

# Focusing on Persons: Colorizing Old Images Learning from Modern Historical Movies

Xin Jin, Zhonglan Li, Ke Liu  
Department of Cyber Security,  
Beijing Electronic Science and  
Technology Institute  
Fengtai District, Beijing 100070  
China

Dongqing Zou\*  
zoudongqing@sensetime.com  
SenseTime Research and Tetras.AI  
Beijing 100080, China  
Qing Yuan Research Institute,  
Shanghai Jiao Tong University  
Shanghai 200240, China

Xiaodong Li, Xingfan Zhu,  
Ziyin Zhou, Qilong Sun,  
Qingyu Liu  
Department of Cyber Security /  
Cryptography, Beijing Electronic  
Science and Technology Institute  
Fengtai District, Beijing 100070  
China



**Figure 1: We propose a colorization method focusing on persons that considers fine grained semantic parsing and correct color of images. Our method performs better on historical images.**

## ABSTRACT

In industry, there exist plenty of scenarios where old gray photos need to be automatically colored, such as video sites and archives. In this paper, we present the HistoryNet focusing on historical person’s diverse high fidelity clothing colorization based on fine grained semantic understanding and prior. Colorization of historical persons is realistic and practical, however, existing methods do not perform well in the regards. In this paper, a HistoryNet including three parts, namely, classification, fine grained semantic

parsing and colorization, is proposed. Classification sub-module supplies classifying of images according to the eras, nationalities and garment types; Parsing sub-network supplies the semantic for person contours, clothing and background in the image to achieve more accurate colorization of clothes and persons and prevent color overflow. In the training process, we integrate classification and semantic parsing features into the coloring generation network to improve colorization. Through the design of classification and parsing subnetwork, the accuracy of image colorization can be improved and the boundary of each part of image can be more clearly. Moreover, we also propose a novel Modern Historical Movies Dataset (MHMD) containing 1,353,166 images and 42 labels of eras, nationalities, and garment types for automatic colorization from 147 historical movies or TV series made in modern time. Various quantitative and qualitative comparisons demonstrate that our method outperforms the state-of-the-art colorization methods, especially on military uniforms, which has correct colors according to the historical literatures.

\*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '21, October 20–24, 2021, Virtual Event, China

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8651-7/21/10...\$15.00

<https://doi.org/10.1145/3474085.3481544>

## CCS CONCEPTS

• **Computing methodologies** → **Image segmentation.**

## KEYWORDS

Fine grained semantic parsing, Colorization, HistoryNet, MHMD

### ACM Reference Format:

Xin Jin, Zhonglan Li, Ke Liu, Dongqing Zou, and Xiaodong Li, Xingfan Zhu, Ziyin Zhou, Qilong Sun, Qingyu Liu. 2021. Focusing on Persons: Colorizing Old Images Learning from Modern Historical Movies. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*, October 20–24, 2021, Virtual Event, China. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3474085.3481544>

## 1 INTRODUCTION

Image colorization is a classic image editing problem and has always been a research hotspot in the field of image editing. Many black-and-white military photos were left in the war years. The color restoration of these photos can better understand the history and have great significance. Also, automatic colorization has a very wide range of applications in industry. There are a large number of images that need to be colored with modern technology to present better visual effects in some video websites such as YouTube. Many archives and museums also have similar requirements. However, so many colorization methods are not suitable for repairing these photos, they do not consider the fine grained semantic parsing of military uniform and the classification of clothing, which makes the coloring result poor. In this paper, the superiority of HistoryNet network to the colorization of person’s clothing is shown mainly through the colorization of military uniform, which has more practical significance.

Traditional colorization methods [3, 6, 25, 26] propose methods with semantics of input text and language description. Although these methods obtain the semantic segmentation information of reference images, they ignore the boundary of person and background, and also don’t consider the human parsing of each part. With the advances of deep learning, many researchers use CNN or GAN to extract information of grayscale images for colorization such as [5, 18, 20, 23, 27, 28, 36, 40]. These methods often realize the natural color matching, which means the colorization is a subjective problem. However, the colorization of historical military uniform is realistic and practical. In addition, most of the existing modern color image datasets lack the old content or information of real gray historical images, especially the colors of the garments of the historical persons. This will also lead to the poor performance of current colorization methods in military uniforms, as shown in the second line of Figure 1.

To address these shortcomings, we propose a new HistoryNet architecture and a MHMD dataset. First, there are a great variety of military uniform in history and we can obtain the eras, nationalities and garment types information of an image. According to these, we design the labels of images and classification subnetwork. Classification submodule supplies classifying of images according to the eras, nationalities and garment types labels. In addition, we have designed a classifier submodule in discriminator, which joint classification subnetwork to constrain loss for getting more suitable colorization. Inspired by InfoGAN [8], we have designed

classifier subnetwork that can better present the classification information. Second, we have designed fine grained semantic parsing subnetwork, which supply the semantic understanding of historical person’s clothing. Through this subnetwork, the historical person’s each part (such as face, arms, hair and so on) will be separate and obtain a clear boundary finally. In the training process, based on U-net [30], we connect the up-sample layers of image features on the parsing subnetwork with the generator network of colorization. And in the future, we also hope to apply relevant research methods to video colorization. Third, we propose a new dataset called Modern Historical Movies Dataset (MHMD) for automatic image colorization. We randomly select 2,000 images from ImageNet [11] and MHMD datasets to calculate their average values on H channel of HSV. MHMD dataset focus on blue and red yellow of color area and ImageNet dataset is distributed in all color areas, especially in yellow and blue areas. This shows that MHMD dataset considers the old content and information of real gray historical images, which are mostly dark blue and red yellow. However, the distribution of ImageNet dataset is more balanced in H value. Therefore, our method can perform well in colorizing old images. Finally, the contributions of this work include:

- (1) Semantic parsing subnetwork can accurately obtain the semantic information of various parts of person, which makes the image colorization boundary and the human semantic parsing more accurate.
- (2) In HistoryNet, we have designed classifier and classification subnetworks. Info information of classifier subnetwork and classification labels jointly achieve HistoryNet coloring accuracy and classification labels more precise.
- (3) We propose a dataset of persons colorization which is called MHMD: Modern Historical Movies Dataset. MHMD contains 1,353,166 images and the 42 labels of eras, nationalities, and garment types, which have great influences on the colorization of old images.

