Automatic Colorization of Images from Chinese Black and White Films Based on CNN

Yu Chen, Yeyun Luo, Youdong Ding, Bing Yu
Shanghai Film Academy
Shanghai University
Shanghai, China
youyingyin@i.shu.edu.cn

***Abstract*—The colorization of black and white films was a hot topic in the 1980s. Some black-and-white movies regained their luster through colorization. Although people are controversial about the artistic value of film colorization, it is no doubt that color images can enhance visual effects. Inspired by the recent colorization methods using deep learning, we propose a novel colorization model which combines two Convolutional Neural Networks and uses multi-scale convolution kernels to get better spatial consistency. Most of the current datasets used in the colorization networks are not applicable to colorizing images from Chinese black and white films. The main reason is that the objects in these films are very different from today's. To address this, we extract a large number of images from Chinese color films of the last century as a training dataset. Experiments demonstrate that our model can obtain pretty good results of colorizing images from Chinese black and white films.**

Keywords—colorization; VGG-16; CNN;

I. INTRODUCTION

In addition to the creators themselves wanting to convey ideas with black and white images, many black-and-white films were produced with restricted shooting equipment. Coloring a black-and-white movie can evoke our memories and can also be an artistic creation. Coloring films with traditional methods is basically manual, which is a time-consuming and labor-intensive job. Since film is composed of a sequence of images, coloring an image is the basis for coloring the film, which is our goal in this paper. With the advent of digital image processing, a large number of colorization methods have emerged. One of them requires the users to draw scribbles [1] to propagate color to neighboring pixels in the image, which requires a large number of user interactions. Another approach colorizes a grayscale image by transferring the color information from the reference image to the target image [2]. The recent methods colorize grayscale images by training CNNs on large -scale image datasets.

Inspired by the recent colorization approaches, we present a network which combines a colorization model based on [3] with some features extracted from the VGG-16 [4] pre-trained model. Iizuka et al. [5] use high-level semantic information (categories of scenes) to help build their networks, however, those high-level features may not be able to provide enough information about the contents since humans are the main objects in films. We use the low and middle features extracted from VGG-16 instead. Also, we expect to get better spatial consistency by using convolution kernels with different sizes.

The previous colorization models are mostly trained on ImageNet, Places, SUN, etc. Those modern pictures are very different from those of the last century. We extract a large number of color images from Chinese color films of the last century as an additional dataset to fine-tune our model.

II. RELATED WORK

Welsh et al. [2] proposes a colorization approach which transfers the color from a reference image to the target image. Basically, local descriptors [6, 7] are used to create mapping functions between reference images and target images. Manual intervention [8, 9] is added in later researches.

Another colorization approach called scribble-based method is proposed by Levin et al. [1]. Users need to manually assign colors to the specified area. These assigned colors are propagated under a premise that neighboring pixels should have similar color if their intensities are similar. This similarity is extended to texture and pattern in [10, 11]. Huang et.al [12] exploit an edge detection scheme to prevent color bleeding.

In the past few years, a lot of automatic methods [3, 13, 14, 15, 16] have been proposed. Desphande et al. [13] colorize an image by minimizing a quadratic objective function trained with a LEARCH framework. Cheng Z. et al. [14] present a deep neural network that leverages variety of features, including DAISY feature [17], semantic feature, and the image patch feature. Iizuka, Serra et al. [5] present a network which combines both global priors (the classification of images) and local image features. Larsson et al. [15] develop a network based on VGG-16 with the classification layer discarded and predict a color probability distribution for each pixel. The inherent multi-model problem of colorization is regarded as a task of classification in [3], which also exploits a class rebalancing method to get colorful images. We draw our network based on [3]. What is more, we combine it with a VGG-16 pre-trained model. Zhang et al. [16] present an approach which combines automatic colorization method with user interactions, allowing users to modify the color at any location. In this paper, our objective is fully automatic colorization.

