Eliminating Gradient Conflict in Reference-based Line-Art Colorization

Zekun Li[1], Zhengyang Geng[2], Zhao Kang[1]*, Wenyu Chen[1], and Yibo Yang[3]

[1] University of Electronic Science and Technology of China, Chengdu, China
kunkun0w0@std.uestc.edu.cn;{zkang,cwy}@uestc.edu.cn
[2] Peking University, School of AI, Beijing, China
ZhengyangGeng@gmail.com
[3] JD Explore Academy
ibo@pku.edu.cn

Abstract. Reference-based line-art colorization is a challenging task in computer vision. The color, texture, and shading are rendered based on an abstract sketch, which heavily relies on the precise long-range dependency modeling between the sketch and reference. Popular techniques to bridge the cross-modal information and model the long-range dependency employ the attention mechanism. However, in the context of reference-based line-art colorization, several techniques would intensify the existing training difficulty of attention, for instance, self-supervised training protocol and GAN-based losses. To understand the instability in training, we detect the gradient flow of attention and observe gradient conflict among attention branches. This phenomenon motivates us to alleviate the gradient issue by preserving the dominant gradient branch while removing the conflict ones. We propose a novel attention mechanism using this training strategy, Stop-Gradient Attention (SGA), outperforming the attention baseline by a large margin with better training stability. Compared with state-of-the-art modules in line-art colorization, our approach demonstrates significant improvements in Fréchet Inception Distance (FID, up to 27.21%) and structural similarity index measure (SSIM, up to 25.67%) on several benchmarks. The code of SGA is available at https://github.com/kunkun0w0/SGA.

Keywords: GAN, Attention Mechanism, Stop-gradient

1 Introduction

Reference-based line-art colorization has achieved impressive performance in generating a realistic color image from a line-art image [32,59]. This technique is in high demand in comics, animation, and other content creation applications [55,2]. Different from painting with other conditions such as color strokes [53,12], palette [56], or text [25], using a style reference image as condition input not only provides richer semantic information for the model but also eliminates

---

* Corresponding author

the requirements of precise color information and the geometric hints provided by users for every step. Nevertheless, due to the huge information discrepancy between the sketch and reference, it is challenging to correctly transfer colors from reference to the same semantic region in the sketch.

Several methods attempt to tackle the reference-based colorization by fusing the style latent code of reference into the sketch [31,38,36]. Inspired by the success of the attention mechanism [41,46], researchers adopt attention modules to establish the semantic correspondence and inject colors by mapping the reference to the sketch [28,55,54]. However, as shown in Figure 1, the images generated by these methods often contain color bleeding or semantic mismatching, indicating considerable room for improving attention methods in line-art colorization.
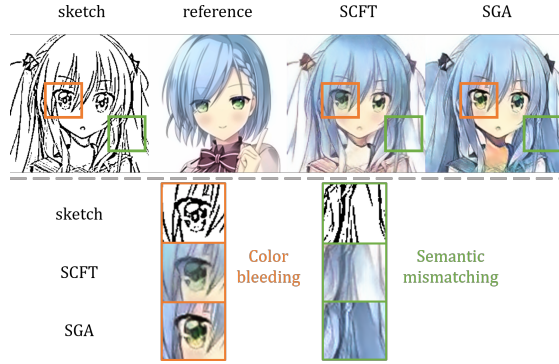

Fig. 1: The comparison between the images produced by SCFT [28] and SGA (Ours). SCFT subjects to color bleeding (orange box) and semantic mismatching (green box).

There are many possible reasons for the deficiency of line-art colorization using attention: model pipeline, module architecture, or training. Motivated by recent works [13,5] concerning the training issues of attention models, we are particularly interested in the training stability of attention modules in line-art colorization. It is even more challenging to train attention models in line-art colorization because state-of-the-art models [28] deploy multiple losses using a GAN-style training pipeline, which can double the training instability. Therefore, we carefully analyze the training dynamics of attention in terms of its gradient flow in the context of line-art colorization. We observe the gradient conflict phenomenon, namely, a gradient branch contains a negative cosine similarity with the summed gradient.

To eliminate the gradient conflict, we detach the conflict one while preserving the dominant gradient, which ensures that the inexact gradient has a positive cosine similarity with the exact gradient and meet theory requirements [14,50]. This training strategy visibly boosts the training stability and performance compared with the baseline attention colorization models. Combined with archi-

tecture design, this paper introduces **S**top-**G**radient **A**ttention, **SGA**, whose training strategy eliminates the gradient conflict and helps the model learn better colorization correspondence. SGA properly transfers the style of reference images to the sketches, establishing accurate semantic correspondence between sketch-reference pairs. Our experiment results on several image domains show clear improvements over previous methods, *i.e.,* up to 27.21% and 25.67% regarding FID and SSIM, respectively.

Our contributions are summarized as follows:

- We reveal the gradient conflict in attention mechanism for line-art colorization, *i.e.,* a gradient branch contains a negative cosine similarity with the summed gradient.
- We propose a novel attention mechanism with gradient and design two attention blocks based on SGA, *i.e.,* cross-SGA and self-SGA.
- Both quantitative and qualitative results verify that our method outperforms state-of-the-art modules on several image datasets.

2   Related Work

Reference-based Line-Art Colorization. The reference-based line-art colorization is a user-friendly approach to assist designers in painting the sketch with their desired color [31,28,55,2]. Early studies attempt to get the style latent code of reference and directly mix it with sketch feature maps to generate the color image [31,38]. To make better use of reference images, some studies propose spatial-adaptive normalization methods [36,60].