## 2 RELATED WORK

**Reference Image-based Methods** Manga also called Japanese comic is popular all over the world. Most of them are monochrome. Therefore, many manga colorization methods appeared which are based on a reference image of sketch and line art. [31] propagate the colors of the reference manga to the target manga by representing images as graphs and matching graphs. [41] integrate residual U-net to apply the style to the grayscale sketch with auxiliary classifier generative adversarial network (AC-GAN) [29]. [10] propose a novel deep conditional adversarial architecture for scribble based anime line art colorization. [17] propose a manga colorization method based on conditional Generative Adversarial Networks (cGAN) and require only a single colored reference image for training. [13] colorize a whole page (not a single panel) semiautomatically, with the same color for the same character across multiple panels. In addition, [38, 39] focus on the texture and luminance information of the reference image to achieve colorization. [19, 34, 43] are other methods based on reference image. [19] colorize images from pixel level. [34] propose a dual conditional generative adversarial network which considers contour and color style of images.

**Semantics-based Methods** Researchers have propose many semantic-based methods to deal with the colorization problem of grayscale images. [3, 6, 25, 26] colorize gray images based on semantics of input text and language description. [48, 49] both propose a method based on scene sketches and semantic segmentation. [33, 44] learn object-level semantics to guide image colorization. [44] propose to exploit pixelated object semantic to guide image colorization, which also consider the semantic categories of objects. [12] achieve auto image colorization by learning from examples. [22] propose a Tag2Pix GAN architecture which takes a grayscale line art and color tag information as input to produce a quality colored image.

**Deep Learning-based Methods** Cheng et al. [9] first propose a neural network method for automatic colorization. With the development of CNN, [18, 23, 27] use CNN to extract information of images. In the work of Iizuka et al. [18], the colorization network can obtain both local and global features of the image. In the research of Mouzon et al. [27], the distribution statistical method is combined with the variational method to calculate the possible color probability for each pixel of the image. [23] train a VGG to learn the color histogram of each pixel.

Many existing GAN have the ability to learn the probability distribution of high dimensional spatial data, which can be applied to colorization tasks. The method in [20] use conditional GAN to map the input grayscale image to the output colorized image. Nazeri et al. [28] attempt to fully generalize the colorization procedure using a conditional DCGAN. [5] leverage the conditional GAN to automatically obtain a variety of possible colorization results through multiple sampling of the input noise. [36] propose an adversarial learning colorization approach coupled with semantic information. [40] present a novel memory-augmented colorization model Memo-Painter that can produce high-quality colorization with limited data.

### 3 ARCHITECTURE AND TRAINING LOSSES

According to the description of the HistoryNet network structure, we define the total loss function as:

$$L(G_\theta, D_w) = L_r(G_{\theta_1}^1) + \lambda_{cls} L_{cls}(G_{\theta_2}^2) \quad (1)$$

$$+ \lambda_{par} L_{par}(G_{\theta_3}^3) + \lambda_g L_g(G_{\theta_1}^1, D_w) + \lambda_{info} L_{info}(G_{\theta_1}^1)$$

The first three terms in the formula are the loss values of the generator, we denote them by  $G_\theta = (G_{\theta_1}^1, G_{\theta_2}^2, G_{\theta_3}^3)$ , where  $\theta = (\theta_1, \theta_2, \theta_3)$  stand for all the generator parameters. The last two terms are the loss values of the discriminator which we denote by  $D_w$ .

**Generator Network** The generator network mainly generates  $(a, b)$  channel images. We define its loss function as:

$$L_r(G_{\theta_1}^1) = E_{(L, a_r, b_r) \sim P_r} \left[ \|G_{\theta_1}^1(L) - (a_r, b_r)\|_2^2 \right] \quad (2)$$

Where  $(L, a_r, b_r)$  is the representation of the color image in the CIE  $L_{ab}$  color space, and  $P_r$  is the distribution of the color image.  $\|\cdot\|_2$  is the Euclidean distance. By calculating the Euclidean distance between  $G_{\theta_1}^1(L)$  and  $(a_r, b_r)$ , the resulting image can better perceive the color difference with the real image in the  $L_{ab}$  color

space.  $L_2$  loss has enough ability to constrain the network to obtain more realistic colorization results.