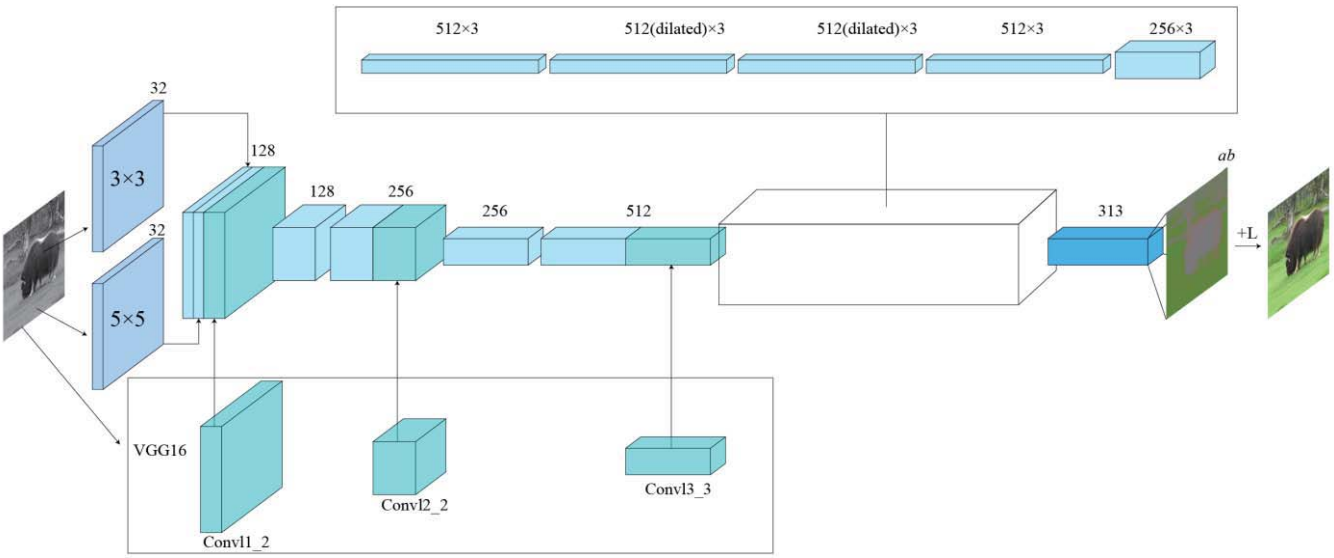


Fig. 1. Our network is based on [3]. We combine it with some layers extracted from VGG-16.

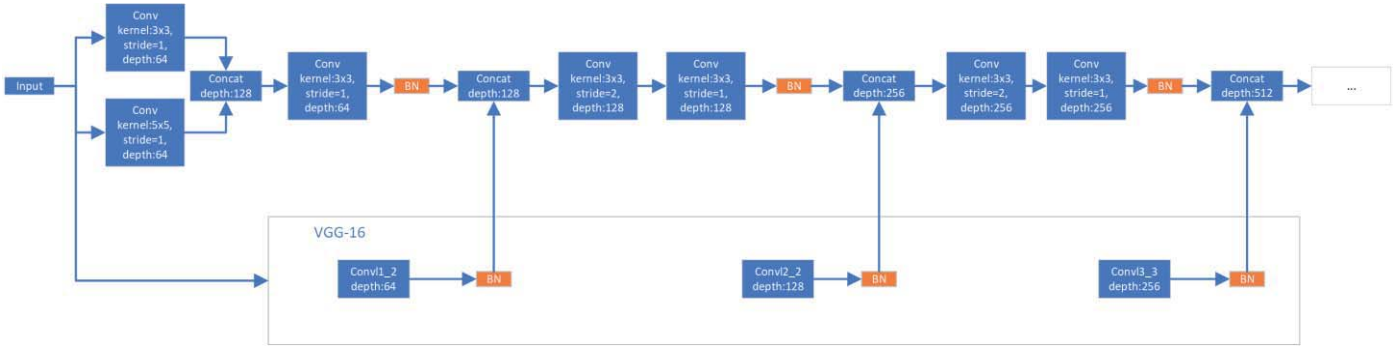


Fig. 2. Detail of the part we combine [3] with VGG-16.

III. The Proposed Neural Network

A. Neural network

We draw our model based on the network proposed by [3]. Fig.1 shows the architecture. Instead of using only a 3×3 convolution kernel to process convolution on the first layer, we use two convolution kernels of 3×3 and 5×5, which proven in the experiment section to be able to get better spatial consistency.

