Attention-Guided NIR Image Colorization via Adaptive Fusion of Semantic and Texture Clues

Xingxing Yang  
Jie Chen  
Department of Computer Science, Hong Kong Baptist University, Hong Kong

Zaifeng Yang  
Department of Electronics and Photonics, Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore

Zhenghua Chen  
Department of Machine Intelligence, Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore

Abstract
Near infrared (NIR) imaging has been widely applied in low-light imaging scenarios, such as remote sensing and night surveillance devices. However, it is difficult for human and algorithms to perceive the real scene in the colorless NIR domain. While Generative Adversarial Network (GAN) has been widely employed in various image colorization tasks, it is challenging for a direct mapping mechanism, such as a conventional GAN, to transform an image from the NIR to the RGB domain with correct semantic reasoning, well-preserved textures, and vivid color combinations concurrently. In this work, we propose a novel Attention-based NIR image colorization framework via Adaptive Fusion of Semantic and Texture clues (AAFSTNet), aiming at achieving these goals within the same framework. In AAFSTNet, the tasks of texture transfer and semantic reasoning are carried out in two separate network blocks. Specifically, the Texture Transfer Block (TTB) aims at extracting texture features from the NIR image’s Laplacian component and transferring them for subsequent color fusion. The Semantic Reasoning Block (SRB) extracts semantic clues and maps the NIR pixel values to the RGB domain. Finally, a Fusion Attention Block (FAB) is proposed to adaptively fuse the features from the two branches and generate an optimized colorization result.

In order to enhance the network’s learning capacity in semantic reasoning as well as mapping precision in texture transfer, we have proposed the Residual Coordinate Attention Block (RCAB), which incorporates coordinate attention into a residual learning framework, enabling the network to capture long-range dependencies along the channel direction and meanwhile precise positional information can be preserved along spatial directions. RCAB is also incorporated into FAB to facilitate accurate texture alignment during fusion. Both quantitative and qualitative evaluations show that the proposed method outperforms state-of-the-art NIR image colorization methods.

Keywords: Near-Infrared image colorization, Generative Adversarial Network, coordinate attention, texture transfer, adaptive fusion

1. Introduction
The near-infrared (NIR) spectrum (from 780 nm to 1000 nm) is adjacent to the visible spectrum (from 380 nm to 780 nm), which cannot be perceived by human eyes. NIR imaging
advantages traditional RGB imaging due to its high sensitivity to light. The intensity of images depends on the absorption and reflection of the materials in addition to the color itself, which benefits it in capturing details of scene in some poor light conditions such as remote sensing Hunt et al. (2010), night time video surveillance Govardhan and Pati (2014) and object detection Tu et al. (2019), while RGB cameras capture noisy and dark scene in the same condition.

Lack of color makes it difficult for people to perceive NIR images intuitively, and thus NIR image colorization draws great attentions with the rapid development of deep learning techniques in recent years. Although this issue shares some similarities with color transfer (e.g., Oliveira et al. (2015)) and grayscale image colorization (e.g., Zhang et al. (2016)), NIR image colorization is a much more challenging task. On the one hand, color transfer usually refers to transfer the color of a specific RGB image to another, which generates multi-channel output from multi-channel input (e.g., 3 channel in RGB). While the NIR image is monochromatic, the reduction of input image channels requires to distinguish the differences of each channels to produce reasonable results Limmer and Lensch (2016). On the other hand, image colorization, for example, grayscale image colorization, where the required intensity information is already given by the grayscale image input, thus the remaining problem is to estimate the chrominance in the RGB domain. While the NIR image colorization needs to estimate not only the intensity but also the chrominance because of the different spectral bands. Moreover, even for a same scene, the change of the light sources and the environment (such as thermal change) will produce NIR images with large variation in intensities. Accordingly, there are some notable cues for Image colorization and Color Transfer compared with NIR image colorization, including optimization, feature extraction and some prerequisites in common, but the mapping from the NIR domain to the RGB domain is more challenging for the reduction of input channels and difference of spectrum. As a result, improving the mapping capacity of the neural network is of vital importance in NIR image colorization.
Generally, NIR image colorization can be formulated as a bidirectional mapping problem, where both the color features in the RGB domain and texture information in the NIR domain should be maintained in the generated results. Unlike grayscale image colorization, which is a single-to-many mapping problem, NIR image colorization is a many-to-many mapping problem, which is a much more challenging and ambiguous process, as both the intensity and the chrominance information need to be estimated. On the one hand, the different intensity objects in the NIR domain may have the same chrominance counterpart in the RGB domain, which we name many-to-single mapping. On the other hand, the same intensity objects in the NIR domain may have different chrominance counterparts in the RGB domain, which we name single-to-many mapping, as illustrated in Fig. 1. If the mapping capacity of network is limited, it will fail to distinguish the color information in the RGB domain thus leads to incorrect color prediction of objects in the first case, as shown in Fig. 2. While mode collapse usually happens in the second case, as objects of the same intensity have counterparts in different colors, which leads the generator generate an "averaged-color" result to trick the discriminator, as shown in Fig. 2. As a result, the overall chrominance of generated images is slightly yellow. Therefore, improving the mapping capacity of generator is one of the most critical factors in generating a more vivid result.