**Parsing Network** While many methods consider the semantic segmentation for colorization, they are not suitable for the historical person's diverse clothing colorization, thus we propose fine grained semantic understanding for this issue. Under the guidance of human parsing as ground truth, the parsing feature of the image is obtained by continuously up-sampling and based on U-net [30], we concatenate the up-sampling information of parsing network with the generator network (the blue  $G_1$ ). Through this way, the generator network can obtain the fine grained semantic parsing information of images. For example, the person's face, hat, hands and clothes are separated to ensure the colorization boundary is clear and accurate. As shown in Figure 6. The parsing network loss we defined is:

$$L_{par}(G_{\theta_3}^3) = E_{(S, a_r, b_r) \sim P_r} \left[ \|G_{\theta_3}^3(S) - (a_r, b_r)\|_2^2 \right] \quad (3)$$

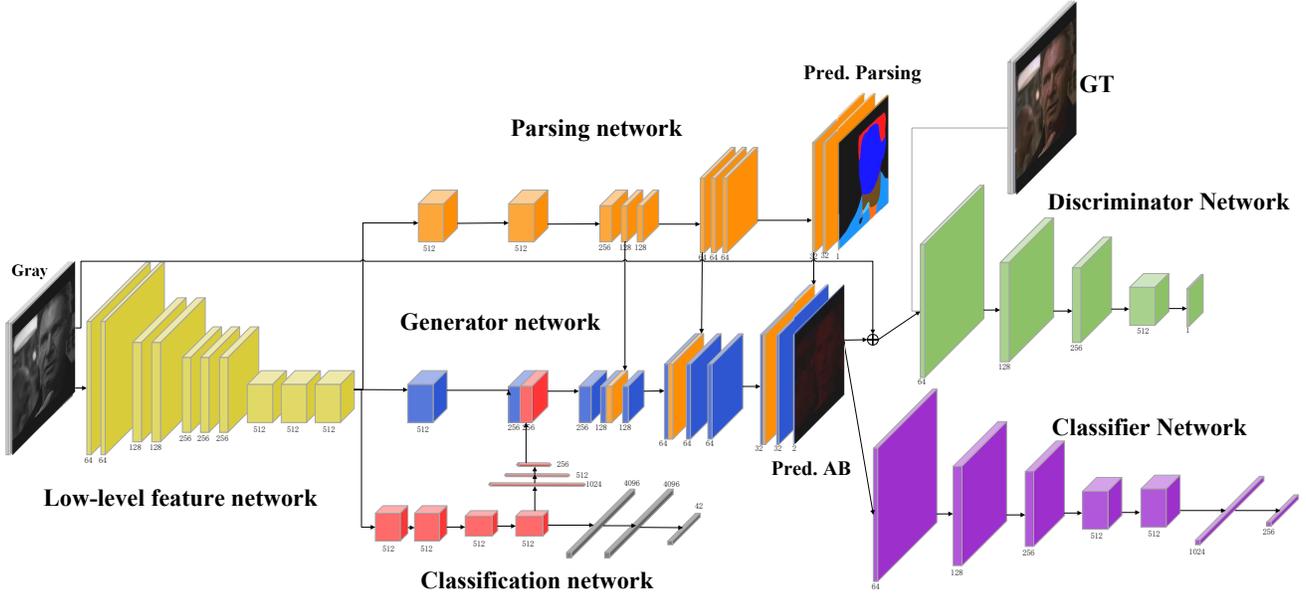
In  $(S, a_r, b_r)$ ,  $S$  is the image before parsing, and  $(a_r, b_r)$  is the parsing image. Calculating the Euclidean distance between the image generated by the parsing network and the actual parsing image and minimize it. In this way, the image input to the parsing network is closer to the real image and finally the colorization result is more accurate.

**Classification network** As shown in red in Figure 2, classification network is designed to obtain the high-level features of the image and the category label information of the image colorization. The convolution of four modules are then divided into two parts of full connection layers. The gray fully connected layers on the right side obtain a 42-dimensional vector that represents the 42 classification labels in the dataset we developed. 42 labels consider the eras, nationalities and garment types of historical person's images. The detailed labels of classification improve the accuracy of image colorization such as the navy clothing is mainly white. The second part (as shown in the red part) is concatenated together with the blue  $G_1$  convolution layer above. This part can obtain 256-dimensional vector which is corresponding to classifier network to constrain the loss for correct color classification of images. The loss function is defined as:

$$L_{cls}(G_{\theta_2}^2) = E_{L \sim P_{rg}} \left[ KL(y_v \| G_{\theta_2}^2(L)) \right] \quad (4)$$

Where  $P_{rg}$  is the distribution of the input grayscale image.  $y_v \in R^m$  is the classification vector obtained by the VGG network classifying the images in the dataset, and  $m$  is the number of image classifications.  $KL$  divergence calculate the loss due to  $y_v$  fitting  $G_{\theta_2}^2(L)$ . The input grayscale image is processed by the classification network so that the coloring network can choose color more correctly.

**Classifier Network** In general colorization methods, the loss of colorization is required to be minimized as much as possible, so that the colorization result is closer to the real image. For example, when we want to colorize a certain object blue sky, no matter how many images are trained, the final average color value is blue so that the colorization image is still blue of sky. However, due to the nationality variations, garment types and era differences, these factors will hinder the calculation of color loss value of historical person's



**Figure 2: HistoryNet network structure diagram.** Overview of our model, it combines a low-level feature network  $G_0$  (in yellow), a Generator network  $G_1$  (include yellow, blue, red and orange layers), a classification network  $G_2$  (in red), a parsing network  $G_3$  (in orange) and a Discriminator network. The discriminator network is divided into two parts which are  $D_1$  (in green) and  $D_2$  (in purple).

images especially military uniform. According to the traditional method of calculating the minimum loss between the colorization image and the real image, the color of the final colorization image will tend to be gray after averaging. So, classifier network adopt InfoGAN [8], which can make good use of the information of image latent code to achieve better performance of images. After obtaining the  $(a, b)$  channel of image through the generator, we input it into the classifier network, and finally obtain a 256-dimensional vector through the fully connected layers. The  $KL$  divergence between the 256-dimensional vector obtained in classification network  $G_2$  is expected to minimize the fitting loss so that the generated image can better present the classification information.