Most of CNN classification models, in addition to be used for classification tasks, also contain more information that could be extracted. Zeiler and Fergus [18] show the intermediate layers of a CNN visually. As the network level increases, we can see the approximate outline of many objects. These intermediate layers are able to provide much representation information. To obtain prior information for our colorization model, we utilize some features extracted from VGG-16, which has a simple architecture and pretty good classification performance. We extract a few layers (Convl_2, Convl2_3, and Convl3_3) by feeding images to VGG-16 network during training. Since VGG-16 is trained without batch normalization [19], we first execute batch normalization before concatenate these layers to [3]. More detail about the part that we combine [3] with VGG-16 is shown in Fig.2. Since the VGG-16 network is trained on color images, the gray images perform poorer than color images on the classification task, which means fewer representations are learned. As described in [15], before combine VGG-16 with the main network, we fine-tune it with grayscale images as input.

B. Loss

We convert the color space to CIE Lab since it is designed to be perceptually linear. For a colorization task, *L* channel is the input and the objective is the *ab* channels.

Color prediction is a multi-model problem because many objects may have many possible colors. To solve this problem, Zhang et al. [3] predict the possible color distribution for each pixel in the image. The *ab* values in color gamut are quantized into blocks, as depicted in Fig.3. In total, $Q = 313$ values are kept. For a given input $\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$, our goal is learning a function $\hat{\mathbf{Z}} = \mathcal{G}(\mathbf{X})$ to maps $\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$ to a probability distribution $\hat{\mathbf{Z}} \in [0,1]^{H \times W \times Q}$. In order to calculate the loss,

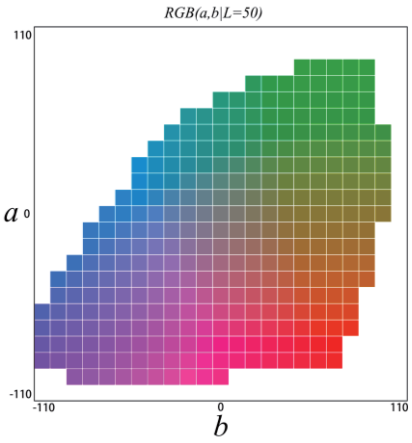


Fig. 3. Quantized ab color space. In total, 313 *ab* values are kept.

Fig. 4. The colorization results by setting different values of *T*.

$\mathbf{Z} = \mathcal{H}_{gt}^{-1}(\mathbf{Y})$ is defined to converts ground truth color $\mathbf{Y}$ to vector $\mathbf{Z}$. This conversion is done by using a Gaussian kernel $\sigma = 5$ to assign different weights to the 5 nearest *ab* values to $\mathbf{Y}_{h,w}$ based on the distance between them and $\mathbf{Y}_{h,w}$.

Reference [3] use a class-rebalancing method to generate colorful images. That procedure works great on most outdoor scenes, such as landscapes, however, it leads to colorful but spatially inconsistent results in our most cases. We discard the class-rebalancing process instead.

The cross entropy loss is defined:

$$\mathrm{L}(\hat{\mathbf{Z}}, \mathbf{Z}) = -\sum_{h,w}\sum_{q}\mathbf{Z}_{h,w,q}\log\left(\hat{\mathbf{Z}}_{h,w,q}\right) \tag{1}$$

Where $\mathbf{Z}_{h,w,q}$ represents the possibility that the *ab* value of the pixel at (h,w) is q in ground truth image, and $\hat{\mathbf{Z}}_{h,w,q}$ corresponds to the predicted possibility.

C. Distribution to Point

Once we get the distribution $\hat{\mathbf{Z}}$ as described in the previous section, $\mathcal{H}$ is defined to map $\hat{\mathbf{Z}}$ to $\hat{\mathbf{Y}}$ in *ab* space. There are two options to calculate the final color:

1) Mode: Using the mode of the predicted distribution as the final color. Sometimes it results in spatial inconsistency.

2) Mean: Taking the mean of the predicted distribution, which may lead to desaturated results.