Different from the aforementioned methods that adopt latent vectors for style control, [28,55,54] learn dense semantic correspondences between sketch-reference pairs. These approaches utilize the dot-product attention [41,46] to model the semantic mapping between sketch-reference pairs and inject color into sketch correctly. Although traditional non-local attention is excellent in feature alignment and integration between different modalities, the model cannot learn robust representation due to the gradient conflict in attention's optimization. Thus, our work proposes the stop-gradient operation for attention to eliminate the gradient conflict problem in line-art colorization.

Attention Mechanism. The attention mechanism [49,41] is proposed to capture long-range dependencies and align signals from different sources. It is widely applied in vision [46,52], language [41,10], and graph [42] areas. Due to the quadratic memory complexity of standard dot-product attention, many researchers from the vision [6,7,29,57,13] and language [24,44,37] communities endeavor to reduce the memory consumption to linear complexity. Recently, vision transformer [11] starts a new era for modeling visual data through the attention mechanism. The booming researches using transformer substantially change the trend in image [39,30,45,48], point cloud [16,58], gauge [19], and video [34,1] processing.

Unlike existing works concerning the architectures of attention mechanism, we focus on the training of attention modules regarding its gradient flow. Although some strategies have been developed to improve the training efficiency [39,3] for vision transformer, they mainly modify the objective function to impose additional supervision. From another perspective, our work investigates the gradient issue in the attention mechanism.

Stop-Gradient Operation. Backpropagation is the foundation for training deep neural networks. Recently some researchers have paid attention to the gradient flow in the deep models. Hamburger [13] proposes the one-step gradient to tackle the gradient conditioning and gradient norm issues in the implicit global context module, which helps obtain stable learning and performance. SimSiam [4] adopts the one-side stop-gradient operation to implicitly introduce an extra set of variables to implement Expectation-Maximization (EM) like algorithm in contrastive learning. VQ-VAE [35] also encourages discrete codebook learning by the stop-gradient supervision. All of these works indicate the indispensability of the gradient manipulation, which demonstrates that the neural network performance is related to both the advanced architecture and the appropriate training strategy.

Inspired by prior arts, our work investigates the gradient conflict issue for training non-local attention. The stop-gradient operation clips the conflict gradient branches while preserving correction direction for model updates.

3   Proposed Method

3.1   Overall Workflow

As illustrated in Fig. 2, we adopt a self-supervised training process similar to [28]. Given a color image I, we first use XDoG [47] to convert it into a line-art image I_s. Then, the expected coloring result I_{gt} is obtained by adding a random color jittering on I . Additionally, we generate a style reference image I_r through applying the thin plate splines transformation on I_{gt}.

In the training process, utilizing I_r as the reference to color the sketch I_s, our model first uses encoder E_s and E_r to extract sketch feature $f_s \in \mathbb{R}^{c \times h \times w}$ and reference feature $f_r \in \mathbb{R}^{c \times h \times w}$. In order to leverage multi-level representation simultaneously for feature alignment and integration, we concatenate the feature maps of all convolution layers outputs after using 2D adaptive average pooling function to down-sample them into the same spatial size.

To integrate the content in sketch and the style in reference, we employ our SGA blocks. There are two types of SGA blocks in our module: cross-SGA integrates the features from different domains and self-SGA models the global context of input features. Then several residual blocks and a U-net decoder Dec with skip connections to sketch encoder E_s are adopted to generate the image I_{gen} by the mixed feature map f_{gen}. In the end, we add an adversarial loss [15] by using a discriminator D to distinguish the output I_{gen} and the ground truth I_{gt}.
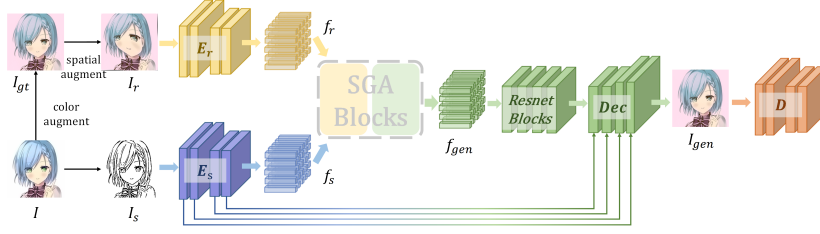
Fig. 2: The overview of our reference-based line-art colorization framework with a discriminator $\boldsymbol{D}$: Given the sketch I_s, the target image I_{gt} and the reference I_r obtained through original image I, we input I_s and I_r into encoder E_s and E_r to extract feature maps f_s and f_r. The SGA blocks, which contain cross-SGA and self-SGA, integrate f_s and f_r into the mixed feature map f_{gen}. Then f_{gen} is passed through several residual blocks and a U-net decoder Dec with skip connection to generate the image I_{gen}. The I_{gen} is supposed to be similar to I_{gt}.

3.2 Loss Function

Image Reconstruction Loss. According to the Section 3.1, both generated images I_{gen} and ground truth images I_{gt} should keep style consistency with reference I_r and outline preservation with sketch I_s. Thus, we adopt L_1 regularization to measure the difference between I_{gen} and I_{gt}, which ensures that the model colors correctly and distinctly:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{I_s, I_r, I_{gt}} \left[\| \boldsymbol{G}(I_s, I_r) - I_{gt} \|_1 \right] \tag{1}$$

where $\boldsymbol{G}(I_s, I_r)$ means coloring the sketch I_s with the reference I_r.