During this way, final colorization images can have correct color for each part of garment. In addition, the interior coloring of each part of the garment is continuous. As shown in Figure 7. The loss we designed is:

$$L_{info} \left( G_{\theta_1}^1 \right) = E_{L \sim P_{rg}} \left[ KL \left( infoGAN || G_{\theta_1}^1 (L) \right) \right] \quad (5)$$

**Discriminator Network** As shown in Figure 2,  $D_1$  is based on the Markov discriminator architecture. The PatchGAN [20] discriminator can track and capture the high-frequency structure of the generated image, thereby making up for the high-frequency information loss caused by the use of  $L_2$  loss in  $G_1$ . For this reason, we define each patch as true or fake.

In this paper, based on WGAN [2], we design the loss of discriminator. WGAN [2] uses the Earth-Mover distance to minimize the possible and true distribution of the generator. Using this feature of the WGAN [2] network can avoid gradient disappearance and

mode collapse during the training process and eventually achieve stable training and obtain better-colored images. Also, we use Kantorovich - Rubinstein duality [21, 35] and add the gradient penalty term to constrain the  $L_2$  norm of the discriminator relative to its input, thus defining  $D_w \in D$ , where  $D$  denotes the set of 1-Lipschitz functions.

$$L_g \left( G_{\theta_1}^1, D_w \right) = E_{\tilde{I} \sim P_r} \left[ D_w \left( \tilde{I} \right) \right] - E_{(a,b) \sim P_{G_{\theta_1}^1}} \left[ D_w \left( L, a, b \right) \right] - E_{\hat{I} \sim P_{\hat{I}}} \left[ \left( \left\| \nabla_{\tilde{I}} D_w \left( \hat{I} \right) \right\|_2 - 1 \right)^2 \right] \quad (6)$$

Where  $P_{G_{\theta_1}^1}$  is the model distribution in  $G_{\theta_1}^1 (L)$  of  $L \sim P_{rg}$ . As in [15],  $P_{\hat{I}}$  is a straight-line uniform sampling which is between the pairs of points sampled along the data distribution  $P_r$  and the generator distribution  $P_{G_{\theta_1}^1}$ .

According to the formula (1) (total loss function) proposed above, we train the networks  $G$  and  $D$  by calculating the following expressions. During the training process, We take the hyperparameter values  $\lambda_{cls}, \lambda_{par}, \lambda_g, \lambda_{info}$  as: 0.003, 0.003, 0.1, 0.003.

$$\min_{G_{\theta}} \max_{D_w \in D} L \left( G_{\theta}, D_w \right) \quad (7)$$

## 4 DATASETS

The colorization of historical person's clothing is of practical significance, especially the colorization of historical military uniform. At present, most of the existing modern color image datasets contain

modern objects or scenes but lack old content or information of real gray historical images, especially the colors of the garments of the historical persons. Therefore, we build a dataset called MHMD: Modern Historical Movies Dataset. MHMD can meet the requirements of garment types, eras and nationalities. First of all, we search for 147 historical movies and TV series in modern time. Second, after preprocessing, the MHMD dataset obtains 1,353,166 images, including 1.2M images on training dataset and 100,001 images on testing dataset. We classify the images into 42 labels according to above types, as shown in Figure 3. The 42 labels are divided by us to have solved the problem of different military uniforms of different countries in different periods. Detailed classification of labels is helpful to obtain accurate information for network structure. On this basis, we design the HistoryNet network structure to realize the colorization of historical military uniforms and the restoration of old historical photos. More details about MHMD can be found on <https://github.com/BestiVictory/MHMD>.

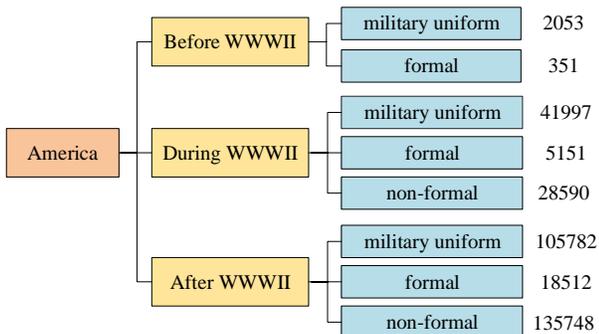


Figure 3: The picture shows the U.S. military uniform, formal and non formal labels and the corresponding number of pictures before, during and after World War II in MHMD.

#### 4.1 Collecting Methods

We present MHMD of approximately 1.3M images with image-level labels and localization instructions. We mainly download TV series, movies and documentaries from various websites. Because we are focusing on the colorization of historical person’s images and in order to ensure the quality of the films (such as definition, etc.), we have formulated the film selection criteria: color films taken after the 1990s, the content of the film is mainly about history, war, person and objects of various countries before 1990.

#### 4.2 Processing Methods

First, we cut the film into images, and use the program to remove black and white images, images with too low pixel blur. For the goal of focusing on persons, we use Yolov3 to judge whether there is a person in the image and delete the images without people or too many people. Through these preprocessing, the diversity of the content and the clarity of the image are guaranteed. Secondly, through an image, we can obtain the eras, nationalities and garment types information of it. According to this, we have designed 42 kinds

of labels and divide them into three categories: era, nationality and garment type. The era label is divided into: before, during and after World War II. Nationality labels are: China (divided into the Communist Party and the Kuomintang), Japan, the United States, Germany, Britain and Russia. There are three types of garment type: military, formal (such as suits) and informal. Thirdly, We randomly extract 1% of the images from the dataset, which is the same distribution as the original dataset, and then we label the 1% images manually. The 1% images and data augmentation of these images (such as horizontal image mirroring and Gaussian blur) will be inputted to ResNet [16] for training. After the training, we input other images in the dataset into ResNet [16] for automatic classification to obtain labels. The accuracy of classification can reach 98% or more. The categories of MHMD and the comparison with other colorization datasets are shown in table 1.

