74%	13.20%	10.15%
Larsson et al. [20]	8.12%	4.06%	5.58%
Iizuka et al. [21]	7.11%	9.64%	7.61%
Ours	**69.04%**	**73.10%**	**77.65%**

Table 2. Qualitative score compared with image translation methods.

Method \ Dataset	COCO	VisualGenome	Landscape
CycleGAN	16.75%	14.72%	11.68%
DCGAN	10.65%	10.15%	10.66%
Pix2Pix	8.62%	7.11%	7.11%
Ours	**63.95%**	**68.02%**	**70.56%**

Table 3. Naturalness evaluation.

Method	Naturalness (Mean)
Zhang et al. [22]	87.53%
Larsson et al. [20]	85.58%
Iizuka et al. [21]	89.13%
CycleGAN	79.46%
DCGAN	75.71%
Pix2Pix	71.24%
Ours	**93.21%**
Ground truth	98.86%

Evaluating the results of colorization methods is a very subjective challenge, and both quantitative and qualitative evaluations are difficult. As for qualitative evaluation, it is very difficult to make qualitative analysis on such a highly ill-posed problem as colorization. The peak signal-to-noise ratio(PSNR) is widely used in the field of image processing, and many colorization methods (Larsson et al. [20]; He et al. [17]) also use PSNR to evaluate image quality The comparison results are shown in Table 4. Our method has a higher PSNR than other methods, which proves that our method can produce more realistic and higher-quality images.

Table 4. Quantitative evaluation.

Method	PSNR (dB)
Zhang et al. [22]	22.90
Larsson et al. [20]	24.25
Iizuka et al. [21]	23.86
Ours	**24.43**
Ground truth	NA

6. Limitations and Discussion

Colorization is a highly ill-posed and ambiguous problem. We can easily infer the colors of the oceans and forests, but there is often no unique solution to the colors of the clothes people wear. In addition, the limitation of our method is that we can misjudge objects (see Figure 7, row 1, the land

was wrongly colored green) and color the artifacts incorrectly (see Figure 7, row 2, the kite was incorrectly colored).

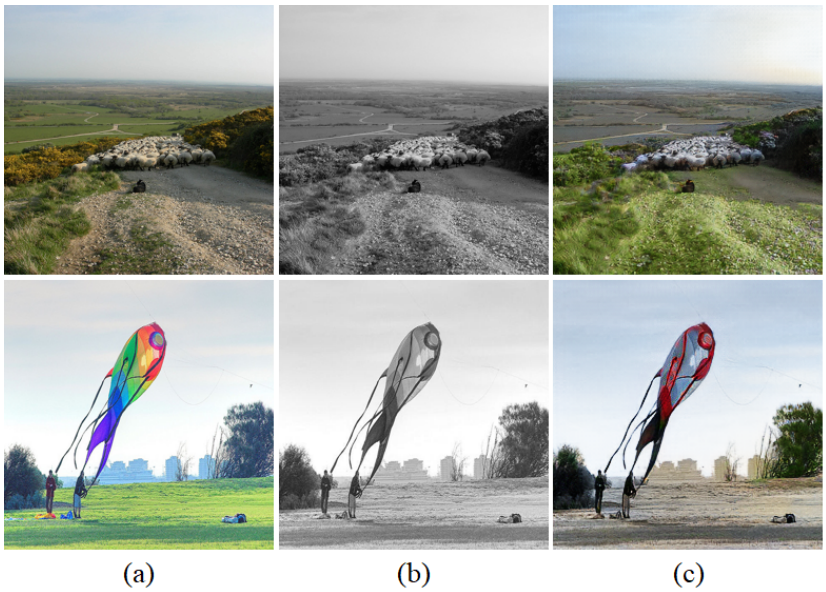

Figure 7. Failure cases: (**a**) Ground truth, (**b**) Targets, (**c**) Our results.

Learning-based methods are data-driven. As long as the dataset is large enough and the content is rich enough, the better the colorized image quality is theoretically. At the same time, the same object will present different colors in different environments, weather and seasons, i.e., changeable lighting conditions also bring challenges to colorization. Meanwhile, to make the colors of the artifacts in the dataset representative and universal may still need to be manually labeled. However, the possibility of semantic colorization is not explored in this paper.

7. Conclusions

In this paper, we proposed a novel end-to-end colorization framework that integrates attention mechanism and AGIN which is a leranable adaptive normalization function. Attention maps produced by auxiliary classifier serve as a guide for the generator to focus on details that are easily overlooked. The addition of attention mechanism solves the problem of color bleeding well. Furthermore, AGIN plays an integral role in flexibly and freely producing vivid colors when the dataset contains images with complex content and diverse scenes. Mass experimental results verify that our framework can transfer reasonable and visually pleasing color to black and white images. In addition, our method is superior to other state-of-the-art GAN-based colorization methods.

In the future research, from the perspective of algorithm, the network structure can continue to be optimized to reduce the model training time, and the video colorization algorithm can also be taken into consideration. From the perspective of dataset, a new dataset can be built to train the network, which can be applied in the colorization of legacy photos and old movies.