Adversarial Loss. In order to generate a realistic image with the same outline as the prior sketch I_s, we leverage a conditional discriminator $\boldsymbol{D}$ to distinguish the generated images from real ones [21]. The least square adversarial loss [33] for optimizing our GAN-based model is formulated as:

$$\mathcal{L}_{adv} = \mathbb{E}_{I_{gt}, I_s} \left[\| \boldsymbol{D}(I_{gt}, I_s) \|_2^2 \right] + \mathbb{E}_{I_s, I_r} \left[\| (1 - \boldsymbol{D}(\boldsymbol{G}(I_s, I_r), I_s)) \|_2^2 \right] \tag{2}$$

Style and Perceptual Loss. As shown in previous works [28,22], perceptual loss and style loss encourage a network to produce a perceptually plausible output. Leveraging the ImageNet pretrained network, we reduce the gaps in multi-layer activation outputs between the target image I_{gt} and generated image I_{gen} by minimizing the following losses:

$$\mathcal{L}_{perc} = \mathbb{E}_{I_{gt}, I_{gen}} \left[\sum_l \| \phi_l(I_{gt}) - \phi_l(I_{gen}) \|_1 \right] \tag{3}$$

$$\mathcal{L}_{\text{style}} = \mathbb{E}_{I_{gt}, I_{gen}} \left[\| \mathcal{G}(\phi_l(I_{gt})) - \mathcal{G}(\phi_l(I_{gen})) \|_1 \right] \tag{4}$$

where ϕ_l represents the activation map of the l_{th} layer extracted at the relu from VGG19 network, and $\mathcal{G}$ is the gram matrix.

Overall Loss   In summary, the overall loss function for the generator $\boldsymbol{G}$ and discriminator $\boldsymbol{D}$ is defined as:

$$\min_{\boldsymbol{G}} \max_{\boldsymbol{D}} \mathcal{L}_{\text{total}} = \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{\text{rec}} + \lambda_2 \mathcal{L}_{\text{perc}} + \lambda_3 \mathcal{L}_{\text{style}} \tag{5}$$

3.3   Gradient Issue in Attention

In this section, we use SCFT [28], a classic attention-based method in colorization, as an example to study the gradient issue in attention. $\boldsymbol{Q} \in \mathbb{R}^{n \times d}$ is the feature projection transformed by $\mathbf{W}_\text{q}$ from the input $\boldsymbol{X} \in \mathbb{R}^{n \times d}$. The feature projections $\boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{n \times d}$ from input $\boldsymbol{Y} \in \mathbb{R}^{n \times d}$ are transformed by $\mathbf{W}_\text{k}$ and $\mathbf{W}_\text{v}$. Given the attention map $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, the classic dot-product attention mechanism can be formulated as follows:

$$\boldsymbol{Z} = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}})\boldsymbol{V} + \boldsymbol{X} = \boldsymbol{A}\boldsymbol{V} + \boldsymbol{X} \tag{6}$$

(a) SCFT   (b) SCFT w/ stop-grad

Fig. 3: Stop-gradient in attention module. The $\boldsymbol{g}_{skip}$, $\boldsymbol{g}_Q$, $\boldsymbol{g}_K$ and $\boldsymbol{g}_V$ separately represent the gradient along their branches. The stop-gradient operation (**stop-grad**) truncates the backpropagation of conflict gradients existing in attention map calculation.

Previous works [5,13,3,39] present the training difficulty of vision attention: instability, worse generalization, *etc* . For line-art colorization, it is even more challenging to train the attention models, as the training involves GAN-style loss and reconstruction loss, which are understood to lead to mode collapse [15] or trivial solutions. Given a training schedule, the loss of colorization network can shake during training and finally deteriorate.
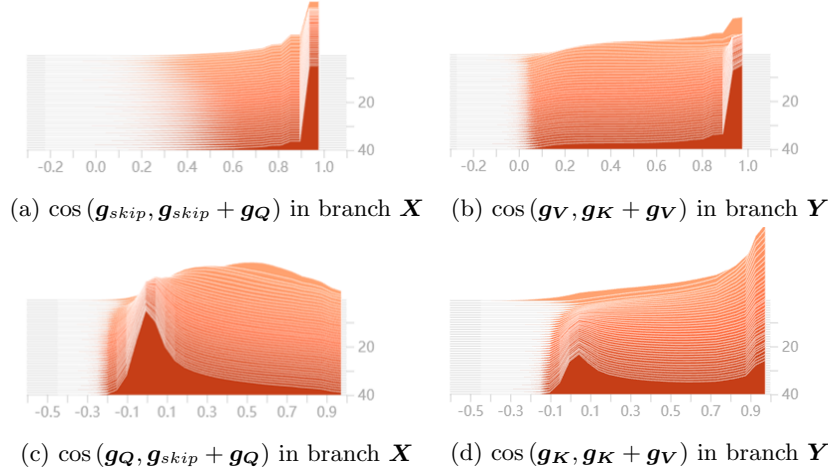
(a) $\cos{(\boldsymbol{g}_{skip}, \boldsymbol{g}_{skip} + \boldsymbol{g_Q})}$ in branch $\boldsymbol{X}$ (b) $\cos{(\boldsymbol{g_V}, \boldsymbol{g_K} + \boldsymbol{g_V})}$ in branch $\boldsymbol{Y}$

(c) $\cos{(\boldsymbol{g_Q}, \boldsymbol{g}_{skip} + \boldsymbol{g_Q})}$ in branch $\boldsymbol{X}$ (d) $\cos{(\boldsymbol{g_K}, \boldsymbol{g_K} + \boldsymbol{g_V})}$ in branch $\boldsymbol{Y}$

Fig. 4: The histograms of the gradient cosine value distribution in 40 epochs. A large cosine value means that the network mainly uses this branch of gradient to optimize the loss function.