## 5 EXPERIMENTS

In this section, we evaluate our methods quantitatively and qualitatively on MHMD we designed. We compare our experimental results with, Deoldify [1], Iizuka et al. [18], Larsson et al. [23], ChromaGAN [37], Su et al. [33], and use LPIPS [42], PSNR and SSIM indicators to quantitatively compare our experimental results with those of the most advanced methods. Finally, we carry out ablation experiments on parsing network and classifier network, and prove that they have positive effects on the improvement of network performance.

### 5.1 Implementation Details

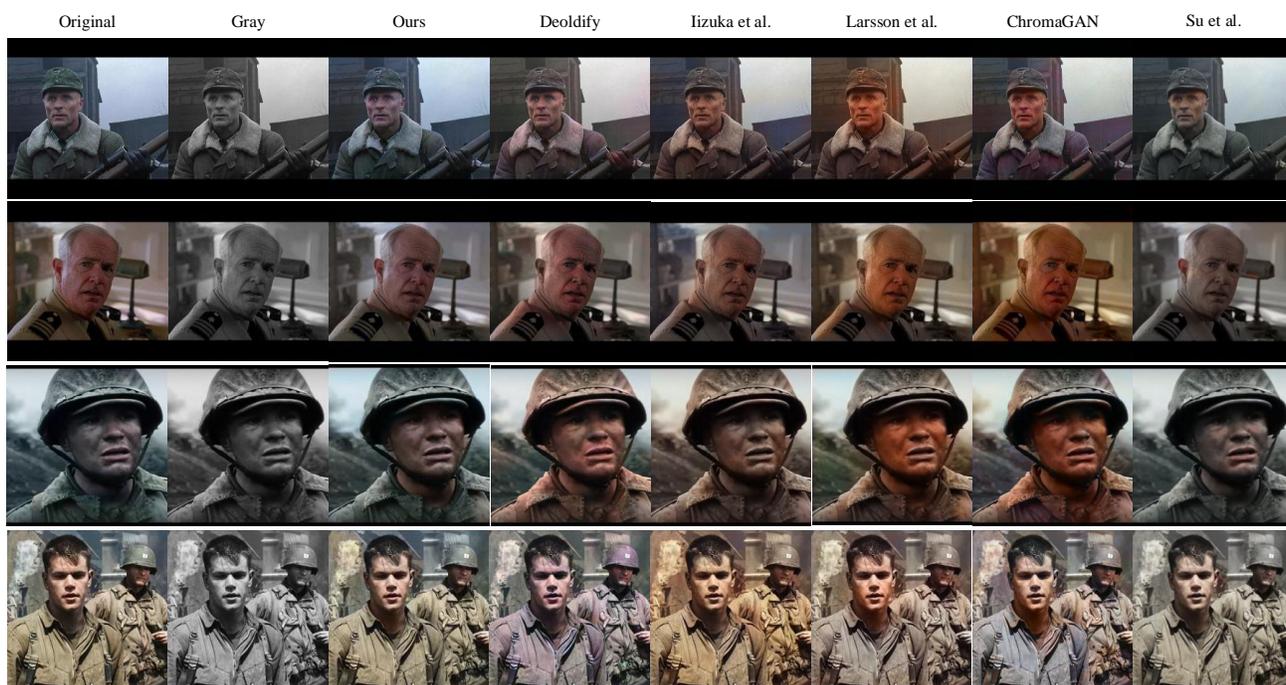
We train HistoryNet for a total of eight epochs and set the batch size to 16, on the 1.3M images from MHMD resized to 224×224. A single epoch takes approximately 28 hours on a Nvidia titan X pascal GPU. We minimize our objective loss using Adam optimizer with learning rate equal to  $2 \times 10^{-5}$  and momentum parameters  $\beta_1 = 0.5$  and  $\beta_2 = 0.999$ . We alternate the optimization of the generator  $G_0$  and discriminator  $D_1$ . The first stage of the network (displayed in yellow in Figure 2), takes as input a grayscale image of size 224 × 224, and is initialized using the pre-trained weights of the VGG-16 [32] trained on ImageNet.

### 5.2 Quantitative Comparisons

We use three metrics, LPIPS [42], PSNR and SSIM. LPIPS [42] is a measurement method of perceptual similarity proposed by Zhang et al. that can achieve good perceptual judgment in challenging models of visual prediction and other tasks. The lower the LPIPS [42] value, the smaller the perceived difference between the image and the original image. PSNR and SSIM are two commonly used indexes in image quality evaluation. PSNR is peak signal to noise ratio, which is usually used as an evaluation index of the quality of an image before and after compression. The higher the PSNR value is, the better the quality of the generated image will be. SSIM is structural similarity index, which is an index to measure the similarity between two images. Table 2 shows the comparison of our experimental results with Deoldify [1], Iizuka et al. [18], Larsson et al. [23], ChromaGAN [37], Su et al. [33]. In the above methods, We have successfully applied the network structures of ChromaGAN

**Table 1: Comparison of MHMD with other colorization datasets**

Dataset	Scene	The labels of the dataset				Total		
		Era	Nationality	Garment Type				
MHMD	×	Before WWII	66,900	Chinese	753,473	Military	707,771	1,353,166
				American	934,415			
		During WWII	547,318	Russian	45,291	Formal	104,763	
				German	59,015			
		After WWII	738,948	Japanese	110,562	Informal	540,632	
				English	46,641			
ImageNet [11]	✓	×	×	×	×	about 1,300,000		
COCO-Stuff [4]	✓	×	×	×	×	about 164,000		
Places205 [45]	✓	×	×	×	×	20,500		



**Figure 4: Contrastive experimental diagram. Some qualitative results, from left to right: Original, Gray, Ours, Deoldify [1], Iizuka et al. [18], Larsson et al. [23], ChromaGAN [37], Su et al. [33]. The results are comparable.**

[37] and Su et al. [33] to MHMD for training, and the training details are shown in the supplementary.

