Learn to Recover Visible Color for Video Surveillance in a Day

Guangming Wu[1], Yinqiang Zheng[2⋆], Zhiling Guo[1], Zekun Cai[1], Xiaodan Shi[1], Xin Ding[3,4], Yifei Huang[1], Yimin Guo[1], and Ryosuke Shibasaki[1]

[1] The University of Tokyo, Tokyo 113-8654, Japan
{huster-wgm,guozhilingcc,caizekun,shixiaodan,guo.ym,shiba}@csis.u-tokyo.ac.jp,
hyf@iis.u-tokyo.ac.jp
[2] National Institute of Informatics, Tokyo 101-8430, Japan
yqzheng@nii.ac.jp
[3] Wuhan University, Hubei 430072, China
[4] Peng Cheng Laboratory, Shenzhen 518055, China
ding-xin@whu.edu.cn

Abstract. In silicon sensors, the interference between visible and near-infrared (NIR) signals is a crucial problem. For all-day video surveillance, commercial camera systems usually adopt NIR cut filter, and auxiliary NIR LED illumination to selectively block or enhance NIR signal according to the surrounding light conditions. This switching between the daytime and the nighttime mode inevitably involves mechanical parts, and thus requires frequent maintenance. Furthermore, images captured at nighttime mode are in shortage of chrominance, which might hinder human interpretation and high-level computer vision algorithms in succession. In this paper, we present a deep learning based approach that directly generates human-friendly, visible color for video surveillance in a day. To enable training, we capture well-aligned video pairs through a customized optical device and contribute a large-scale dataset, video surveillance in a day (VSIAD). We propose a novel multi-task deep network with state synchronization modules to better utilize texture and chrominance information. Our trained model generates high-quality visible color images and achieves state-of-the-art performance on multiple metrics as well as subjective judgment.

Keywords: Video Surveillance in a Day, Color Recovery, State Synchronization Network

1 Introduction

In recent years, surveillance cameras have been widely used for security and scientific purposes. Most commercial surveillance cameras are based on silicon sensors, usually equipped with an RGB color filter array, which are sensitive to light with a wavelength from about 400 nm to 1000 nm [37], covering both the

⋆ Corresponding author.

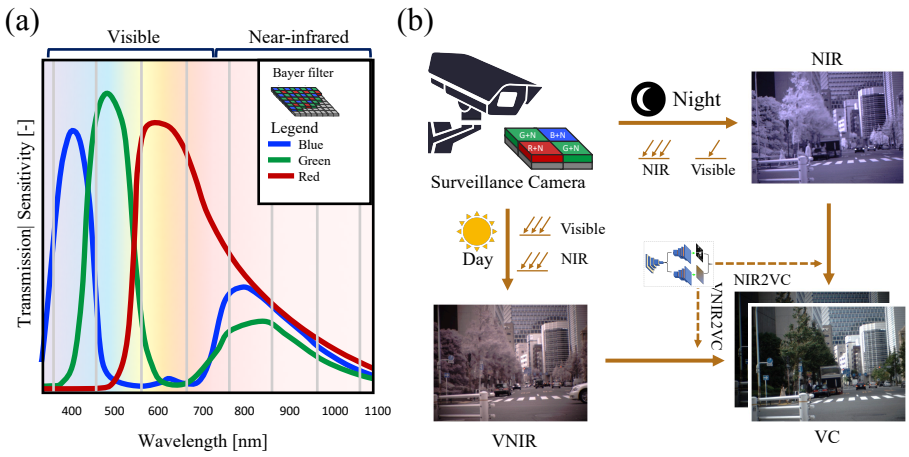

Fig. 1. (a) The spectrum response of the Bayer silicon sensor; and (b) Our integrated pipeline for video surveillance in a day.

visible and near-infrared (NIR) spectrum (as shown in Figure 1 (a)). During daytime, because of the mixture of visible and NIR signal, the captured visible and near-infrared (VNIR) imagery suffers from severe color and contrast degradation [35], as shown in Figure 1 (b). While at nighttime, due to the deficient level of illumination, getting visible color imagery is pretty challenging [44].

For all-day surveillance, the industry practice is to adopt a switchable infrared cut filter (IRCF) to physically block NIR signal at daytime, and to use NIR LEDs, usually centered at 850 *nm* for illumination during nighttime [40]. This switching mechanism is troubled with frequent maintenance of the mechanical parts. Besides, even NIR shares many properties with visible light, NIR imagery contains less color or texture information, which might hinder human interpretation as well as high-level computer vision applications, *e.g.*, visual tracking [8] and object recognition [19].

To resolve the first drawback, dual-sensor camera systems adopt a beam splitter to split the light and then capture visible and NIR images independently [30]. These systems are free from moving parts, and can directly generate paired visible and NIR images without further image processing steps. Another choice is to use a multispectral filter array (MSFA), which separately records visible and NIR signals in a specially mosaiced sensor [28, 20]. The MSFA system can get rid of the mechanical IR cut filter and produce visible as well as NIR images simultaneously through specialized demosaicing [35]. These two solutions capture NIR images independently, which can be used for further visible color image enhancement, such as denoising [10], deblurring [24], and dehazing [38]. However, these two solutions, especially the dual-sensor systems, are relatively expensive and limited to professional usage.