To better understand reasons behind the training difficulty of attention in colorization, we analyze the gradient issue through the classic SCFT model [28]. We visualize the gradient flow back through the attention module in terms of each gradient branch and the summed gradient.

Fig. 4 offers the cosine value between different gradient branches and the total gradient. We separately calculate $\cos{(\boldsymbol{g}_{skip}, \boldsymbol{g}_{skip} + \boldsymbol{g_Q})}$ and $\cos(\boldsymbol{g_Q}, \boldsymbol{g}_{skip} + \boldsymbol{g_Q})$ for each pixel in branch $\boldsymbol{X}$ (means gradient in sketch feature maps f_s), $\cos{(\boldsymbol{g_V}, \boldsymbol{g_K} + \boldsymbol{g_V})}$ and $\cos{(\boldsymbol{g_K}, \boldsymbol{g_K} + \boldsymbol{g_V})}$ in branch $\boldsymbol{Y}$ (means gradient in reference feature maps f_r) to explore the gradient flow of the network during learning.

Note that first order optimization methods usually require the surrogate gradient $\tilde{\boldsymbol{g}}$ for update to be ascent, i.e., $\cos{(\tilde{\boldsymbol{g}}, \boldsymbol{g})} > 0$, where $\boldsymbol{g}$ is the exact gradient. Then the update direction based on the surrogate gradient can be descent direction. The visualization in Fig. 4 implies that the gradient $\boldsymbol{g}_{skip}$ from the skip connection for the branch $\boldsymbol{X}$ and the gradient $\boldsymbol{g_V}$ from $\boldsymbol{V}$ for the branch $\boldsymbol{Y}$ has already become an ascent direction for optimization, denoting that $\boldsymbol{g_Q}$ and $\boldsymbol{g_K}$ from the attention map construct the "conflict gradient" $\not{\boldsymbol{g}}$ in respect of the total gradient $\boldsymbol{g}$, i.e., $\cos{(\not{\boldsymbol{g}}, \boldsymbol{g})} < 0$.

Figs. 4a and 4b show that $\boldsymbol{g}_{skip}$ and $\boldsymbol{g_V}$ are usually highly correlated with the total gradient, where over **78.09%** and **52.39%** of the cosine values are greater than **0.935** in the 40th epoch, respectively. Moreover, these percentages increase during training, indicating the significance of the representative gradient. On the other hand, nearly **30.57%** of $\boldsymbol{g}_q$ in Fig. 4c and **10.77%** of $\boldsymbol{g_K}$ in Fig. 4d have negative cosine values in the 40th epoch. These proportions are **22.81%** and **5.32%** in the 20th epoch, respectively, gradually increasing during training.

The visualization regarding the gradient flows demonstrates that the two gradient branches compete with each other for a dominant position during training process, while $\boldsymbol{g}_{skip}$ and $\boldsymbol{g}_V$ construct an ascent direction and $\boldsymbol{g}_Q$ and $\boldsymbol{g}_K$ remain as the conflict gradient in respect of the total gradient in each branch. According to existing works in multi-task learning [50], large gradient conflict ratios may result in significant performance drop. It motivates us to detach the conflict gradient while preserving the dominant gradient as inexact gradient to approximate the original gradient, illustrated in Fig. 3.

Verified by Figs. 4a and 4b, the gradient after the stop-gradient operation forms an ascent direction of the loss landscape, *i.e.,* $\cos(\tilde{\boldsymbol{g}}, \boldsymbol{g}) > 0$, and thus be valid for optimization [14].

Table 1 shows that the gradient clipping through the stop-gradient operation can effectively improve the model performance. We can also remove $\mathbf{W}_k$ and $\mathbf{W}_q$ since there is no gradient propagating in them and they will not be updated in the training process. The lower FID and higher SSIM mean that model can generate more realistic images with higher outline preservation during colorization after the stop-gradient clipping.

Table 1: Test the Fréchet Inception Distance (FID) and SSIM with different settings of SCFT on anime dataset. ↑ means the higher the better, while ↓ indicates the lower the better.

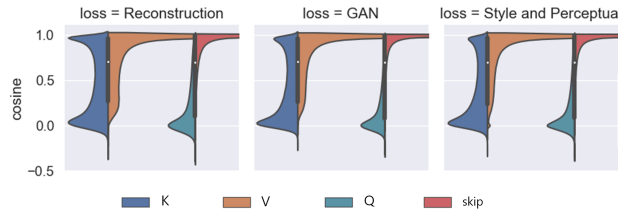| SCFT Setting		FID↓	SSIM↑
stop-grad	$\mathbf{W}_q\,\&\,\mathbf{W}_k$		
✗	✓	44.65	0.788
✗	✗	48.04	0.799
✓	✓	38.20	0.835
✓	✗	**36.78**	**0.841**

Fig. 5: Visualizations of the gradient cosine distribution when using a single loss on the pretrained SCFT model.

In order to investigate the reliability of gradient conflicts, we test the gradient cosine distributions when using a certain loss to confirm the trigger to gradient issue is the dot-product attention. We use the SCFT model to compute the

gradients cosine distribution of each loss to investigate whether loss functions or architectures cause the conflict. Fig. 5 shows that all loss terms cause similar conflicts, implying that the attention architecture leads to gradient conflicts.