Author Contributions: Conceptualization, Y.G.; Data curation, H.L.; Investigation, F.W.; Methodology, Y.G.; Software, Y.G.; Validation, Y.G.; Writing—original draft, Y.G.; Writing—review & editing, Y.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kim, J.; Kim, M.; Kang, H.; Lee, K. U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation. *arXiv* **2019**, arXiv:1907.10830.
2. Wu, Y.; He, K. Group Normalization. In Proceedings of the Computer Vision—ECCV 2018—5th European Conference, Part XIII, Munich, Germany, 8–14 September 2018; pp. 3–19.
3. Ulyanov, D.; Vedaldi, A.; Lempitsky, V.S. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv* **2016**, arXiv:1607.08022.
4. Koleini, M.; Monadjemi, S.A.; Moallem, P. Film Colorization Using Texture Feature Coding and Artificial Neural Networks. *J. Multim.* **2009**, *4*, 240–247. [CrossRef]
5. Cheng, Z.; Yang, Q.; Sheng, B. Deep Colorization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 415–423.
6. Putri, V.K.; Fanany, M.I. Sketch plus colorization deep convolutional neural networks for photos generation from sketches. In Proceedings of the 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Yogyakarta, Indonesia, 19–21 September 2017; pp. 1–6.
7. Vitynskyi, P.; Tkachenko, R.; Izonin, I.; Kutucu, H. Hybridization of the SGTM neural-like structure through inputs polynomial extension. In Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 21–25 August 2018; pp. 386–391.
8. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Proceedings of the 4th International Conference on Learning Representations (ICLR 2016), San Juan, Puerto Rico, 2–4 May 2016.
9. Isola, P.; Zhu, J.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976.
10. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019), Long Beach, CA, USA, 16–20 June 2019; pp. 4401–4410.
11. Zhu, J.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 2242–2251.
12. Levin, A.; Lischinski, D.; Weiss, Y. Colorization using optimization. *ACM Trans. Graph.* **2004**, *23*, 689–694. [CrossRef]
13. Zhang, R.; Zhu, J.; Isola, P.; Geng, X.; Lin, A.S.; Yu, T.; Efros, A.A. Real-time user-guided image colorization with learned deep priors. *arXiv* **2017**, arXiv: 1705.02999.
14. Sangkloy, P.; Lu, J.; Fang, C.; Yu, F.; Hays, J. Scribbler: Controlling Deep Image Synthesis with Sketch and Color. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 6836–6845.
15. Welsh, T.; Ashikhmin, M.; Mueller, K. Transferring color to greyscale images. *ACM Trans. Graph.* **2002**, *21*, 277–280. [CrossRef]
16. Tai, Y.; Jia, J.; Tang, C. Local Color Transfer via Probabilistic Segmentation by Expectation-Maximization. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20–26 June 2005, San Diego, CA, USA, 2005; pp. 747–754.
17. He, M.; Chen, D.; Liao, J.; Sander, P.V.; Yuan, L. Deep exemplar-based colorization. *ACM Trans. Graph.* **2018**, *37*, 1–16. [CrossRef]
18. Zhang, B.; He, M.; Liao, J.; Sander, P.V.; Yuan, L.; Bermak, A.; Chen, D. Deep Exemplar-Based Video Colorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019), Long Beach, CA, USA, 16–20 June 2019; pp. 8052–8061.
19. Yoo, S.; Bahng, H.; Chung, S.; Lee, J.; Chang, J.; Choo, J. Coloring With Limited Data: Few-Shot Colorization via Memory Augmented Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019), Long Beach, CA, USA, 16–20 June 2019; pp. 11283–11292.
20. Larsson, G.; Maire, M.; Shakhnarovich, G. Learning Representations for Automatic Colorization. In Proceedings of the Computer Vision—ECCV 2016—14th European Conference, Part IV, Amsterdam, The Netherlands, 11–14 October 2016; pp. 577–593.

21. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Trans. Graph.* **2016**, *35*, 1–11. [CrossRef]
22. Zhang, R.; Isola, P.; Efros, A.A. Colorful Image Colorization. In Proceedings of the Computer Vision—ECCV 2016—14th European Conference, Part III, Amsterdam, The Netherlands, 11–14 October 2016; pp. 649–666.
23. Messaoud, S.; Forsyth, D.A.; Schwing, A.G. Structural Consistency and Controllability for Diverse Colorization. In Proceedings of the Computer Vision—ECCV 2018—15th European Conference, Part VI, Munich, Germany, 8–14 September 2018; pp. 603–619.
24. Cao, Y.; Zhou, Z.; Zhang, W.; Yu, Y. Unsupervised Diverse Colorization via Generative Adversarial Networks. In Proceedings of the Machine Learning and Knowledge Discovery in Databases—European Conference (ECML PKDD 2017), Part I, Skopje, Macedonia, 18–22 September 2017; pp. 151–166.
25. Zhao, J.; Han, J.; Shao, L.; Snoek, C.G.M. Pixelated Semantic Colorization. *arXiv* **2019**, arXiv:1901.10889.
26. Zhou, B.; Khosla, A.; Lapedriza, À.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June, 2016; pp. 2921–2929.
27. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 618–626.
28. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning (ICML 2015), Lille, France, 6–11 July 2015; pp. 448–456.
29. Ba, L.J.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
30. Nam, H.; Kim, H. Batch-Instance Normalization for Adaptively Style-Invariant Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018 (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018; pp. 2563–2572.
31. Huang, X.; Belongie, S.J. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 1510–1519.
32. Dumoulin, V.; Shlens, J.; Kudlur, M. A Learned Representation For Artistic Style. In Proceedings of the 5th International Conference on Learning Representations (ICLR 2017), Toulon, France, 24–26 April 2017.
33. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015.
34. Lin, T.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. ECCV (5). *Lect. Notes Comput. Sci.* **2014**, *8693*, 740–755.
35. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D.A.; et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [CrossRef]