As can be seen from table 2, our method have better perform in these metrics. The lower LPIPS value indicates that our results are more similar to the source image. The higher PSNR and SSIM values means that the quality of the generated image is better and is similar to the original image.

### 5.3 Ablation Experiments

We have trained ChromaGAN on MHMD dataset as a baseline. As can be seen from table 3, LPIPS, PSNR and SSIM perform better after the parsing and classifier networks are added. This is enough

**Table 2: Quantitative comparison of experimental**

Method	LPIPS↓	PSNR↑	SSIM↑
Iizuka et al. [18]	0.134	25.779	0.956
Larsson et al. [23]	0.147	24.527	0.946
Deoldify [1]	0.127	26.321	0.957
ChromaGAN [36]	0.118	29.487	0.951
Su et al. [33]	0.132	25.951	0.941
HistoryNet	<b>0.101</b>	<b>30.638</b>	<b>0.962</b>

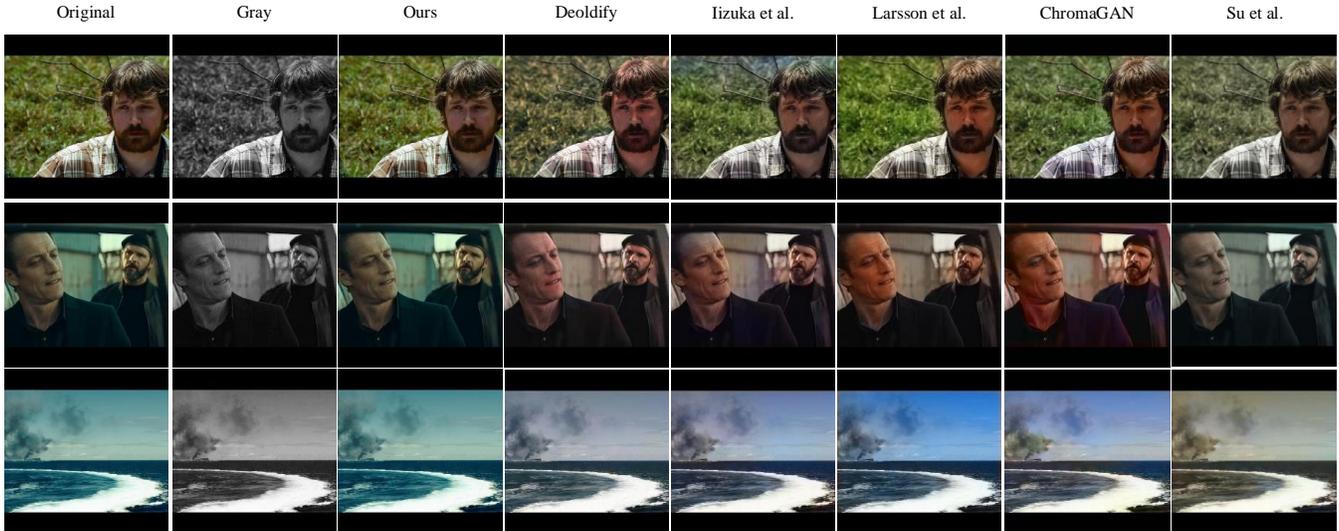


Figure 5: The colorization comparison of natural scenery and person on HistoryNet. Our method also works well in natural scenes.

to show that the parsing network plays a positive role in image colorization and classifier network can be better used in image colorization network after obtaining the information of image latent layer. It can be seen from the last row of table 3 that our proposed method can perform well on all three parameters, which fully shows that our method has a positive effect on colorization. We also perform three ablation tests: Baseline + Parsing, Baseline + Classifier and our method.

Table 3: Ablation experiments

Method	LPIPS↓	PSNR↑	SSIM↑
ChromaGAN	0.123	27.093	0.946
Baseline+Parsing	0.121	28.992	0.948
Baseline+Classifier	0.119	29.828	0.951
HistoryNet	<b>0.107</b>	<b>30.585</b>	<b>0.959</b>

#### 5.4 Qualitative comparisons

**Effect of fine grained semantic parsing** There exist kinds of image segmentation methods in the literatures [7, 14, 24, 46, 47]. In this paper, we use Human Parsing [14] and Deeplab-v3 [7] separately for semantic segmentation. Deeplab-v3 [7] only separate persons and background, but can't focus on details of persons' each part, such as face and hands. Therefore, we adopt instance-level human parsing [14] for semantic segmentation, which is focus on recognizing each semantic part for example arms and hair. Due to the limit of equipment and time, we separately adjust 75,837 images of human parsing and Deeplab-v3 [7] manually and use them as the ground truth for parsing network. Through training on HistoryNet, it can be seen in Figure 6 that human parsing focus on fine grained semantic parsing so its segmentation accuracy is better than semantic segmentation. Because of this, the result of

using human parsing as ground truth training is better than that of Deeplab-v3. For example, in the third line of the Figure 6, the person's hands are separated by Human Parsing that the right hand of the person is light green, the left one is light blue. By taking human parsing as the ground truth of parsing network, parsing network can accurately guide the acquisition and colorization of color information of each part. Therefore, the boundary of Parsing Result is more clear and the colorization effect is closer to Original images. However, Semantic Result can not achieve accurate segmentation, which leads to the color fusion of hands and background and presents the same green of the background.