Rather than adding an extra set of imaging sensors or modifying the color filter array, we propose a software solution by training deep convolutional networks (DCN) to realize the automatic translation from the VNIR or NIR input to visible color output. With our proposed model, visible color information can be extracted from the mixed VNIR imagery during the daytime. While at nighttime, NIR images will be colorized into visible color images. To train such a model, it is a big challenge to get sufficient data with ground truth image pairs, *i.e.*, paired VNIR&VC images of daytime, and NIR&VC images of nighttime. Currently, publicly available datasets partially contain either VNIR&VC [34] or NIR&VC [7, 6] image pairs. They are limited to small-scale static scenes only[29], thus inappropriate for surveillance usages, for which moving objects like vehicles and pedestrians are of central importance. Besides, because these paired images are captured from different light sources or view angles, there are obvious distortions and misalignments in each pair [11]. To address these problems, we propose a novel optical system with a beam splitter followed by two geometrically aligned and temporally synchronized sensors. We add a NIR cut filter to capture VNIR&VC video pairs and a NIR bandpass filter to capture NIR&VC video pairs. We also introduce large-scale video surveillance in a day (VSIAD) dataset, which is likely to boost other researches.

In order to fully exploit the potential of the dataset, efficient and generalized algorithms are very critical. In recent years, deep convolutional networks (DCNs), including various generative adversarial networks (GANs), have shown promising results for various image-to-image translation tasks [14, 48]. Among them, there is a relatively similar topic termed image colorization, which aims to colorize low chrominance images into visible color images. Since texture information is well provided, chrominance recovery is the only issue to be addressed [45, 12]. However, in our task, due to the complexity of light sources, only learning chrominance is not sufficient. Hence, we propose a novel multi-task fully convolutional network with state synchronization modules, to learn proper texture and chrominance information from multispectral inputs. To evaluate our approach, we conduct comparison experiments on the newly captured VSIAD dataset. Inspired by the existing researches [21, 23], peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [43], learned perceptual image patch similarity (LPIPS) [46] as well as human judgment are used for our image quality evaluation. The results demonstrate that the proposed network achieves considerable translation accuracy in both VNIR2VC and NIR2VC tasks, and outperforms the state-of-the-art image colorization techniques.

The main contributions of this study can be summarized as follows:

- We design a novel optical system to capture well-aligned VNIR&VC and NIR&VC image pairs and contribute a large-scale dataset, video surveillance in a day (VSIAD).
- We demonstrate a software solution of recovering visible color for all-day video surveillance, in contrast to existing hardware solutions that require switchable filters, multispectral filter arrays, or dual sensors.

– We propose a novel multi-task fully convolutional network with state synchronization modules to ensure the consistency of the generated texture and chrominance information.

2 Related work

To our best knowledge, there is no other research at present trying to handle both VNIR2VC and NIR2VC in a unified network for video surveillance in a day. Instead, several similar studies are working on each topic separately.

VINR2VC. VNIR2VC is a typical imaging problem that aims to extract vivid visible color images from multispectral VNIR images. Zhang *et al.*[47] proposed a dual-camera system using a 45° hot mirror to separate visible from NIR light, and then captured them independently. Additionally, Kise *et al.*[18] designed a triple camera system equipped with interchangeable optical filters. Rather than using two or more camera systems, multispectral filter arrays (MSFA) offer an alternative option. Lu *et al.*[28] presented a customized 4×4 CFA through spatial-domain optimization, which enables the extraction of visible and NIR image pairs from single RAW measurements. Similarly, Chen *et al.*[4] introduces a four-channel bayer pattern (*i.e.*. R, G, B, and NIR) to record visible and NIR signal independently.

Nevertheless, the above-mentioned solutions, which require customized hardware, are relatively expensive and limited to professional usage.

NIR2VC. NIR2VC is slightly different from grayscale image colorization. In grayscale image colorization, the grayscale input and corresponding color output are derived from the same visible color image. Texture information from input and output are almost identical, and chrominance is the only factor to be learned [5, 22]. Zhang *et al.*[45] turned grayscale image colorization as classification of chrominance values. Further, Iizuka *et al.*[12] proposed a multi-task network that combines color prediction and scene classification to achieve more natural results.

Different from grayscale image colorization, NIR2VC is also subject to texture recovering from NIR to visible light (as shown in Figure 1 (a) and Figure 2 (c)). Recently, Berg *et al.*[1] utilized an additional structure loss that can minimize the texture difference between thermal infrared (TIR) and grayscale. To avoid misalignments between images pairs, Mehri *et al.*[31] and Nyberg *et al.*[33] adopt modified CycleGANs [48] for unpaired thermal infrared to visual color (TIR2VC) translation. Because of the loose connection between texture and chrominance, their result suffers from severe blurring as well as mismatching of texture and chrominance.