To get a balance between these two options, a parameter *T* is added to readjust the softmax distribution. $\mathcal{H}$ is defined as below:

$$\mathcal{H}(\mathbf{Z}_{h,w}) = \mathbb{E}\left[f_T(\mathbf{Z}_{h,w})\right],\ f_T(\mathrm{z}) = \frac{\exp(\log(\mathrm{z})/T)}{\sum_q \exp(\log(\mathrm{z}_q)/T)} \tag{2}$$

As we can see, setting $T = 1$ means take the mean of the distribution, while setting $T \to 0$ is equivalent to taking the mode of the distribution. Fig.4 shows the colorization results by setting different values of *T*, $T = 0.5$ works fine for our model in most cases.

IV. Experiment And Analysis

A. Making of Dataset and Training

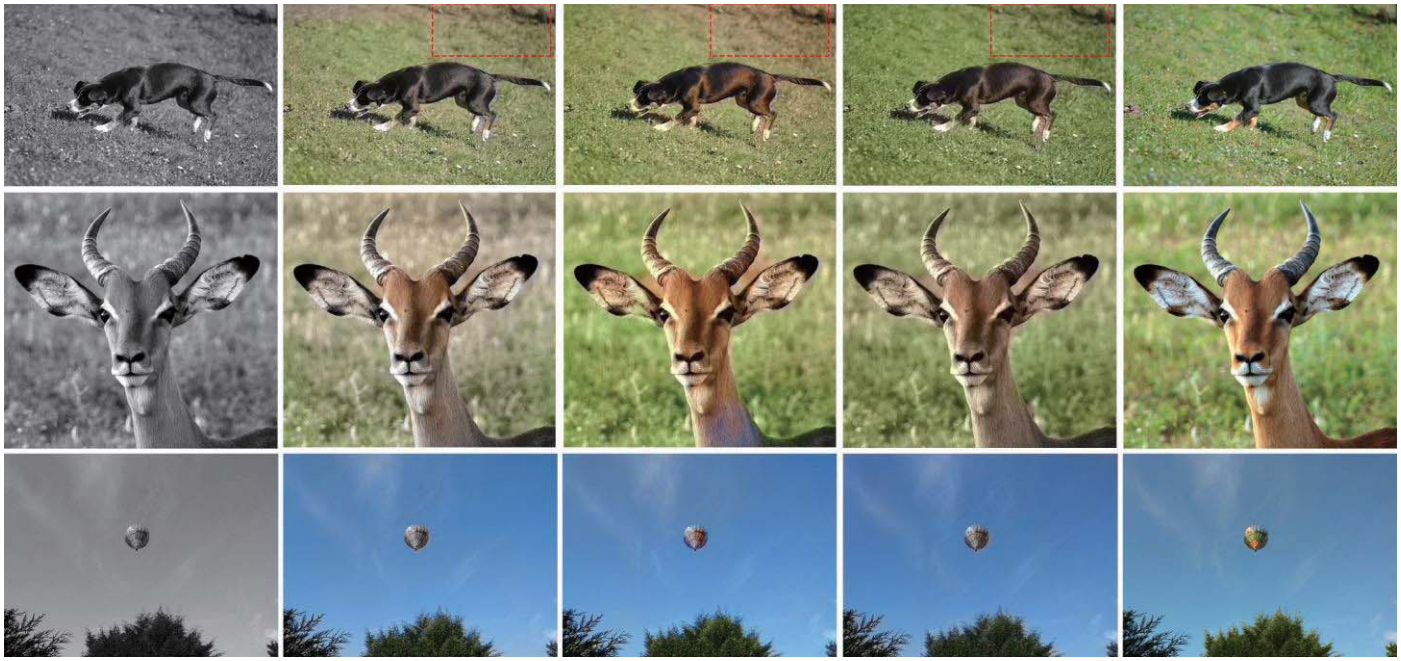
Most previous colorization networks are trained on ImageNet, Places, SUN, etc. These modern images are significantly different from those in Chinese black and white films. We extract a large number of color images from Chinese color films of the last century as an additional dataset. Since the camera equipment was not advanced at the time, the clarity of the aged films was poor and the resolution was low, leading to a challenge for choosing films. In order to simplify training and reduce running times, we finally selected thirty-five films from the 70s to the 90s. We first adjust their white balance and exposure, then select 3 to 5 frames per shot, and finally extract approximately 100,000 images after filtering some low brightness images, which we call data A later.