Figure 6: The picture shows that the segmentation of human parsing is more accurate than semantic segmentation.

**Ablation experiments** In HistoryNet, there are two main sub-modules: parsing and classifier subnetworks, so we do the corresponding ablation experiments. As shown in Figure 7, baseline consists of  $G_0$ ,  $G_1$ ,  $G_2$  and  $D_1$  in Figure 2 of HistoryNet structure. From the Figure 7, we can see that parsing can better segment the

boundary, so as to make the colorization more accurate. For example, in the first column, Baseline+Parsing can accurately separate the two clothes and the boundary of the clothes is more precise, while Baseline+Classifier does not perform due to the lack of fine grained semantic parsing of person. Although the parsing results have better boundary of person, the color of the images is not correct and nature, such as the third column in the figure. Compared with Baseline+Classifier, Baseline+Parsing is not classified, so the color of the neck is not natural.



**Figure 7: Some qualitative comparisons of ablation experiments. Parsing submodule can obtain the clear boundary through fine grained semantic parsing; Classifier submodule can help choose correct color for colorization.**

**Results on old photos** We design MHMD suitable for colorizing historical images, especially military uniforms and suits of historical images. In addition, our method can restore old photos, remove the background noise and colorize the yellowing background to the normal color. As shown in Figure 8. We also colorize some legacy black and white images. As shown in the Figure 9, we can see our method is still applicable and has good effect for black and white photos.



**Figure 8: HistoryNet can correct the color and finally achieve a better colorization result. For example, this historical photo appears yellow and green as a whole.**

**Comparisons with state-of-the-art** Figure 4 shows the experimental results of the five best current methods. It can be seen



**Figure 9: HistoryNet also apply for some legacy black and white images.**

from the Figure 4 that our method has a good performance on the accuracy of the image color and boundary of each part, while other methods have problems such as inaccurate colorizing of the images and discontinuous lumpiness of the color on the clothes. From the first and fourth line of Figure 4, we can see that each element (such as clothes, face etc.) in our colorization results have a clear boundary and achieve the consistency within the region block, that is, the clothes part is of the same color, From this comparative experiment, we can see the advantages of designing parsing and classifier submodules in HistoryNet. Parsing subnetwork can solve the problem of boundary segmentation and classifier subnetwork can achieve the consistency and continuity of the overall coloring of the persons clothing.

In addition, our method is not only suitable for the colorization of military uniform, but also for the colorization of other natural landscapes and characters, as shown in the Figure 5.

## 6 CONCLUSIONS

In this paper, we propose a new HistoryNet architecture, which contains parsing, classification and classifier subnetworks. Semantic parsing subnetwork can help the colorization boundary more accurate. Classifier subnetwork can help choose correct color. In addition, we propose a dataset called MHMD that focus on the real gray historical images. To the best of our knowledge, the proposed MHWD dataset is the largest dataset of historical image colorization. The MHWD can be accessed by request. Through relevant qualitative and quantitative comparison, our method is superior to the state-of-the-art colorization network in LPIPS, PSNR, SSIM. Another purpose of our work is to inspire more researchers to make the gray image colorization technologies more useful in historical image/video colorization. In the future, related research methods will also be used in video coloring.

## ACKNOWLEDGEMENTS

We thank the ACs, reviewers, and the annotators of our dataset such as Jisen Huang and Xuntao Zhou. This work is partially supported by the National Natural Science Foundation of China (62072014), the Beijing Natural Science Foundation (L192040), and the Advanced Discipline Construction Project of Beijing Universities (20210041Z0401).