Deep learning based low light enhancement. Low light enhancement, which aims to enhance image quality under deficient illuminance condition, is significant in video surveillance. Chen *et al.*[3] built a See-in-the-Dark (SID) dataset captured by various exposure time for training a model to brighten extreme dark images. To see motion in the dark, Chen *et al.*[2] and Jiang *et al.*[15] introduced learning-based pipelines to recover texture and chrominance informa-

tion from dynamic scenes. Theoretically, these approaches can also be applied for video surveillance in a day without NIR illuminant. However, in practice, RAW images or videos required as input in their systems, are not available in the majority of existing surveillance cameras.

Image-to-image translation. Both VNIR2VC & NIR2VC can also be viewed as a specific form of image-to-image translation that enables the mapping between an input image and a corresponding output image. Isola *et al.*[14] built a general image-to-image translation framework using conditional adversarial networks. Zhu *et al.*[48] introduced cycle-consistent adversarial networks (cycle-GAN) to get rid of aligned image pairs. For high-resolution image synthesis, Wang *et al.*[42] proposed novel adversarial loss, as well as new multi-scale generator and discriminator architectures. Due to the texture difference between NIR and VC, the NIR2VC task can not be properly addressed with general image-to-image translation methods.

Very recently, Lv *et al.*[29] introduced an integrated enhancement solution for 24-hour colorful imaging. However, their approach is mainly based on indoor static scenes that are inappropriate for surveillance usages, for which moving objects like vehicles and pedestrians are of central importance.

3 Dataset

To enable training, we introduce a novel dataset, video surveillance in a day (VSIAD), which contains ground-truth image pairs of both VNIR&VC and NIR&VC, taken with our co-axis optical imaging system. For data preprocessing, we registered the captured image pairs with feature matching and geometric transformation.

3.1 Data Capturing

The optical imaging system mainly consists of one beam splitter, and two IRCF-free CCD cameras (FLIR GS3-U3-15S4C). A key feature of our system is that we can switch between daytime and nighttime mode easily.

- **Daytime mode.** As shown in Figure 2 (a), light is firstly divided into two branches by a beam splitter. One beam goes to the bayer sensor that yields color imagery containing both visible and NIR information (VNIR). The other one will first pass through the NIR-cut filter to filter out NIR information and then reach the sensor to generate an image of visible color (VC).
- **Nighttime mode.** Note that it is impossible to capture moving objects in low light condition by an ordinary camera. Therefore, we capture NIR and VC pairs at daylight. As shown in Figure 2 (b), nighttime mode utilizes a similar architecture to daytime. To simulate the NIR image captured by a surveillance camera with 850 nm LED illumination, we use an 850 nm bandpass filter, with an FWHM of 50 nm, to filter out visible information.

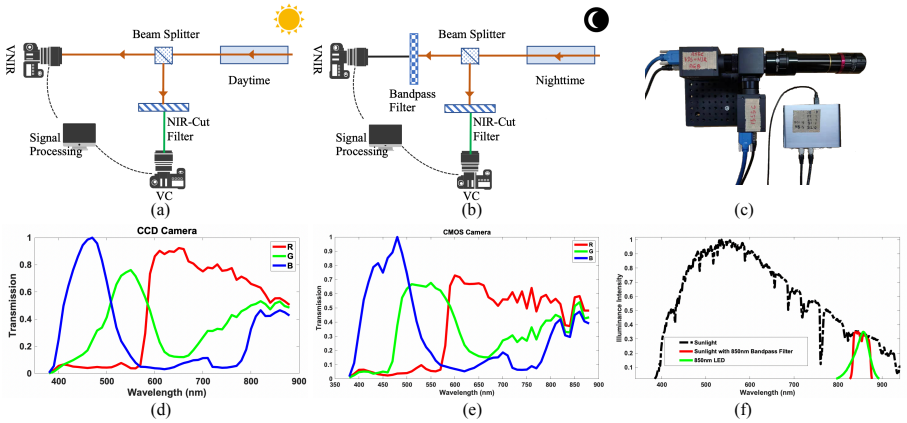

Fig. 2. Overview of the co-axis optical system. (a) The architecture of the daytime mode; (b) The architecture of the nighttime mode; (c) The physical devices; (d) The spectral response of the CCD camera; (e) The spectral response of the CMOS camera; and (f) The spectral distributions of sunlight, sunlight with 850 nm bandpass filter, and 850 nm LED illuminant.

The rationale of this choice is verified by the similarity between the spectral distribution of the conventional 850 LED illuminants and the daylight spectrum filtered by the aforementioned NIR bandpass filter, as shown in Figure 2 (f). Because of the various distributions of light sources (*i.e.*, sunlight at daytime, and LED illuminant at nighttime), there is a slight difference as compared to real-world conditions.

With the proposed imaging system, 80 video clips (40,000 images) were captured from several street spots. All images are saved in 8-bit BMP format with 1384×1032 pixels. The numbers of video clips of VNIR&VC from daytime and NIR&VC from nighttime are set to be equal for the daytime-nighttime balancing.