This paper presents an attention-based NIR image colorization network via adaptive fusion of texture and semantic clues based on Generative Adversarial Network (GAN). A novel stage-fusion architecture is developed to learn both semantic and texture information
sufficiently. Specifically, it involves a texture transfer block (TTB) and a semantic reasoning block (SRB) to individually learn the high-frequency details and contextual information, followed by a fusion attention block (FAB) to fuse the two branches feature and learn the hybrid representation. The proposed method can effectively solve the “single-to-many” and “many-to-single” problems, as illustrated in Fig. 2. Both extensive quantitative experiments and visual results demonstrate significantly stronger mapping capacity compared with the state-of-the-art methods, and show the potential of this proposed framework in the other similar applications, like image fusion, low-light image enhancement. In summary, the contributions of this work are given as follows:

- We propose a two-branch fusion network based on GAN, with the two separately focus on transferring texture features from the NIR domain and semantic reasoning for colorization over two different networks. This mechanism proves to be efficient in generating semantically correct colorization results with fine textures preserved.

- We propose a fusion attention mechanism to adaptively fuse the texture/edge information with colorization outputs in a natural and semantically congruent manner.

- We have introduced the coordinate attention mechanism into NIR image colorization, which improves the model’s mapping capacity and helps the network to focus on important clues across channel and spatial dimensions.

The rest of the paper is organized as follows: Section 2 introduces some related works to NIR image coloration. Section 3 presents our proposed framework. In Section 4, both quantitative experiments and visual comparison are implemented and compared with the state-of-the-art methods, and ablation studies are conducted. Finally, Section 5 concludes this paper.

2. Related Work

2.1. Generative Synthesis in NIR Image Colorization

NIR image colorization can be viewed as a typical generative task, where Generative Adversarial network plays an important role in recent years. Since the mapping process is ambiguous and bidirectional (both texture information in the NIR domain and color information in the RGB domain are required to be estimated), many efforts are made on the generators. Mirza and Osindero (2014) proposed context-guided GAN (CGGAN) to concatenate a conditional code into the random input distribution in order to constraint the output of generator from random generations. In addition, in order to generate more stylized results, style transfer frameworks usually introduce style component into normalization layers, for instance, Conditional Batch Normalization Dumoulin et al. (2017), AdaIN Huang and Belongie (2017b) and SPADE Park et al. (2019). However, most of these methods are based on paired data where the NIR input has its corresponding RGB ground truth. However, such paired data are not easy to obtain due to the time-consuming manual registration of images from two different imaging equipment. To address this problem and make full use of the available abundant unpaired data, Mehri and Sappa (2019) utilized CycleGAN with a UNet-based generator to implement NIR image colorization for unpaired data.
2.2. Attention Mechanisms in Computer Vision Tasks