ImageNet is a widely used training dataset for automatic colorization. Many approaches also use it as an evaluation criterion. Instead of using only the extracting data from old films to train our model, we choose a subset of ImageNet (data B), which contains 500 categories, with 500 photos for each, and use 250,000 images to train our models at the beginning. It is worth mentioning that we used almost the same hyperparameters as [3].

B. Results and Analysis

After training on the data B for about 150,000 iterations, we test our model on some images from ImageNet validation dataset, which have never seen by our model. The result turns out to be pretty good for some of the images, especially for those contain grass and sky. Fig. 5 illustrates the results. Benefit by the features extracted from pre-trained model VGG-16, our network performances as well as [3] while we train our model on only 250,000 images. Due to the kernels with different sizes we used to process the convolution on the input images, as shown in columns 1, our model produces better spatial consistency.

We fine-tune our model with data A for about 60,000 iterations to enable the network to fit old films better and then we test our model on some images from Chinese color old films, which have the ground truth as reference. Fig.6 shows the result using our method along with [3]. In most cases, our model performs better than [3]. However, as we can see from those images, the results of our mode prone to blue tone, which is unsurprising, as most of the Chinese old color films we have used to extract images have cool tone. The tone biases too seriously to be completely corrected. Finally, we test our model on some images from Chinese black and white films, which



(a) (b) (c) (d) (f)

Fig. 5. Comparison of the results of our tests on ImageNet with [3]. (a) The grayscale input images. (b) The results from [3] without color rebalance. (c) The results from [3] with color rebalance. (d) Our results.(f) The corresponding ground truth images. All images are from the internet.

(a) (b) (c) (d) (f)

Fig. 6. Comparison of the results of our tests on Chinese color old films with [3]. (a) The grayscale input images. (b) The results from [3] without color rebalance. (c) The results from [3] with color rebalance. (d) Our results. (f) The corresponding ground truth images. All images are from the internet.

(a) (b) (c) (d)

Fig. 7. Comparison of the results of our tests on Chinese black and white films with [3]. (a) The grayscale input images. (b) The results from [3] without color rebalance. (c) The results from [3] with color rebalance. (d) Our results. The testing images have no ground truth as reference. All images are from the internet.

have no ground truth as reference. Fig.7 shows the results. Notice that the plant in row 3 is colorized to grayish green by our model. That is due to the low saturation of the Chinese color old films, which are used to create our training dataset.

C. *Subjective evaluation*

Comparing the result of colorization to ground truth by using quantitative metrics is often hard to reveal the visual realism. We use similar evaluation methods proposed by [5], which conduct a survey by asking the question "Do you believe this image is from color films?" to assess the reliability of the ground truth images from color films and the results of testing our model and [3] on images from Chinese black and white films. Images are shown to the users one-by-one within 3 seconds. We show 10 images per type to 30 users. The results are shown in Fig. 8. As we can see, a median of 76.7% respondents consider that the images colorized by our fine-tuning model are extracted from the color films, and the results of our model without fine-tuning only get a median of 48.3%. Notice that the ground truth images get a median of 88.3% reliability because some of those aged images have artifacts in some degree. And also, this is a subjective test. Nevertheless, the subjective evaluation indicates that our model is capable of creating plausible colorization results on images from Chinese black and white films.

V. Conclusion

We have presented a novel automatic colorization neural network, which uses multi-scale convolution kernels and combines low and middle features extracting from VGG-16. Experiments prove that our model is able to perform pretty good colorization on images from Chinese black and white films without any user interventions. In order to address the problem that the current training datasets for colorization do not applicable to historical old photographs, we establish an image dataset by extracting frames from Chinese color films of the last century. Since dataset is as important as network