3.2 Data Preprocessing

Although the two imaging sensors are well-aligned with similar positions, there are inevitably pixel-level rotation or translation misalignments. To address this issue, we employ projective transformation to wrap the VC image based on scale-invariant feature transform (SIFT)[27] features of corresponding VNIR or NIR images. The window with a size of 1200×900 is used to crop the center area of overlapping registered image pairs to eliminate boundary aliases and artifacts. Later, the cropped images are resized to 640×480 to reduce storage.

Figure 3 shows sample image pairs from our VSIAD dataset. The upper row contains four sets of VNIR&VC image pairs taken by the daytime mode. Because of interference of the NIR signal, VNIR images result in apparent color

Fig. 3. A subset of the VSIAD dataset. The upper row is VNIR&VC image pairs taken by the daytime mode, while the remaining row is NIR&VC image pairs taken by nighttime mode.

degradation. Sampled image sets of NIR&VC image pairs taken by the nighttime mode are presented at the bottom row. Due to the NIR spectrum characteristic, captured NIR images present a strong signal on green plants but a weak signal on dyes (*e.g.*, color on clothes, and character of the signpost).

4 State Synchronization Network

Inspired by existing end-to-end fully convolutional networks [26, 25], we design a novel state synchronization network (SSN), which utilizes parallel decoders and state synchronization modules (SSMs) to estimate texture and chrominance information separately. Differing with existing colorization methods that use grayscale information as input, our model directly uses RGB values of VNIR or NIR to prevent information loss during the conversion. We note that, although the camera responses of three color channels around 850 nm (see Figure2 (c)) are quite similar, they are indeed slightly different. Thus, visible color recovery from NIR (*i.e.*, NIR2VC) is less ill-posed than retrieving chrominance from a single-channel grayscale image.

Network Architecture. As shown in Figure 4, the proposed SSN consists of one encoder and two parallel decoders with four state synchronization modules. The encoder follows the design of the classic ResNet [9] using sequential basic residual blocks [9] and max-pooling layers. The parallel decoders share the identical architecture except for the final prediction layer. Similar to the encoder, the decoder applies sequential deconvolutional layers [32], residual blocks, and skip connections [36] to refine to original height and width gradually. To avoid interference within batch samples, batch normalization layers [13] are replaced by instance normalization [41].

State Synchronization Module (SSM). For an image, the texture and chrominance information are highly correlated (*e.g.*, tree texture usually cor-

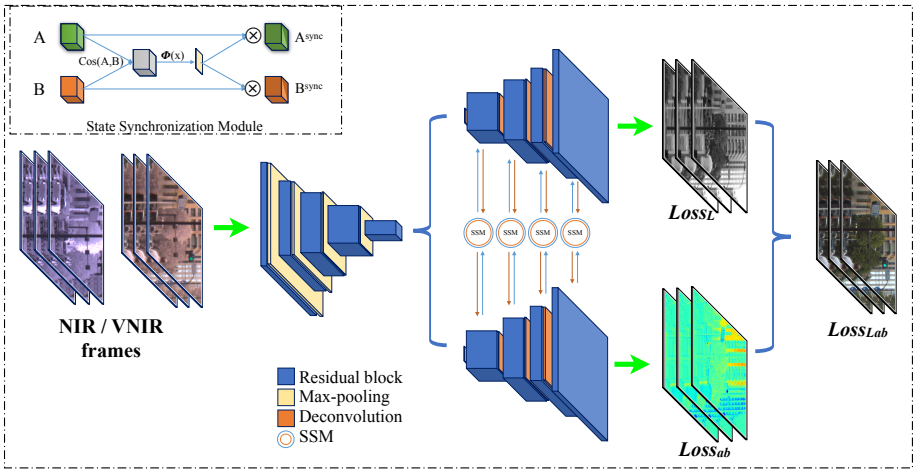

Fig. 4. Architecture of the proposed state synchronization network (SSN). Input VNIR or NIR frames will be translated into visible color images by the network.

relates with green color). If we train L and ab independently, consistency of texture and chrominance will be misconducted.

To avoid this, we design the state synchronization module (SSM) to update the state of the parallel decoders continuously. As shown in Figure 4, features from L or ab decoder are denoted as $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. Within the SSM, the cosine similarly ($\boldsymbol{CS}$) followed by 2D convolution with $k \times k$ gaussian kernel ($\boldsymbol{G}$) are applied to the generated state map of both features ($\boldsymbol{S_{map}}$). Specifically,

$$\boldsymbol{CS} = \frac{\boldsymbol{A} \bullet \boldsymbol{B}}{||\boldsymbol{A}|| \times ||\boldsymbol{B}||} \tag{1}$$

$$\boldsymbol{S_{map}} = \sum_i \sum_j \boldsymbol{CS}_{i:i+k,j:j+k} \bullet \boldsymbol{G} \tag{2}$$