Attention mechanisms have demonstrated their superiority in a variety of computer vision tasks as they can help models to more precisely focus on the objects of interest, for instance, image segmentation [Huang et al. (2020), Hou et al. (2020), Fu et al. (2019)] and image classification [Hu et al. (2019a), Hu et al. (2019b), Bello et al. (2020), Woo et al. (2018)]. One of the most popular early attention mechanisms is SENet Hu et al. (2019b), which squeezes each 2D feature map into 1D weighted channel descriptors and then excites them into the same channel dimension with input to efficiently allocate inter-dependencies among channels. However, they ignored the spatial inter-dependencies of feature maps, which is also crucial for computer vision tasks. Subsequently, Woo et al. (2018) designed Convolutional Block Attention Module (CBAM) to introduce spatial attention mechanism via channel-wise global pooling and convolution layers with large-size kernels, while it sacrifices the channel inter-dependencies. It is worth noting that both texture information and color information are required to be excited in the final generated image for NIR image colorization, which means that it can benefit from both channel attention and spatial attention. Hou et al. (2021) proposed Coordinate Attention to split spatial information into two dimensions and then individually generates 2D channel and one-dimension spatial attention descriptor, and finally multiples the two descriptors together to generate full channel-spatial attention maps.

2.3. Grayscale Image Colorization and Color Transfer

Grayscale image colorization and color transfer are more intuitive and straightforward than NIR image colorization. On the one hand, grayscale image is a linear transfer of its corresponding RGB image, which means the intensity is already given by the grayscale input, so it only requires to estimate the chrominance. For instance, Iizuka et al. (2016) presented an end-to-end framework for automatic coloring of grayscale images, which combines both global priors and local image features. Based on Convolutional Neural Network (CNN), the proposed deep network has a fusion layer, which can effectively combine the local information obtained from each patch with the global information obtained from the whole image. However, this method cannot be directly applied in NIR image colorization because the intensity prior is absent in the latter case. On the other hand, the key point of color transfer is to analyze the color distribution of an input image and then transfer the color of reference image to a target color distribution. For example, Shih et al. (2013) transferred color from videos with a similar scene as the input image, which achieves impressive results in mapping the daytime images into nighttime style. Unfortunately, this framework cannot be applied for NIR image colorization either, because NIR image colorization is a single-channel-to-multi-channel mapping process instead of identical channels.

2.4. NIR Image Colorization

Different from grayscale image colorization and color transfer, NIR image colorization is a more ambiguous problem and need more cues to map the NIR domain into the RGB domain. Suárez et al. (2018) proposed to utilize DCGAN to implement NIR image colorization, which demonstrates that GAN is are competent in this task. However, the architecture is too rough to learn the high-level features in both the NIR domain and the RGB domain, resulting in some blurs and color distortions in the generated results. Suárez et al. (2017)
proposed a triplet framework based on DCGAN to learn the three-channel information from RGB images separately. The color information of objects is well predicted. However, it still fails to maintain the abundant texture information in the NIR domain. Note that these methods mentioned above are strictly based on paired data. Mehri and Sappa (2019) utilized CycleGAN based on unpaired data, and Yang and Chen (2020) further proposed an alternatively trained CycleGAN by introducing cross-scale skip connections into the decoder of the UNet-based generator, which can generate more vivid colorized results, but still fails to generate a result that is "more loyal" to the ground truth and lose some texture information.

We claim that all these methods simply map the NIR domain to the RGB domain will lose some texture information of the NIR domain inevitably. It is intuitive because the objective function will force the generated images to be as close to the ground truth as possible, as the abundant texture information in NIR input is absent in the RGB ground truth, which inspires us to solve this problem in two directions: one is to map the texture information, the other is to map the color information.

Figure 3: (a) Illustration of the architecture of our Proposed AAFSTNet for NIR image colorization. (b) shows the detail for the Residual Coordinate Attention Block (RCAB), and (c) shows the detail for the Semantic Reasoning Block (SRB).
3. Proposed Approach

We propose a novel Attention-based NIR image colorization framework via Adaptive Fusion of Semantic and Texture clues (AAFSTNet). AAFSTNet focuses on improving the mapping capacity to learn both the high-level color features in the RGB domain and low-level texture features in the NIR domain for correct semantic reasoning, well-preserved textures, and vivid color combinations concurrently. The system diagram of AAFSTNet is illustrated in Fig. 3, unlike the widely used single branch framework which directly transfers images from the NIR to the RGB domain, we propose a novel two-branch parallel feature extraction module to learn the semantic features in the RGB domain and texture information in the NIR domain via the semantic reasoning block (SRB) and the texture transfer block (TTB), respectively. It is observed that directly transferring NIR images to the RGB domain will inevitably lose texture information in the NIR domain and thus leads to a certain degree of ambiguity, as shown in Fig. 5. Finally, we further propose a fusion attention block (FAB) to fuse the texture information and semantic information and refine the feature maps to generate an optimal result.