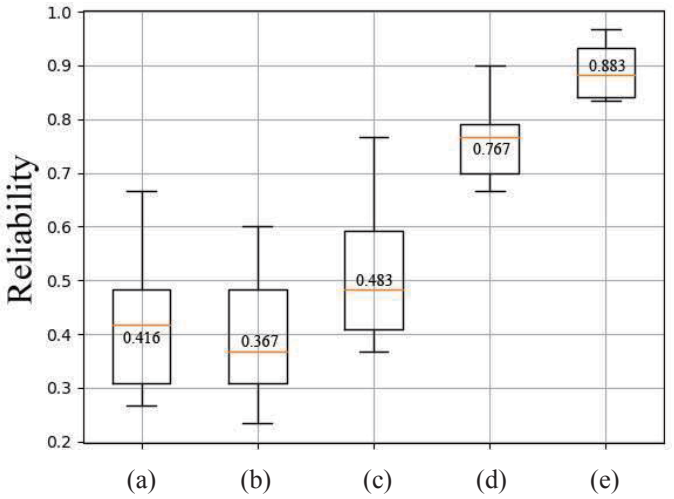


Fig. 8. Results of subjective evaluation assessing the reliability of the ground truth images (e), the results of [3] without color rebalance (a), the results of [3] with color rebalance (b), the results of our model without fine-tuning (c) and the results of our fine-tuning model (d).

structure, one of our future works is to expand our training dataset to enhance the generalization ability of our model. On the other hand, objects such as clothe could take on many possible colors, to have a fully controlling of the colorization results, some degree of human intervention need to be added . Our final objective is building a colorization system for Chinese black and white films.

REFERENCES

[1] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," in ACM Transactions on Graphics (ToG), vol. 23, no. 3, pp. 689-694. ACM, 2004.

[2] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," in ACM Transactions on Graphics (TOG), vol. 21, no. 3, pp. 277-280. ACM, 2002.

[3] R. Zhang, P. Isola, A.A. Efros, "Colorful image colorization." in European Conference on Computer Vision, pp. 649-666. Springer, Cham, 2016.

[4] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in ICLR, 2015.

[5] S. Iizuka, E. Simo-Serra, H. Ishikawa, "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, no. 4 (2016): 110.

[6] G. Charpiat, M. Hofmann, and B. Sch¨olkopf, "Automatic image colorization via multimodal predictions," In European conference on computer vision, pp. 126-139. Springer, Berlin, Heidelberg, 2008.

[7] Y.Morimoto, Y.Taguchi, T. Naemura, "Automatic colorization of grayscale images using multiple images on the web," in SIGGRAPH'09: Posters, p. 32. ACM, 2009..

[8] R. Irony, D. Cohen-Or, and D. Lischinski, "Colorization by example," in Rendering Techniques, pp. 201-210. 2005.

[9] A. Y.-S. Chia, S. Zhuo, R. K. Gupta, Y.-W. Tai, S.-Y. Cho, P. Tan, and S. Lin, "Semantic colorization with internet images," in ACM Transactions on Graphics (TOG), vol. 30, no. 6, p. 156. ACM, 2011.

[10] Q. Luan, F. Wen, D. Cohen-Or, L. Liang, Y.-Q. Xu, and H.-Y. Shum, "Natural image colorization," in Proceedings of the 18th Eurographics conference on Rendering Techniques, pp. 309-320. Eurographics Association, 2007.

[11] Y. Qu, T.-T. Wong, and P.-A. Heng, "Manga colorization," in ACM Transactions on Graphics (TOG), vol. 25, no. 3, pp. 1214-1220. ACM, 2006.

[12] Y.-C. Huang, Y.-S. Tung, J.-C. Chen, S.-W. Wang, and J.-L. Wu, "An adaptive edge detection based colorization algorithm and its applications," in In Proceedings of the 13th annual ACM international conference on Multimedia, pp. 351-354. ACM, 2005.

[13] A. Deshpande, J. Rock, and D. Forsyth, "Learning large-scale automatic image colorization," in Proceedings of the IEEE International Conference on Computer Vision, pp. 567-575. 2015.

[14] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in Proceedings of the IEEE International Conference on Computer Vision, pp. 415-423. 2015.

[15] G. Larsson, M. Maire, G. Shakhnarovich, "Learning representations for automatic colorization," in European Conference on Computer Vision, pp. 577-593. Springer, Cham, 2016.

[16] R. Zhang, J.Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, A. A. Efros. "Real-Time User-Guided Image Colorization with Learned Deep Priors" in SIGGRAPH, 2017.

[17] E. Tola, V. Lepetit, and P. Fua, "A fast local descriptor for dense matching," in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1-8. IEEE, 2008.

[18] M.D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision, pp. 818-833. Springer, Cham, 2014.

[19] S. Ioffe,C.Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proceedings of ICML, pp. 448–456, 2015.