Then, $\boldsymbol{S_{map}}$ is applied to update feature $\boldsymbol{A}$ and $\boldsymbol{B}$ as $\boldsymbol{A}^{sync}$ and $\boldsymbol{B}^{sync}$ through hadamard product by each channel (c) as follows

$$\boldsymbol{A^{sync}}_c = \boldsymbol{A}_c \odot \boldsymbol{S_{map}} \tag{3}$$

$$\boldsymbol{B^{sync}}_c = \boldsymbol{B}_c \odot \boldsymbol{S_{map}} \tag{4}$$

Objective function. After parallel decoders with SSMs, predictions of L and ab are generated separately. Structural dissimilarity (DSSIM) [43] and L1 distance between these predictions and ground truths are denoted as $\mathcal{L}_L$ and $\mathcal{L}_{ab}$, respectively.

$$\mathcal{L}_L = \frac{1 - \boldsymbol{SSIM}(L^{pred} - L^{gt})}{2} \tag{5}$$

$$\mathcal{L}_{ab} = |ab^{pred} - ab^{gt}| \tag{6}$$

Finally, independent predictions of L and ab are concatenated as predicted Lab. To evaluate the consistency of texture and corresponding chrominance, a perceptual loss ($\mathcal{L}_{Lab}$) [16] is calculated.

$$\mathcal{L}_{Lab} = \sum_{layer} |\boldsymbol{VGG}(Lab^{pred}) - \boldsymbol{VGG}(Lab^{gt})|_{layer} \tag{7}$$

where layer $\in [relu1\_2, relu2\_2, relu3\_3, relu4\_3]$ of pre-trained VGG16 network [39]. Thus, the final objective function is formulated as:

$$\mathcal{L}_{final} = \alpha \times \mathcal{L}_L + \beta \times \mathcal{L}_{ab} + \lambda \times \mathcal{L}_{Lab} \tag{8}$$

In our experiments, the configuration of the parameters are set as: $k = 3$, $\alpha = 1$, $\beta = 1$, $\lambda = 10$, respectively.

5 Experimental Setup

We split 20,000 image pairs in our VSIAD dataset into training, validation, and testing. Their ratios are 60:20:20 so that there are 12,000 image pairs for training, 4000 for validation, and 4000 for testing. The numbers of VNIR&VC and NIR&VC pairs are set to be equal by each set. We select a batch size of 8 and randomly crop 256×256 patches from a full-resolution VNIR or NIR image as input for training. We implement the proposed networks using PyTorch 1.0 (https://github.com/huster-wgm/VSIAD) and train it with NVIDIA Tesla V100 . The proposed model is trained for 100,000 iterations with 100 validations performed by every 1,000 iterations. In our experiment, parameters are optimized by the Adam optimizer [17] using initial learning rate = $1e^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e^{-8}$.

5.1 Baselines

For comparison, we choose several representative image colorization methods: Iizuka *et al.*[12], which jointly learns global and local features to exploit classification labels for better colorization; Berg *et al.*[1] that introduces a combination loss for generated texture and chrominance information; Mehri *et al.*[31] and Nyberg *et al.*[33] which apply modified CycleGANs [48] for unpaired thermal infared to visual color (TIR2VC) translation; and pix2pixHD [42], a general high-resolution image-to-image translation framework.

For Iizuka *et al.*'s model, we first try to fine-tune their model on visible images in our VSIAD dataset. However, due to the lack of classification labels,

the performance gained from fine-tuning is minimal (± 0.2 for PSNR). Thus, we directly use the pre-trained model for comparisons. As for Berg *et al.*'s model, we carefully implement and train it from scratch using the standard setup in the literature. Thanks to the publicly available training code, we fine-tune Mehri *et al.*'s, Nyberg *et al.*'s, and pix2pixHD models using our VSIAD dataset.

6 Results and Discussions

6.1 Quantitative Evaluation

To evaluate our method and the baselines, evaluation metrics, including pixel-based PSNR, structure-based SSIM, and learning-based LPIPS, are adopted. Note that the lower score of LPIPS indicates better image quality.

The relative performances of different methods over testing data are listed in Table 1. In general, values of PSNR, SSIM, and LPIPS in the VNIR2VC task are higher than those in the NIR2VC task.

Compared with other baselines, pix2pixHD [42] and our SSN present significantly higher scores in PSNR and SSIM as well as lower scores in LPIPS. As for pixel-based metrics, pix2pixHD has a higher PSNR value in the VNIR2VC task. However, in the NIR2VC task, our model performs as good as pix2pixHD. For more generalized metrics, SSIM and LPIPS, our model outperforms all baselines in both VNIR2VC and NIR2VC tasks. These numbers are consistent with our qualitative observation (see details in Section 6.2). Besides, comparing with pix2pixHD, our model shows a relatively smaller performance gap between VNIR2VC and NIR2VC tasks. These results indicate that our proposed network can handle both VNIR2VC and NIR2VC tasks efficiently and accurately.