3.1. Two-branch Parallel Feature Extraction Module

As shown in Fig. 3 (c), the two-branch parallel feature extraction module consists of texture transfer block (TTB) and semantic reasoning block (SRB). The SRB consists of an encoder-decoder architecture based on UNet with embedded attention mechanism in each encoder and decoder layer. For each encoder, there is a down-sampling operation using a convolution layer with the kernel size of $4 \times 4$ and stride size of 2. After that, an attention mechanism is embedded in the first three layers of encoder and the last three layers of decoder to tell the network "what" and "where" to attend, called residual coordinate attention block (RCAB), which will be introduced in Section 3.2, followed by an instance normalization Huang and Belongie (2017a) (except the first encoder layer) and a leaky ReLU (0.2) activation function. It is worth noting that the attention mechanism becomes less and less important as the network deepens because of the reduction of spatial dimension. In the innermost layer, the dimension of the feature map is $512 \times 1 \times 1$, which can be viewed as a channel descriptor without spatial information. Therefore, we only embed RCAB in the shallow layers, which is also more computationally efficient. Here we adopt instance normalization rather than batch normalization because the former is more suitable for style transfer. For each decoder, there is an up-sampling operation using de-convolution layer with the kernel size of $4 \times 4$ and stride size of 2. The rest processes are as the same as encoder layers. Additionally, we design cross-scale connections in decoder layers to further improve the learning capacity of UNet, which helps to restore more high-frequency information from deep layers and contributes to better color prediction. To be specific, we use bilinear interpolation to upscale the deep decoder layers to the shallow decoder layers.

The TTB is designed to extract texture features of the input NIR images and then will be fused with semantic features, as shown in Fig. 3. As mentioned in Section 1, directly mapping NIR image into the RGB domain will lose texture information of the NIR domain, as the objective function will force the generator to generate an image that is closest to the RGB ground truth, which lacks the abundant texture information compared with the corresponding NIR image. Inspired by Paris et al. (2015), we utilize a Laplacian filter to extract the texture information of NIR image and then fuse it into the feature...
maps extracted by SRB to generate a more vivid result. Generally, the Laplacian filter has been widely used in edge detection as it can highlight regions of rapid intensity change. Specifically, our TTB leverages a $5 \times 5$ fixed-weights Gaussian smoothing filter to reduce its sensitivity to noise, and then the Laplacian component can be generated by using the NIR input to subtract the blurred image. Finally, a $1 \times 1$ convolution layer is adopted to expand the dimension as the same as the feature maps generated by SRB. The whole process can be formulated as follows:

$$L = I - G(I),$$

where $I$ represents the NIR input, and $G(\cdot)$ represents the Gaussian smoothing filter.

### 3.2. Fusion Attention Block

Considering the superiority of attention mechanisms in image classification and segmentation tasks, it motivates us to propose a novel attention mechanism, called residual coordinate attention block (RCAB), which can highlight both channel interest and spatial interest to improve the mapping capacity of our model for NIR image colorization. Coordinate attention was first used in Hou et al. (2021) for efficient mobile network design, which proves to be most effective in object detection. In our context, we face the similar challenge of highlighting objects of interest; therefore, we propose a novel RCAB, which relies on the CAB module to predict the residual of high-level information to further improve the learning capacity of semantic reasoning. Meanwhile, precisely highlighting the semantic features will benefit our network in more accurately fusing the texture features from TTB with the semantic features from SRB. As the color characteristics mainly depend on the relationship between different channels, which will be degraded by the popular Batch Normalization operation, so we adopt Instance Normalization to normalize the feature map. As shown in Fig. 3 (b), unlike the traditional SENet Hu et al. (2019b) that transforms the feature maps into 1D channel descriptor using 2D spatial global pooling and CBAM Woo et al. (2018) that transforms the feature maps into 2D spatial descriptor via 1D channel-wise global pooling, our RCAB factorizes the feature maps along two dimensions of spatial domain via x-wise and y-wise average pooling. Thus both channel descriptor and spatial descriptor can be preserved along these two directions. Finally, the two descriptors are multiplied to generate a full coordinate attention, where long-range dependencies of both channel and spatial domain can be encoded to increase the representational capacity of objects of interest. Therefore, we embed our RCAB in the FAB and SRB. The RCAB is embedded in the first three encoder layers and the last three decoder layers in the UNet of SRB, because we find that as the UNet network deepens, the role of RCAB becomes less important due to the reduction of spatial dimensions.