3.4 Stop-Gradient Attention

Combining with the training strategy, we propose the **S**top-**G**radient **A**ttention (SGA). As Fig. 6a illustrates, in addition to the stop-gradient operation, we also design a new feature integration and normalization strategy for SGA. Treating stop-gradient attention map $\boldsymbol{A}$ as a prior deep graph structure input, inspired by [27,43], features can be effectively aggregated from adjacency nodes and the node itself:

$$\boldsymbol{Z} = \sigma(\boldsymbol{X}\mathbf{W}_{\mathrm{x}}) + \widehat{\boldsymbol{A}}\sigma(\boldsymbol{Y}\mathbf{W}_{\mathrm{y}}) \tag{7}$$

where σ is the leaky relu activate function and $\widehat{\boldsymbol{A}}$ is the attention map normalized
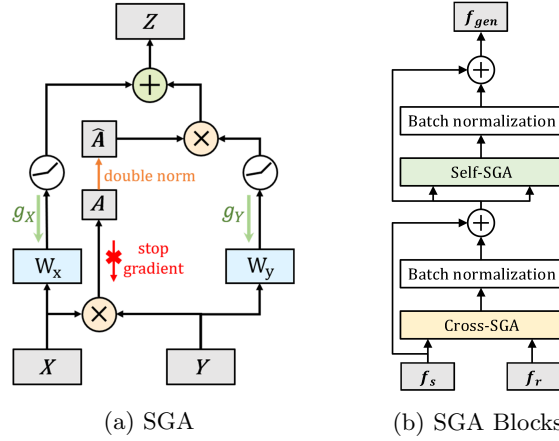

Fig. 6: SGA computes the attention map with stop-gradient, which truncates the gradient propagation of $\boldsymbol{g}_{att}$ and adopts a double normalization technique in addition. In our colorization network, we stack two types of SGA to integrate features: cross-SGA (yellow box) and self-SGA (green box).

by double normalization method analogous to Sinkhorn algorithm [9]. Different from softmax employed in classic non-local attention, the double normalization makes the attention map insensitive to the scale of input features [17]. The

normalized attention map $\hat{A}$ can be formulated as follows:

$$A = XY^{\top} \tag{8}$$

$$\tilde{A}_{ij} = \exp\left(A_{ij}\right) / \sum_{k} \exp\left(A_{ik}\right) \tag{9}$$

$$\widehat{A}_{ij} = \tilde{A}_{ij} / \sum_{k} \tilde{A}_{kj} \tag{10}$$

where $\widehat{A}_{ij}$ means correlation between i_{th} feature vector in X and j_{th} feature vector in Y. The pseudo-code of SGA is summarized in Algorithm 1.

Furthermore, we design two types of SGA, called cross-SGA and self-SGA. Both of their calculation are based on Algorithm 1. As shown in Fig. 6b, the only difference between them is whether the inputs are the same or not. Cross-SGA calculates pixel correlation between features from different image domains and integrates features under a stop-gradient attention map. Self-SGA models the global context and fine-tunes the integration. For stable training, we also adopt batch normalization layer and shortcut connections [18]. Combining above techniques, our SGA blocks integrate the sketch feature f_s and reference feature f_r into generated feature f_{gen} effectively.

Algorithm 1 SGA Pseudocode pytorch

```
# input:
# X: feature maps -> tensor(b, wh, c)
# Y: feature maps -> tensor(b, wh, c)

# output:
# Z: feature maps -> tensor(b, wh, c)

# other objects:
# Wx, Wy: embedding matrix -> nn.Linear(c,c)
# A: attention map -> tensor(b, wh, wh)
# leaky_relu: leaky relu activation function

with torch.no_grad():
    A = X.bmm(Y.permute(0, 2, 1))
    A = softmax(A, dim=-1)
    A = normalize(A, p=1, dim=-2)

X = leaky_relu(Wx(X))
Y = leaky_relu(Wy(Y))

Z = torch.bmm(A,Y) + X
```

---

4   Experiments

4.1   Experiment Setup

Dataset.   We test our method on popular anime portraits [40] and Animal FacesHQ (AFHQ) [8] dataset. The anime portraits dataset contains 33323 anime faces for training and 1000 for evaluation. AFHQ is a dataset of animal faces consisting of 15,000 high-quality images at 512×512 resolution, which contains three categories of pictures, *i.e.,* cat, dog, and wildlife. Each class in AFHQ provides 5000 images for training and 500 for evaluation. To simulate the line-art drawn by artists, we use XDoG [47] to extract sketch inputs and set the

parameters of XDoG algorithm with $\phi = 1 \times 10^9$ to keep a step transition at the border of sketch lines. We randomly set σ to be 0.3/0.4/0.5 to get different levels of line thickness, which generalizes the network on various line widths to avoid overfitting. And we set $p = 19, k = 4.5, \epsilon = 0.01$ by default in XDoG.

Implementation Details.   We implement our model with the size of input image fixed at 256×256 for each dataset. For training, we set the coefficients for each loss terms as follows: $\lambda_1 = 30, \lambda_2 = 0.01$, and $\lambda_3 = 50$. We use Adam solver [26] for optimization with $\beta_1 = 0.5, \beta_2 = 0.999$. The learning rate of generator and discriminator are initially set to 0.0001 and 0.0002, respectively. The training lasts 40 epochs on each dataset.