6.2 Qualitative Results

Qualitative comparison of our model against baselines on both VNIR2VC and NIR2VC tasks are shown in Figure 5 and 6, respectively. The sequential input images, including VNIR and NIR, are derived from the same location but

Table 1. Performance comparison on both VNIR2VC and NIR2VC tasks. Metric with '↑' means the higher the better image quality, while '↓' means the opposite.

Method	PSNR ↑		SSIM ↑		LPIPS ↓	
	VNIR2VC	NIR2VC	VNIR2VC	NIR2VC	VNIR2VC	NIR2VC
Iizuka *et al.*[12]	14.812	14.465	0.662	0.513	0.321	0.460
Berg *et al.*[1]	20.188	16.543	0.755	0.622	0.236	0.370
Mehri *et al.*[31]	22.359	14.025	0.779	0.491	0.223	0.454
Nyberg *et al.*[33]	21.096	16.474	0.754	0.573	0.174	0.360
pix2pixHD [42]	**25.003**	19.654	0.790	0.641	0.139	0.287
Ours	24.336	**19.690**	**0.836**	**0.698**	**0.109**	**0.248**

Fig. 5. Qualitative result from the baselines and our SSN model on the VNIR2VC task. The sequential frames are derived from the same location but different viewpoints.

different viewpoints. Within these images, moving vehicles and pedestrians are coarsely presented.

Compared to the baselines mentioned above, our method, as well as the fine-tuned pix2pixHD model, generates images with higher color fidelity. On the VNIR2VC task, images generated by pix2pixHD and our model show somewhat similar chrominance and slight sharpness differences of texture. Generally, images from pix2pixHD are sharper than those from us. However, their images are too sharp that perceptually unnatural (*e.g.*, leaves in trees at the 2^{nd} row, Figure 5).

On the NIR2VC task, even with some artifacts, our method shows significantly better translated images than pix2pixHD (*e.g.*, cars from the 2^{nd} and 6^{th} rows, Figure 6). Considering both VNIR2VC and NIR2VC, which is critical for video surveillance in a day, our method yields the most consistent visual result using both VNIR and NIR inputs.

6.3 Perceptual Experiments

We evaluate the perceptual quality of the generated images through blind testing. In each inquiry, we present the participant with videos. At every frame, VNIR&VC (or NIR&VC) image pairs and corresponding images generated from ours or a baseline model are organized side by side. The participants are asked to pick up the one that is more close to the original visible color video. In the experiment, 134 feedbacks are collected and listed in Table 2. Videos generated by our SSN achieve a significantly higher preference rate under blind, subjective judgment.

Fig. 6. Qualitative result from the baselines and our SSN model on the NIR2VC task. The sequential frames are derived from the same location but different viewpoints.

Table 2. Preference rates of videos generated by different methods.

Tasks	Preference Rate		
	Ours	pix2pixHD [42]	No preference
VNIR2VC	**73.5**%	4.4%	22.1%
NIR2VC	**82.3**%	1.5%	16.2%

6.4 Generalization Analysis

To evaluate the robustness and generalization capability of the proposed method, we test our trained model on real-world time-elapse images captured from a static viewpoint. We also use a CMOS camera (FLIR BFS-U3-63S4C) and remove its IR-cut filter, which is different from the CCD camera (FLIR GS3-U3-15S4C) in training data capture. Despite their difference, we can see that their spectral response curves are quite similar (*e.g.*, Figure 2 (d) and (e)).

As shown in Figure 7, even with some artifacts, our model can generate proper visible images from VNIR or NIR images captured by the CCD camera at most times. Because of the difference in spectral response curves, VNIR/NIR images captured by CMOS show significantly different color style when compared with those images taken by CCD (1^{st} vs. 3^{rd} row). Despite this, our model can produce pretty natural visible images during daytime (*e.g.*, at 09:00, 12:00, and 15:00) using the CMOS camera. As for the nighttime (*e.g.*, at 18:00 and 21:00), CMOS results are less satisfactory. We note that cross-camera color recovery at night is extremely challenging, because of the inevitable interference by light