Instead of fusing the features maps directly, we propose an additional fusion attention block (FAB) to adaptively fuse the extracted texture information and semantic information and refine to generate a delicate result, as shown in Fig. 3. It consists of a $3 \times 3$ convolution layer with stride size of 1 and a leaky ReLU (0.2), followed by a RCAB and another $3 \times 3$ convolution layer with stride size of 1 to decrease the dimension into 3 channels. Finally, a sigmoid function is adopted to normalize the result. The FAB can not only benefit the network in maintaining more abundant texture information but also refine it to generate more natural colorful results, as demonstrated in Fig. 5.
3.3. Objective Functions

The generator of the GAN model is illustrated in Fig. 3, while for the discriminator, conventional PatchGAN is adopted. The detailed losses will be described in this subsection.

3.3.1. GAN Loss

The GAN loss can be expressed as:

\[ \mathcal{L}_{GAN}(G,D,x,y) = \mathbb{E}_{y \sim p_{data}}[\log D(y)] + \mathbb{E}_{x \sim p_{data}}[\log(1 - D(G(x)))] \quad (2) \]

Here \( \mathbb{E}_{x \sim p_{data}}(x) \) and \( \mathbb{E}_{y \sim p_{data}}(y) \) are the distributions of NIR images and generated fake RGB images, respectively, and \( \log(1 - D(G(x))) \) and \( \log D(y) \) are the probabilities of the discriminators for fake and real data.

3.3.2. Pair Consistent Loss

As we only use paired NIR and RGB data to train our model, for the pair consistent loss on the paired data, they can be calculated as:

\[ \mathcal{L}_{pair}(G,x,y) = \mathbb{E}_{x \sim p_{paired\ data}}||(G(x)) - y||_1 \quad (3) \]

Here \( \mathcal{L}_{pair}(G,x,y) \) is the pair consistent loss for the generated fake RGB images and the ground truth.

3.3.3. Total Loss

The total losses for our model can be expressed as follows:

\[ \mathcal{L} = \mathcal{L}_{GAN} + \lambda \mathcal{L}_{pair} \quad (4) \]

where \( \lambda \) indicates the contribution of the pair consistent loss to the whole objective function.

3.4. Dataset and Data Augmentation

The VCIP 2020 Grand Challenge dataset provides pixel-aligned NIR-RGB image pairs and 1020 unpaired RGB images with similar scene categories. Only 372 NIR images are available in the paired data with ground truth, while 1392 RGB images are available in both paired and unpaired data. Compared with the numbers of RGB images, the number of the NIR images is far from enough. Moreover, as we can observed from the NIR images, the pixel values of a same object such as tree, sky and land can be varied in a wide range, which makes it more difficult for the network to learn the difference between objects. Furthermore, it can also be observed that some objects such as trees and sky locate at the edges and only half or less than half of the objects can be shown in the photos, which makes it challenging for the network to learn the semantic correctly.

In our training phase, in order to make full use of the available data and address the problems mentioned above, the 372 NIR and their corresponding paired RGB images will be augmented in terms of scaling, mirroring, random size cropping and contrast adjustment:

- **Contrast adjustment**: we randomly adjust the contrast from 0.5 to 1.5 times of the original images of 80% of the paired data.
Table 1: Quantitative comparison among different methods on the VCIP 2020 Grand Challenge Dataset. Higher PSNR and SSIM values and lower AE values indicate better performance. The best and second best results are highlighted in red and blue, respectively.

| Method        | PSNR↑ | SSIM↑ | AE↓  |
|---------------|-------|-------|------|
| ATCycleGAN    | 19.59 | 0.59  | 4.33 |
| NIR-GNN       | 17.50 | 0.60  | 5.22 |
| MFF           | 17.39 | 0.61  | 4.69 |
| SST           | 14.26 | 0.57  | 5.61 |
| AAFSTNet (Ours) | 20.40 | 0.62  | 3.91 |

- **Cropping and scaling**: After contrast adjustment, we adopt bicubic interpolation to scale the original NIR images, which will be randomly cropped to a size of 200 × 200 and resized to 256 × 256.