Evaluation Metrics.   In evaluation process, we randomly select reference images and sketch images for colorization as Fig. 7 shows. The popular Fréchet Inception Distance (FID) [20] is used to assess the perceptual quality of generated images by comparing the distance between distributions of generated and real images in a deep feature embedding. Besides measuring the perceptual credibility, we also adopt the structural similarity index measure (SSIM) to quantify the outline preservation during colorization, by calculating the SSIM between reference image and original color image of sketch.
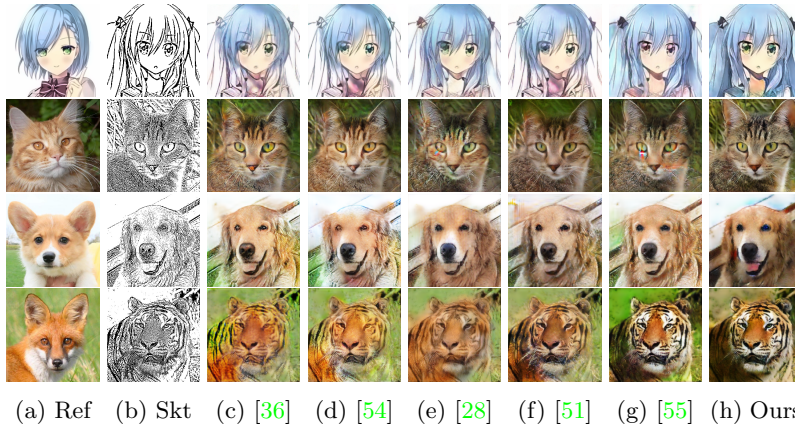

(a) Ref   (b) Skt   (c) [36]   (d) [54]   (e) [28]   (f) [51]   (g) [55]   (h) Ours

Fig. 7: Visualization of colorization results. "Ref" stands for "reference". "Skt" indicates "sketch". Compared with other methods, SGA shows correct correspondence between the sketch and reference images.

4.2   Comparison Results

We compare our method with existing state-of-the-art modules include not only reference-based line-art colorization [28] but also image-to-image translation,

i.e., SPADE [36], CoCosNet [54], UNITE [51] and CMFT [55]. For fairness, in our experiments, all networks use the same encoders, decoder, residual blocks and discriminator implemented in SCFT [28] with aforementioned train losses. Table 2 shows that SGA outperforms other techniques by a large margin. With respect to our main competitor SCFT, SGA improves by 27.21% and 25.67% on average for FID and SSIM, respectively. This clear-cut improvement means that SGA produces a more realistic image with high outline preservation compared with previous methods. According to Fig. 7, the images generated by SGA have less color-bleeding and higher color consistency in perceptual.

Table 2: Quantitative comparison with different methods. Boldface represents the best value. Underline stands for the second score.

Method	anime		cat		dog		wild	
	FID↓	SSIM↑	FID↓	SSIM↑	FID↓	SSIM↑	FID↓	SSIM↑
SPADE [36]	57.55	0.681	36.11	0.526	76.57	0.631	24.56	0.573
CoCosNet [54]	52.06	0.672	35.02	0.511	68.69	0.603	23.10	0.554
SCFT [28]	44.65	0.788	36.33	0.636	79.08	0.683	24.93	0.633
UNITE [51]	52.19	0.676	**33.26**	0.636	72.38	0.677	23.97	0.592
CMFT [55]	38.94	0.873	37.78	0.813	73.18	0.809	23.90	0.822
SGA	**29.65**	**0.912**	34.35	**0.843**	**54.76**	**0.841**	**15.19**	**0.831**

Furthermore, we explore the superiority of SGA over SCFT in terms of rescaling spectrum concentration of the representations. We compare the accumulative ratios of squared top r singular values over total squared singular values of the unfolded feature maps (*i.e.,* $\mathbb{R}^{C \times HW}$) before and after passing through the attention module, illustrated in Fig. 8. The sum of singular values is the nuclear norm, *i.e.,* the convex relaxation for matrix rank that measures how compact the representations are, which is widely applied in machine learning [23]. The accumulative ratios are obviously lifted after going through SCFT and SGA, which facilitates the model to focus more on critical global information [13]. However, our effective SGA can not only further denoise feature maps but also enforce the encoder before attention module to learn energy-concentrated rep-
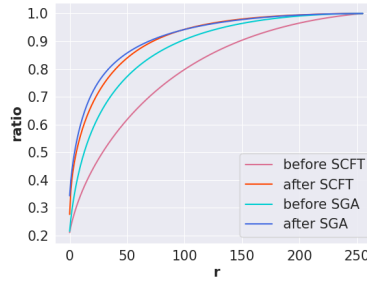

Fig. 8: Accumulative ratio of the squared top r singular values over total squared singular values in feature maps. The ratios of feature maps before and after the attention module in SCFT and SGA are displayed.

resentations, *i.e.,* under the effect of SGA, the CNN encoder can also learn to focus on the global information.

4.3   Ablation Study

We perform several ablation experiments to verify the effectiveness of SGA blocks in our framework, *i.e.,* stop-gradient operation, attention map normalization, and self-SGA. The quantitative results are reported in Table 3, showing the superiority of our SGA blocks.

Specifically, to evaluate the necessity of stop-gradient in non-local attention, we design a variant SGA without stop-gradient. In Table 3, it obtains inferior performance, which verifies the benefit of eliminating gradient conflict through stop-gradient.

Table 3: Ablation study result with different settings. Boldface represents the best value. Underline stands for the second best.

Setting	anime		cat		dog		wilds	
	FID↓	SSIM↑	FID↓	SSIM↑	FID↓	SSIM↑	FID↓	SSIM↑
SGA	**29.65**	0.912	34.35	**0.843**	**54.76**	**0.841**	**15.19**	**0.831**
SGA w/o stop-gradient	36.34	0.876	40.73	0.796	72.34	0.808	19.90	0.791
SGA w/o double-norm	33.42	0.861	34.42	0.811	55.08	0.828	15.95	0.809
SGA w/o self-SGA	31.56	**0.917**	**34.26**	0.842	55.69	0.839	16.36	0.821

Furthermore, we conduct an ablation study on the attention map normalization to validate the advantage of double normalization in our framework. Table 3 demonstrates that SGA with double normalization outperforms that with classic softmax function. Although classic softmax can generate realistic images, it suffers a low outline preservation, *i.e.,* the SSIM measure.