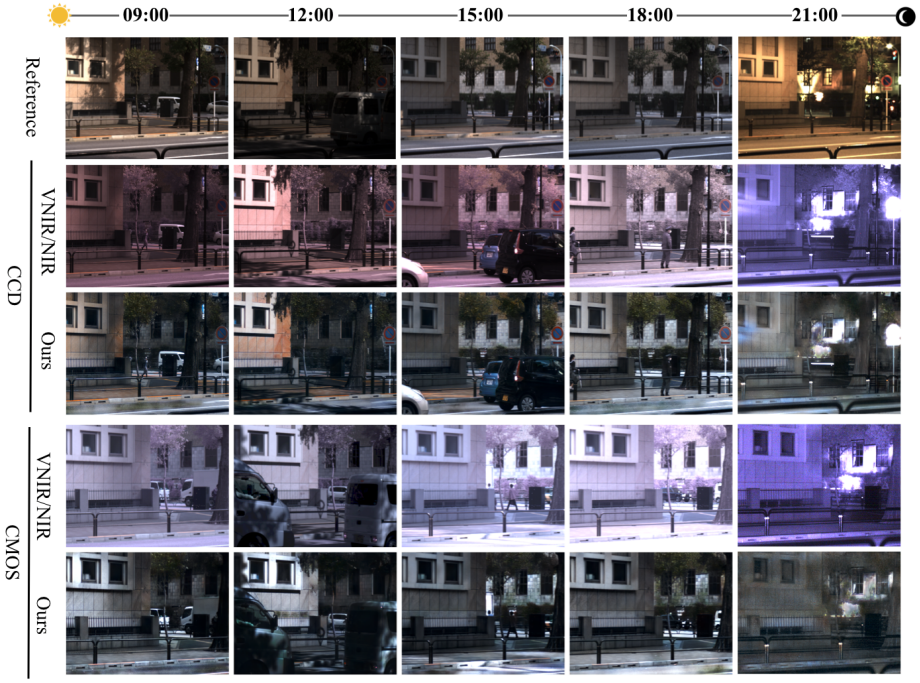

Fig. 7. Time-elapse experiment on both CCD and CMOS cameras. Reference images are captured using a NIR cut filter at daytime and using long exposure at nighttime.

contamination from street light and image noise, which we plan to study further as future work.

6.5 Ablation Experiment

To investigate the effectiveness of different components, we conduct ablation studies on VSIAD. The performance under the different conditions are illustrated in Table 3 and Figure 8.

State Synchronization Module (SSM). In 1^{st} row, removing SSM leads to significant performance decline (*e.g.*, the value of LPIPS increases about 15.6%), which demonstrates the effectiveness and importance of our SSM.

Color. As shown in 2^{nd} row, changing RGB input to grayscale causes performance losses of 2.2% in PSNR, 1.7% in SSIM, and 14.3% in LPIPS.

Texture. As presented at 3^{rd} row, while replacing structural dissimilarity (DSSIM) with L1 distance of $\mathcal{L}_L$ (Eq. 5), a slight performance degradation can be observed.

Normalization. From the 4^{rd} row, if replacing instance normalization (IN) with group normalization (GN), the perceptual performance (*i.e.*, LPIPS) gets worse, while the PSNR and SSIM can be slightly improved.

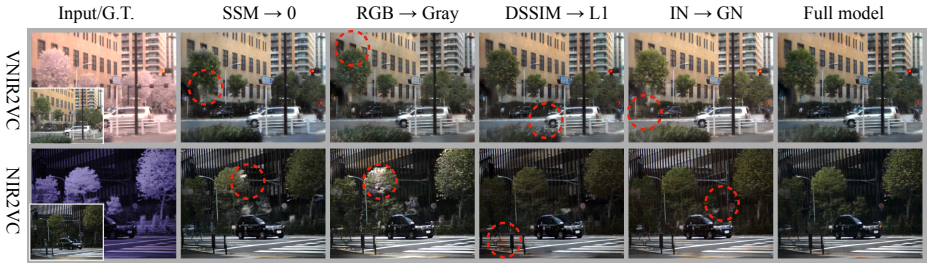

Fig. 8. Representative results of the proposed state synchronization network (SSN) under different conditions.

Table 3. Ablation analysis results. The table reports the mean values of PSNR, SSIM, and LPIPS.

Conditions	PSNR ↑	SSIM ↑	LPIPS ↓
$SSM \rightarrow 0$	21.419	0.747	0.212
$RGB \rightarrow Gray$	21.535	0.754	0.209
$DSSIM \rightarrow L1$	21.910	0.759	0.188
$IN \rightarrow GN$	**23.145**	**0.773**	0.182
Full model	22.013	0.767	**0.179**

7 Conclusion

We have demonstrated the effectiveness of our integrated pipeline for video surveillance in a day. Degraded images, including VNIR and NIR, are directly translated into visible color images through a learned model. In contrast to existing hardware solutions that require switchable filters, multispectral filter arrays, or dual sensors, our approach can directly apply to commercial surveillance cameras that are much more cost-efficient. To enable training, we collect a new dataset that contains well-aligned VNIR&VC and NIR&VC image pairs, and introduce a novel parallel network with state synchronization modules that can keep consistency between texture and chrominance information. We also notice that slight performance degradation happened during cross-camera color recovery, which we plan to study further as future work.

Acknowledgements

This work was supported in part by the JSPS KAKENHI under Grant No. 19K20307. A part of this work was finished during Y. Zheng's visit and X. Ding's internship at Peng Cheng Laboratory.