- **Mirroring**: After contrast adjustment and size scaling, half of the MIR images will be mirrored. Note that 20% of the input NIR images will not be involved in contrast adjustment, cropping and scaling but mirroring.

4. Experimental Results

In this section, we first introduce the implementation details of our model, and then three evaluation metrics will be illustrated, including PSNR, SSIM and AE. Subsequently, we will show both quantitative and visual results compared with other state-of-the-art methods. Meanwhile, in order to further demonstrate the strong learning capacity of our method, it compares with one state-of-the-art method Yang and Chen (2020) on EPFL dataset\(^1\), which ranks the first place of quantitative results in VCIP 2020 Grand Challenges on NIR Image Colorization\(^2\). Finally, we will introduce our ablation study details.

4.1. Implementation Details

We have trained our proposed AFSTNet using a PC with Intel(R) Core(TM) i7-10700 CPU (2.90GHz) and one NVIDIA GeForce RTX 2070 SUPER GPU under Windows10 operation system. All the images have the same size of 256 × 256 and normalized to the range of (0, 1). For the data provided by the Challenge of VCIP 2020, 372 paired NIR-RGB images images are provided. In Eq. 4, the parameter \(\lambda = 100\). In terms of training phase, it takes 50s for one epoch in average. The learning rate is set to \(1 \times 10^{-3}\), and will decay to zero from 550 epoch gradually.

4.2. Evaluation Criteria

In terms of evaluation metrics, we use Peak of Noise-to-Signal Ratio (PSNR), Structural Similarity (SSIM) Wang et al. (2004), and Angular Error (AE) to evaluate the performance.

---

1. Dataset URL: [https://www.epfl.ch/labs/ivrl/research/downloads/rgb-nir-scene-dataset/](https://www.epfl.ch/labs/ivrl/research/downloads/rgb-nir-scene-dataset/)
2. The website of this challenge: [http://www.vcip2020.org/grand_challenge.htm](http://www.vcip2020.org/grand_challenge.htm)
of colorization. PSNR value is calculated as:

$$\text{PSNR} = 10 \cdot \log_{10} \left( \frac{255^2}{\text{MSE}(I_{\text{out}}, I_{\text{gt}})} \right),$$  \hspace{1cm} (5)

where $I_{\text{out}}$ and $I_{\text{gt}}$ represent the RGB colorization result and the RGB ground truth, respectively. MSE indicate mean squared error. To provide a color similarity measure closer to human color perception, we also calculate the SSIM and the Angular Error (AE). AE is defined as:

$$\text{AE} = \cos^{-1} \left[ \frac{I_{\text{out}} \odot I_{\text{gt}}}{||I_{\text{out}}||_2 \cdot ||I_{\text{gt}}||_2} \right],$$  \hspace{1cm} (6)

where $|| \cdot ||_2$ indicates $L_2$ norm, and $\odot$ indicates element-wise multiplication between vectorized images.

### 4.3. Results and Analysis

In this section, we first comprehensively evaluate our module and compare with four state-of-the-art methods for NIR image colorization. Ablation studies are subsequently carried out to individually evaluate the contributions of each sub module. To further elaborate the learning capacity of our method, we compare with ATCycle, which ranked first in quantitative comparison of VCIP 2020 Grand challenge on NIR Image Colorization, on EPFL dataset, whose scenario is more complicated than the VCIP 2020 Grand Challenge dataset.

We first conduct quantitative comparison between our method and the results generated by the other four state-of-the-art methods: ATCycleGAN [Yang and Chen (2020)], NIR-GNN [Valsesia et al. (2020)], MFF [Yan et al. (2020)] and SST [Wang et al. (2020)], in terms of PSNR, SSIM and AE on VCIP 2020 Grand Challenge dataset. We have retrained ATCycleGAN with only paired data. In addition, as the MMF and SST have used additional online-RGB images without corresponding NIR images provided by the VCIP challenge for training, while we haven’t used those additional data, however as validated by our experiments, our module can achieve even more vivid colorization, which further demonstrates the effectiveness of our model. As shown in Table. 1, obviously, our method outperforms all these state-of-the-art methods in terms of PSNR, SSIM and AE. Higher PSNR and SSIM values indicate more vital features including more abundant texture and more vivid color information, are retained in the generated results. In addition, the smaller value of AE indicates that the more consistent color information of objects in the generated results.