Based on the framework with stop-gradient and double normalization, we make an ablation study on the improvement of self-SGA additionally. Although our model has achieved excellent performance without self-SGA, there is still a clear-cut enhancement on most datasets after employing the self-SGA according to Table 3. The stacks of SGA can help model not only integrate feature effectively, but also fine-tune a better representation with global awareness for coloring.

Extending the training schedule to 200 epochs, Fig. 9 shows that SGA can still perform better with more epochs (29.71 in the 78th epoch) and collapse later than
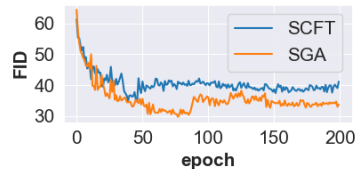

Fig. 9: Cat's FID during 200 epochs training.

SCFT [28], demonstrating the training stability for attention models in line-art colorization.

Additionally, to be more rigorous, we visualize the gradient distributions in the "SGA w/o stop-gradient". Fig. 10 implies the existing of gradient conflicts is a general phenomena in dot-product attention mechanism.
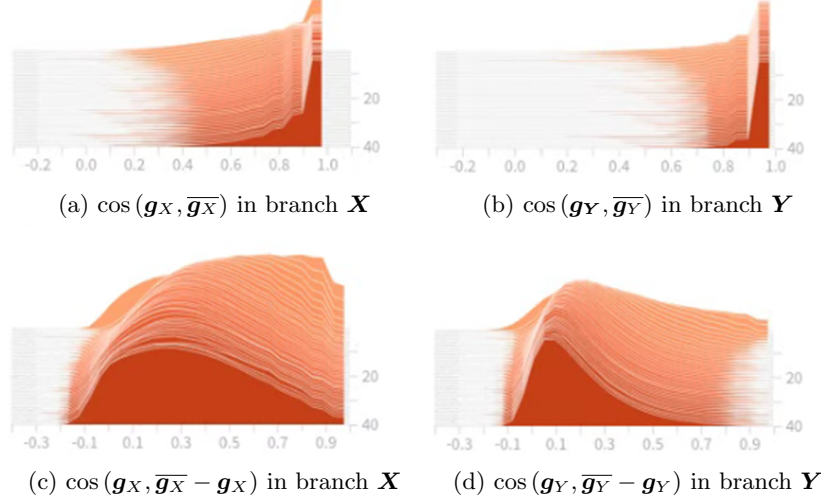

(a) $\cos\left(\boldsymbol{g}_X, \overline{\boldsymbol{g}_X}\right)$ in branch $\boldsymbol{X}$ (b) $\cos\left(\boldsymbol{g}_Y, \overline{\boldsymbol{g}_Y}\right)$ in branch $\boldsymbol{Y}$

(c) $\cos\left(\boldsymbol{g}_X, \overline{\boldsymbol{g}_X} - \boldsymbol{g}_X\right)$ in branch $\boldsymbol{X}$ (d) $\cos\left(\boldsymbol{g}_Y, \overline{\boldsymbol{g}_Y} - \boldsymbol{g}_Y\right)$ in branch $\boldsymbol{Y}$

Fig. 10: The gradient distribution of "SGA w/o stop-gradient". The $\boldsymbol{g}_X$ and $\boldsymbol{g}_Y$ are illustrated in Fig. 6a. The $\overline{\boldsymbol{g}_X}$ and $\overline{\boldsymbol{g}_Y}$ represent the total gradient, similar to the $\boldsymbol{g}_{skip} + \boldsymbol{g}_Q$ and $\boldsymbol{g}_K + \boldsymbol{g}_V$ in Fig. 4.

5   Conclusion

In this paper, we investigate the gradient conflict phenomenon in classic attention networks for line-art colorization. To eliminate the gradient conflict issue, we present a novel cross-modal attention mechanism, **S**top-**G**radient **A**ttention (**SGA**) by clipping the conflict gradient through the stop-gradient operation. The stop-gradient operation can unleash the potential of attention mechanism for reference-based line-art colorization. Extensive experiments on several image domains demonstrate that our simple technique significantly improves the reference-based colorization performance with better the training stability.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6836–6846 (October 2021) 3
2. Casey, E., Perez, V., Li, Z.: The animation transformer: Visual correspondence via segment matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11323–11332 (October 2021) 1, 3
3. Chen, X., Hsieh, C.J., Gong, B.: When vision transformers outperform resnets without pretraining or strong data augmentations. arXiv preprint arXiv:2106.01548 (2021) 4, 6
4. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15750–15758 (June 2021) 4
5. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9640–9649 (2021) 2, 6
6. Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J.: Aˆ2-nets: Double attention networks. In: Advances in Neural Information Processing Systems 31. pp. 352–361 (2018) 3
7. Chen, Y., Rohrbach, M., Yan, Z., Shuicheng, Y., Feng, J., Kalantidis, Y.: Graph-based global reasoning networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 3
8. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 10
9. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26**, 2292–2300 (2013) 9
10. Dai, Z., Yang, Z., Yang, Y., Carbonell, J.G., Le, Q., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. In: Annual Meeting of the Association for Computational Linguistics (ACL). pp. 2978–2988 (2019) 3
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021) 3
12. Dou, Z., Wang, N., Li, B., Wang, Z., Li, H., Liu, B.: Dual color space guided sketch colorization. IEEE Transactions on Image Processing **30**, 7292–7304 (2021) 1
13. Geng, Z., Guo, M.H., Chen, H., Li, X., Wei, K., Lin, Z.: Is attention better than matrix decomposition? In: International Conference on Learning Representations (2021) 2, 3, 4, 6, 12
14. Geng, Z., Zhang, X.Y., Bai, S., Wang, Y., Lin, Z.: On training implicit models. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021) 2, 8
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014) 4, 6
16. Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M.: Pct: Point cloud transformer. Computational Visual Media **7**(2), 187–199 (Apr 2021) 3
17. Guo, M.H., Liu, Z.N., Mu, T.J., Hu, S.M.: Beyond self-attention: External attention using two linear layers for visual tasks. arXiv preprint arXiv:2105.02358 (2021) 9

18. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV (2016) 10

19. He, L., Dong, Y., Wang, Y., Tao, D., Lin, Z.: Gauge equivariant transformer. Advances in Neural Information Processing Systems **34**, 27331–27343 (2021) 3

20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 11

21. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 5

22. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision (2016) 5

23. Kang, Z., Peng, C., Cheng, J., Cheng, Q.: Logdet rank minimization with application to subspace clustering. Computational intelligence and neuroscience **2015** (2015) 12

24. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: International Conference on Machine Learning. pp. 5156–5165. PMLR (2020) 3

25. Kim, H., Jhoo, H.Y., Park, E., Yoo, S.: Tag2pix: Line art colorization using text tag with secat and changing loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9056–9065 (2019) 1

26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015) 11

27. Kipf, T.N., Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks. In: Proceedings of the 5th International Conference on Learning Representations. ICLR '17 (2017) 9

28. Lee, J., Kim, E., Lee, Y., Kim, D., Chang, J., Choo, J.: Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 2, 3, 4, 5, 6, 7, 11, 12, 14

29. Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., Liu, H.: Expectation-maximization attention networks for semantic segmentation. In: International Conference on Computer Vision (2019) 3

30. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. International Conference on Computer Vision (ICCV) (2021) 3

31. LvMin Zhang, Y.J., Liu, C.: Style transfer for anime sketches with enhanced residual u-net and auxiliary classifier gan. In: Asian Conference on Pattern Recognition (ACPR) (2017) 2, 3

32. Maejima, A., Kubo, H., Funatomi, T., Yotsukura, T., Nakamura, S., Mukaigawa, Y.: Graph matching based anime colorization with multiple references. In: ACM SIGGRAPH 2019 (2019) 1

33. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017) 5

34. Neimark, D., Bar, O., Zohar, M., Asselmann, D.: Video transformer network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 3163–3172 (October 2021) 3

35. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: NIPS (2017) 4

36. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 2, 3, 11, 12

37. Roy, A., Saffar, M.T., Vaswani, A., Grangier, D.: Efficient content-based sparse attention with routing transformers. Transactions of the Association for Computational Linguistics **9**, 53–68 (2021) 3

38. Sun, T.H., Lai, C.H., Wong, S.K., Wang, Y.S.: Adversarial colorization of icons based on contour and color conditions. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 683–691 (2019) 2, 3

39. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021) 3, 4, 6

40. Tseng, H.Y., Fisher, M., Lu, J., Li, Y., Kim, V., Yang, M.H.: Modeling artistic workflows for image generation and editing. In: European Conference on Computer Vision. pp. 158–174. Springer (2020) 10

41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017) 2, 3

42. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: International Conference on Learning Representations (ICLR) (2018) 3

43. Wang, R., Yan, J., Yang, X.: Learning combinatorial embedding networks for deep graph matching. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3056–3065 (2019) 9

44. Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-attention with linear complexity. ArXiv **abs/2006.04768** (2020) 3

45. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: IEEE ICCV (2021) 3

46. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7794–7803 (2018) 2, 3

47. Winnemöller, H., Kyprianidis, J.E., Olsen, S.C.: Xdog: an extended difference-of-gaussians compendium including advanced image stylization. Computers & Graphics **36**(6), 740–753 (2012) 4, 10

48. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. arXiv preprint arXiv:2103.15808 (2021) 3

49. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning (ICML). pp. 2048–2057 (2015) 3

50. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. Advances in Neural Information Processing Systems **33**, 5824–5836 (2020) 2, 8

51. Zhan, F., Yu, Y., Cui, K., Zhang, G., Lu, S., Pan, J., Zhang, C., Ma, F., Xie, X., Miao, C.: Unbalanced feature transport for exemplar-based image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15028–15038 (June 2021) 11, 12

52. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. ICML (2019) 3

53. Zhang, L., Li, C., Simo-Serra, E., Ji, Y., Wong, T.T., Liu, C.: User-guided line art flat filling with split filling mechanism. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 1

54. Zhang, P., Zhang, B., Chen, D., Yuan, L., Wen, F.: Cross-domain correspondence learning for exemplar-based image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5143–5153 (2020) 2, 3, 11, 12

55. Zhang, Q., Wang, B., Wen, W., Li, H., Liu, J.: Line art correlation matching feature transfer network for automatic animation colorization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3872–3881 (January 2021) 1, 2, 3, 11, 12

56. Zhang, R.Y., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. ACM Transactions on Graphics **36**(4), 119 (2017) 1

57. Zhang, S., Yan, S., He, X.: LatentGNN: Learning efficient non-local relations for visual recognition. In: International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 97, pp. 7374–7383. PMLR (2019) 3

58. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16259–16268 (2021) 3

59. Zhou, X., Zhang, B., Zhang, T., Zhang, P., Bao, J., Chen, D., Zhang, Z., Wen, F.: Cocosnet v2: Full-resolution correspondence learning for image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11465–11475 (June 2021) 1

60. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: Image synthesis with semantic region-adaptive normalization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 3