References

1. Berg, A., Ahlberg, J., Felsberg, M.: Generating visible spectrum images from thermal infrared. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1143–1152 (2018)
2. Chen, C., Chen, Q., Do, M.N., Koltun, V.: Seeing motion in the dark. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3185–3194 (2019)
3. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3291–3300 (2018)
4. Chen, Z., Wang, X., Liang, R.: Rgb-nir multispectral camera. Optics express **22**(5), 4985–4994 (2014)
5. Cheng, Z., Yang, Q., Sheng, B.: Deep colorization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 415–423 (2015)
6. Choe, G., Kim, S.H., Im, S., Lee, J.Y., Narasimhan, S.G., Kweon, I.S.: Ranus: Rgb and nir urban scene dataset for deep scene parsing. IEEE Robotics and Automation Letters **3**(3), 1808–1815 (2018)
7. Fredembach, C., Süsstrunk, S.: Colouring the near-infrared. In: Color and Imaging Conference. vol. 2008, pp. 176–182. Society for Imaging Science and Technology (2008)
8. Gao, S., Cheng, Y., Zhao, Y.: Method of visual and infrared fusion for moving object detection. Optics letters **38**(11), 1981–1983 (2013)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Honda, H., Timofte, R., Van Gool, L.: Make my day-high-fidelity color denoising with near-infrared. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 82–90 (2015)
11. Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1037–1045 (2015)
12. 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 Transactions on Graphics (TOG) **35**(4), 110 (2016)
13. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
14. 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)
15. Jiang, H., Zheng, Y.: Learning to see moving objects in the dark. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7324–7333 (2019)
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Kise, M., Park, B., Heitschmidt, G.W., Lawrence, K.C., Windham, W.R.: Multispectral imaging system with interchangeable filter design. Computers and Electronics in Agriculture **72**(2), 61–68 (2010)

19. Kleynen, O., Leemans, V., Destain, M.F.: Development of a multi-spectral vision system for the detection of defects on apples. Journal of food engineering **69**(1), 41–49 (2005)
20. Koyama, S., Inaba, Y., Kasano, M., Murata, T.: A day and night vision mos imager with robust photonic-crystal-based rgb-and-ir. IEEE Transactions on Electron Devices **55**(3), 754–759 (2008)
21. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 170–185 (2018)
22. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: European Conference on Computer Vision. pp. 577–593. Springer (2016)
23. Lei, C., Chen, Q.: Fully automatic video colorization with self-regularization and diversity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3753–3761 (2019)
24. Li, W., Zhang, J., Dai, Q.H.: Robust blind motion deblurring using near-infrared flash image. Journal of Visual Communication and Image Representation **24**(8), 1394–1413 (2013)
25. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
26. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
27. Lowe, D.G., et al.: Object recognition from local scale-invariant features. In: iccv. vol. 99, pp. 1150–1157 (1999)
28. Lu, Y.M., Fredembach, C., Vetterli, M., Süsstrunk, S.: Designing color filter arrays for the joint capture of visible and near-infrared images. In: 2009 16th IEEE International Conference on Image Processing (ICIP). pp. 3797–3800. IEEE (2009)
29. Lv, F., Zheng, Y., Li, Y., Lu, F.: An integrated enhancement solution for 24-hour colorful imaging. In: AAAI. pp. 11725–11732 (2020)
30. Matsui, S., Okabe, T., Shimano, M., Sato, Y.: Image enhancement of low-light scenes with near-infrared flash images. Information and Media Technologies **6**(1), 202–210 (2011)
31. Mehri, A., Sappa, A.D.: Colorizing near infrared images through a cyclic adversarial approach of unpaired samples. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 971–979. IEEE (2019)
32. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1520–1528 (2015)
33. Nyberg, A., Eldesokey, A., Bergström, D., Gustafsson, D.: Unpaired thermal to visible spectrum transfer using adversarial training. In: European Conference on Computer Vision Workshops. pp. 657–669. Springer (2018)
34. Özkan, K., Işık, Ş., Topsakal Yavuz, B.: Identification of wheat kernels by fusion of rgb, swir, vnir samples over feature and image domain. Journal of the Science of Food and Agriculture (2019)
35. Park, C., Kang, M.: Color restoration of rgbn multispectral filter array sensor images based on spectral decomposition. Sensors **16**(5), 719 (2016)
36. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)

37. Sadeghipoor, Z., Lu, Y.M., Süsstrunk, S.: A novel compressive sensing approach to simultaneously acquire color and near-infrared images on a single sensor. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1646–1650. IEEE (2013)
38. Schaul, L., Fredembach, C., Süsstrunk, S.: Color image dehazing using the near-infrared. In: 2009 16th IEEE International Conference on Image Processing (ICIP). pp. 1629–1632. IEEE (2009)
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
40. Tessler, N., Medvedev, V., Kazes, M., Kan, S., Banin, U.: Efficient near-infrared polymer nanocrystal light-emitting diodes. Science **295**(5559), 1506–1508 (2002)
41. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
42. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
43. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
44. Zafar, I., Zakir, U., Romanenko, I., Jiang, R.M., Edirisinghe, E.: Human silhouette extraction on fpgas for infrared night vision military surveillance. In: 2010 Second Pacific-Asia Conference on Circuits, Communications and System. vol. 1, pp. 63–66. IEEE (2010)
45. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European conference on computer vision. pp. 649–666. Springer (2016)
46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
47. Zhang, X., Sim, T., Miao, X.: Enhancing photographs with near infra-red images. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
48. Zhu, J.Y., 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. pp. 2223–2232 (2017)