We randomly select 5 images in the testing dataset of VCIP 2020 Grand Challenge dataset for visual comparison. As illustrated in Fig. 4, generally, all methods can generate colorized images. However, the images generated by SST and NIR-GNN show missing textures and color distortion. while the results generated by ATCycleGAN have large area of blur and also have color distortion to certain extent. The results generated by MFF are better, however, still fail to map the color information in RGB domain correctly, for example, the mountains in the third row and last row. In general, the results generated by our method are closer to the RGB ground truth, texture information from the NIR domain have been well transferred to the RGB domain, and the color information is vivid and semantically correct.
Figure 4: Visual comparison among different methods on VCIP 2020 Grand Challenge Dataset. Pictures from left to right correspond to the NIR input images, and the generated results of SST Wang et al. (2020), NIR-GNN Valsesia et al. (2020), ATCycleGAN Yang and Chen (2020), MFF Yan et al. (2020), our method (in red) and RGB ground truth.

Ablation Study on Sub-Modules. To illustrate the performance of our proposed mechanism, including texture transfer block (TTB), fusion attention block (FAB) and residual coordinate attention block (RCAB), we perform a series of ablation experiments, the quantitative comparison are listed in Table. 2, and the corresponding visual results are shown in Fig. 5. Note that all network variations under ablation experiments are trained from scratch, with all network settings identical except for the current target module under evaluation. As can be seen, the FAB contributes the most to the final quantitative results. The main contribution of TTB is in improving the fine texture details. The RCAB can significantly make the outcome more colorful. Especially, as shown in Fig. 5, without any of these three modules, the output images show obvious color mapping errors as shown in the first and third row. As shown in the second row of this figure, without TTB, the RGB images will lose some texture information which leads to blur. Although the difference for quantitative results is not significant when the RCAB is absent, obvious degradation in the qualitative results can be seen, where the color of some salient objects are darkened (e.g., the tree in red boxes of last row in Fig. 5.), and unpleasant artifacts and blur occur (e.g., the
NIR Image Colorization Via Adaptive Fusion

Figure 5: Visual comparison of different pipelines with the full AAFSTNet framework. The main differences have been highlighted in red boxes.

Table 2: Comparison of different pipelines with the original framework. The best results are highlighted in red, and second best results are highlighted in blue.

| Settings       | PSNR↑ | SSIM↑ | AE↓ |
|----------------|-------|-------|-----|
| full AAFSTNet   | 20.40 | 0.62  | 3.91|
| w/o RCAB       | 20.07 | 0.60  | 3.88|
| w/o TTB        | 19.83 | 0.60  | 4.06|
| w/o FAB        | 18.88 | 0.59  | 4.18|

sky in the third row in Fig. 5.), which indicates that the RCAB helps to boost the learning capacity in semantic reasoning as well as mapping precision during texture transfer.

To further illustrate the superior learning capacity of our method, we compare with ATCycleGAN on the EPFL dataset, which contains much more complex semantic elements than the VCIP 2020 dataset. The dataset includes 380 paired data with buildings, street and urban scenarios. As shown in Table. 3, the quantitative differences between the two methods are even larger than those evaluations on the VCIP 2020 dataset, which indicates that our model has strong learning capacity and can be applied to more complicated scenarios. The visual comparison is shown in Fig. 6. As can be seen, the results generated by our method are much clearer with vivid color information.
Figure 6: Visual comparison between our proposed AAFSTNet with ATCycleGAN on the EPFL dataset. The main differences are high in red boxes.

Table 3: Comparison with ATCycleGAN on the EPFL dataset. The best results are highlighted in red.

| Settings        | PSNR↑ | SSIM↑ | AE↓  |
|-----------------|-------|-------|------|
| ATCycleGAN      | 16.89 | 0.46  | 5.61 |
| AAFSTNet (ours) | 17.37 | 0.55  | 5.30 |

5. Conclusion

In this paper, we have proposed a novel NIR image colorization framework to adaptively fuse the texture information transferred from the NIR domain and semantic information in the RGB domain based on GAN. To improve the learning capacity of the generative model, we have proposed a residual coordinate attention mechanism to accurately highlight and target both channel and spatial clues. The results indicate that the proposed method outperforms state-of-the-art methods, achieving 20.40 dB, 0.62 and 3.91 in evaluation metrics of PSNR, SSIM and AE, respectively, over the VCIP 2020 Grand Challenge dataset.
However, our model shows limitations when applied to colorization scenarios without clear semantic context (e.g., failure to vividly generate artificial color over urban targets). We will continue to work on this limitation as our future work.