## REFERENCES

- [1] Jason Antic. 2019. DeOldify. <https://github.com/jantic/DeOldify>.
- [2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875* (2017).
- [3] Hyojin Bahng, Seungjoo Yoo, Wonwoong Cho, David Keetae Park, Ziming Wu, Xiaojuan Ma, and Jaegul Choo. 2018. Coloring with words: Guiding image colorization through text-based palette generation. In *Proceedings of the european conference on computer vision (eccv)*, 431–447.
- [4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1209–1218.
- [5] Yun Cao, Zhiming Zhou, Weinan Zhang, and Yong Yu. 2017. Unsupervised diverse colorization via generative adversarial networks. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 151–166.
- [6] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. 2018. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8721–8729.
- [7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.
- [8] X. Chen, Y. Duan, R. Houthoof, J. Schulman, I. Sutskever, and P. Abbeel. 2016. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Neural Information Processing Systems (NIPS)*.
- [9] Zezhou Cheng, Qingxiang Yang, and Bin Sheng. 2015. Deep colorization. In *Proceedings of the IEEE International Conference on Computer Vision*. 415–423.
- [10] Yuanzheng Ci, Xinzhu Ma, Zhihui Wang, Haojie Li, and Zhongxuan Luo. 2018. User-guided deep anime line art colorization with conditional adversarial networks. In *Proceedings of the 26th ACM international conference on Multimedia*. 1536–1544.
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [12] Aditya Deshpande, Jason Rock, and David Forsyth. 2015. Learning large-scale automatic image colorization. In *Proceedings of the IEEE International Conference on Computer Vision*. 567–575.
- [13] Chie Furusawa, Kazuyuki Hiroshiba, Keisuke Ogaki, and Yuri Odagiri. 2017. Comicolorization: semi-automatic manga colorization. In *SIGGRAPH Asia 2017 Technical Briefs*. 1–4.
- [14] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. 2018. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 770–785.
- [15] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Advances in neural information processing systems*. 5767–5777.
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [17] Paulina Hensman and Kiyoharu Aizawa. 2017. cGAN-based manga colorization using a single training image. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 3. IEEE, 72–77.
- [18] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2016. 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 (2016), 1–11.
- [19] Revital Ironi, Daniel Cohen-Or, and Dani Lischinski. 2005. Colorization by Example.. In *Rendering Techniques*. Citeseer, 201–210.
- [20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.
- [21] Leonid Kantorovitch. 1958. On the translocation of masses. *Management Science* 5, 1 (1958), 1–4.
- [22] Hyunsu Kim, Ho Young Jho, Eunhyeok Park, and Sungjoo Yoo. 2019. Tag2pix: Line art colorization using text tag with secat and changing loss. In *Proceedings of the IEEE International Conference on Computer Vision*. 9056–9065.
- [23] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. 2016. Learning representations for automatic colorization. In *European Conference on Computer Vision*. Springer, 577–593.
- [24] Qing Li, Xiaowu Chen, Yafei Song, Yu Zhang, Xin Jin, and Qiping Zhao. 2014. Geodesic Propagation for Semantic Labeling. *IEEE Trans. Image Process.* 23, 11 (2014), 4812–4825. <https://doi.org/10.1109/TIP.2014.2358193>
- [25] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. 2017. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 1271–1280.
- [26] Varun Manjunatha, Mohit Iyyer, Jordan Boyd-Graber, and Larry Davis. 2018. Learning to color from language. *arXiv preprint arXiv:1804.06026* (2018).
- [27] Thomas Mouzon, Fabien Pierre, and Marie-Odile Berger. 2019. Joint CNN and Variational Model for Fully-automatic Image Colorization. In *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer, 535–546.
- [28] Kamyar Nazeri, Eric Ng, and Mehran Ebrahimi. 2018. Image colorization using generative adversarial networks. In *International conference on articulated motion and deformable objects*. Springer, 85–94.
- [29] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional Image Synthesis with Auxiliary Classifier GANs (*Proceedings of Machine Learning Research, Vol. 70*), Doina Precup and Yee Whye Teh (Eds.). PMLR, International Convention Centre, Sydney, Australia, 2642–2651. <http://proceedings.mlr.press/v70/odena17a.html>
- [30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.
- [31] Kazuhiro Sato, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2014. Reference-based manga colorization by graph correspondence using quadratic programming. In *SIGGRAPH Asia 2014 Technical Briefs*. 1–4.
- [32] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [33] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. 2020. Instance-aware Image Colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7968–7977.
- [34] Tsai-Ho Sun, Chien-Hsun Lai, Sai-Keung Wong, and Yu-Shuen Wang. 2019. Adversarial Colorization Of Icons Based On Structure And Color Conditions. *arXiv preprint arXiv:1910.05253* (2019).
- [35] Cédric Villani. 2008. *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media.
- [36] Patricia Vitoria, Lara Raad, and Coloma Ballester. 2020. ChromaGAN: Adversarial Picture Colorization with Semantic Class Distribution. In *The IEEE Winter Conference on Applications of Computer Vision*. 2445–2454.
- [37] Patricia Vitoria, Lara Raad, and Coloma Ballester. 2020. ChromaGAN: Adversarial Picture Colorization with Semantic Class Distribution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
- [38] Tomihisa Welsh, Michael Ashikhmin, and Klaus Mueller. 2002. Transferring color to greyscale images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*. 277–280.
- [39] Wengqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. 2018. Texturegan: Controlling deep image synthesis with texture patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8456–8465.
- [40] Seungjoo Yoo, Hyojin Bahng, Sunghyo Chung, Junsoo Lee, Jaehyuk Chang, and Jaegul Choo. 2019. Coloring with limited data: Few-shot colorization via memory augmented networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11283–11292.
- [41] Lvmin Zhang, Yi Ji, Xin Lin, and Chunping Liu. 2017. Style transfer for anime sketches with enhanced residual u-net and auxiliary classifier gan. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 506–511.
- [42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.
- [43] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. 2017. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999* (2017).
- [44] Jiaojiao Zhao, Jungong Han, Ling Shao, and Cees GM Snoek. 2019. Pixelated Semantic Colorization. *International Journal of Computer Vision* (2019), 1–17.
- [45] Bolei Zhou, Agata Lapedriza, Jianxiang Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.
- [46] Quan Zhou, Yu Wang, Yawen Fan, Xiaofu Wu, Suofei Zhang, Bin Kang, and Longin Jan Latecki. 2020. AGLNet: Towards real-time semantic segmentation of self-driving images via attention-guided lightweight network. *Appl. Soft Comput.* 96 (2020), 106682. <https://doi.org/10.1016/j.asoc.2020.106682>
- [47] Quan Zhou, Yu Wang, Jia Liu, Xin Jin, and Longin Jan Latecki. 2019. An open-source project for real-time image semantic segmentation. *Sci. China Inf. Sci.* 62, 12 (2019), 227101. <https://doi.org/10.1007/s11432-019-2685-1>
- [48] Changqing Zou, Haoran Mo, Chengying Gao, Ruofei Du, and Hongbo Fu. 2019. Language-based colorization of scene sketches. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–16.
- [49] Changqing Zou, Qian Yu, Ruofei Du, Haoran Mo, Yi-Zhe Song, Tao Xiang, Chengying Gao, Baoquan Chen, and Hao Zhang. 2018. Sketchyscene: Richly-annotated scene sketches. In *European Conf. on Comp. Vis. (ECCV)*. 421–436.