References

Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks, 2020.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style, 2017.

Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation, 2019.

P Govardhan and Umesh C. Pati. NIR image based pedestrian detection in night vision with cascade classification and validation. In IEEE International Conference on Advanced Communications, Control and Computing Technologies, pages 1435–1438, 2014.

Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In IEEE Conference on Computer Vision and Pattern Recognition, pages 4002–4011, 2020.

Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design, 2021.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks, 2019a.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019b.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In IEEE International Conference on Computer Vision, pages 1510–1519, 2017a.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization, 2017b.

Zilong Huang, Xinggang Wang, Yunchao Wei, Lichao Huang, Humphrey Shi, Wenyu Liu, and Thomas S. Huang. Ccnet: Criss-cross attention for semantic segmentation, 2020.

E. Raymond Hunt, W. Dean Hively, Stephen J. Fujikawa, David S. Linden, Craig S. T. Daughtry, and Greg W. McCarty. Acquisition of nir-green-blue digital photographs from unmanned aircraft for crop monitoring. Remote Sensing, 2(1):290–305, 2010.

S. Iizuka, Edgar Simo-Serra, and H. Ishikawa. Let there be color! ACM Transactions on Graphics, 35:1–11, 2016.

Matthias Limmer and Hendrik P. A. Lensch. Infrared colorization using deep convolutional neural networks, 2016.
Armin Mehri and Angel D. Sappa. Colorizing near infrared images through a cyclic adversarial approach of unpaired samples. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 971–979, 2019.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

Miguel Oliveira, Angel Domingo Sappa, and Vitor Santos. A probabilistic approach for color correction in image mosaicking applications. *IEEE Transactions on Image Processing*, 24(2):508–523, 2015.

Sylvain Paris, Samuel W. Hasinoff, and Jan Kautz. Local laplacian filters: Edge-aware image processing with a laplacian pyramid. *Commun. ACM*, 58(3):81–91, February 2015.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization, 2019.

Yichang Shih, Sylvain Paris, Frédéric Durand, and William T. Freeman. Data-driven hallucination of different times of day from a single outdoor photo. *ACM Transactions on Graphics*, 32(6), November 2013.

Patricia L. Suárez, Angel D. Sappa, and Boris X. Vintimilla. Learning to colorize infrared images. In Fernando De la Prieta, Zita Vale, Luis Antunes, Tiago Pinto, Andrew T. Campbell, Vicente Julián, Antonio J.R. Neves, and María N. Moreno, editors, *Trends in Cyber-Physical Multi-Agent Systems. The PAAM Collection - 15th International Conference*, pages 164–172, Cham, 2018. Springer International Publishing.

Patricia L. Suárez, Angel D. Sappa, and Boris X. Vintimilla. Infrared image colorization based on a triplet DCGAN architecture. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 212–217, 2017.

Zhengzheng Tu, Tian Xia, Chenglong Li, Yijuan Lu, and Jin Tang. M3S-NIR: Multi-modal multi-scale noise-insensitive ranking for RGB-T saliency detection. In *IEEE Conference on Multimedia Information Processing and Retrieval*, pages 141–146, 2019.

Diego Valsesia, Giulia Fracastoro, and Enrico Magli. NIR image colorization with graph-convolutional neural networks. In *IEEE International Conference on Visual Communications and Image Processing*, pages 451–454, 2020.

Fengqiao Wang, Lu Liu, and Cheolkon Jung. Deep near infrared colorization with semantic segmentation and transfer learning. In *IEEE International Conference on Visual Communications and Image Processing*, pages 455–458, 2020.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module, 2018.

Longbin Yan, Xiuheng Wang, Min Zhao, Shumin Liu, and Jie Chen. A multi-model fusion framework for NIR-to-RGB translation. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 459–462, 2020.
Zaifeng Yang and Zhenghua Chen. Learning from paired and unpaired data: Alternately trained CycleGAN for near infrared image colorization. In *IEEE International Conference on Visual Communications and Image Processing*, pages 467–470, 2020.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization, 2016.