Supplementary of "PCL"

Yuanyuan Liu[1*], Wenbin Wang[1*], Yibing Zhan[2], Shaoze Feng[1], Kejun Liu[1], Zhe Chen[3†]
[1]School of Computer Science, China University of Geosciences, Wuhan, China
[2]JD Explore Academy, China
[3]The University of Sydney, Australia
{liuyy, wangwenbin, fengshaoze, liukejun}@cug.edu.cn; zhanyibing@jd.com; zhe.chen1@sydney.edu.au

The content of our supplementary material is organized as follows:

- Effect of Dynamic Weight Average in our framework.
- Compare with different backbones in our framework.
- Effect of different pose extractors
- Transfer learning
- Visualization of different features.
- Visualization of reconstructed faces with different features.
- Visualization of self-supervision results on different downstream tasks.
- Detailed description of the backbone B, two subnets and reconstruction network D.

1. Additional Ablation Experiments

1.1. Effect of Dynamic Weight Average

In order to adaptively balance the learning of L_{pose} and L_{face} in our PCL, we employ the Dynamic Weight Average (DWA) [7] to obtain the α_{pose} and α_{face} during the multi-task training. The dynamic weights α_{pose} and α_{face} can be calculated as:

$$\alpha_k(t) = \frac{Kexp(\omega_k(t-1)/T)}{\sum_i exp(\omega_i(t-1)/T)}, \ \omega_k(t-1) = \frac{L_k(t-1)}{L_k(t-2)}, \quad (1)$$

where the $\alpha_k(t)$ represents the weight α_{pose} of pose-related contrastive learning or the weight α_{face} of face contrastive learning in Eq.(7) of our paper at the $t-th$ epoch. T represents a temeprature which controls the softness of task weighting, and K represents the number of tasks (K=2 in this study). Through properly training with DWA, our PCL adaptively introduces pose and face information according to the learning objective, thus obtaining the best learning performance. Table 1 shows the comparison of the results of DWA and manual weighting. Notely, in the manual weighting, we constrain $\alpha_{face} = 1 - \alpha_{pose}$, where $\alpha_{pose} = 0$ represents that PCL only contains pose-unrelated face contrastive learning. Obviously, the result demonstrates that our PCL with DWA training achieved the better learning performance than the other learning schemes.

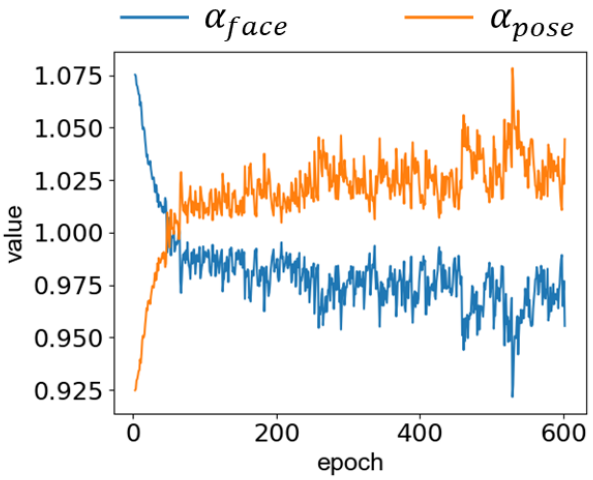

Figure 1. The dynamic learning procedure of the α_{pose} and α_{face} during training.

In addition, Fig 1 presents the learning procedure of the α_{pose} and α_{face} via DWA during training. At the beginning of training, the pose task had the smaller weight due to its fast convergence rate. As training went on, the facial task weight decreased while the pose weight increased, because the DWA attempted to slow down the task that are learned quickly for more balanced learning. When the learning converged, the dynamic weights of the facial expression and pose reached stability, so that promoting both two tasks simultaneously.

*Equally-contributed first authors
†Corresponding author

Table 1. Comparison of DWA and manual weighting in PCL on the RAF-DB dataset.

α_{pose}	0	0.01	0.1	DWA
Acc(%)	73.24	72.52	73.96	**74.47**

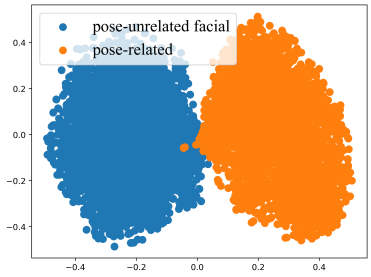

Figure 2. The pose-related and pose-unrelated facial features in 2D space by t-SNE visualization

Table 2. Comparison of different backbones. We conducted FER on the FER2013 dataset, facial recognition on the LFW dataset, and AU detection on the DISFA dataset.

Backbone	Params(M)	FER(Acc.)	Facial Recognition(Acc.)	AU detection(F1)
FaceCycleBackbone [1]	**3.06**	**56.81**	**79.72**	**54.8**
ResNet-18 [4]	11.18	52.3	79.6	40.57
ResNet-34 [4]	21.28	52.1	76.68	37.96

1.2. Comparison of Different Backbones

Table 2 reports model parameters, accuracy (Acc.) and F1 score of three different backbones used in our PCL for different facial tasks. Obviously, the deeper backbone degrades the performance. The possible reason is that the deeper backbone tends to be overfitting during training of the pretext task, resulting in a decrease in generalizability for the downstream linear evaluation tasks.

1.3. Effect of Different Pose Extractors

As shown in Table 3, we used cGAN [8] and an off-the-shelf 2D face alignment method TP-GAN [5] stead of our proposed PDD to generate images with different poses. Our PDD outperformed the cGAN and TP-GAN by 18.97% and 2.76%, respectively. Additionally, we also tried to use vanilla GAN [3], however, the vanilla GAN cannot ensure the generation of the desired pose. The possible reason is that the generator that tries to fool the discriminator may make the distribution of features bias in CL.

Table 3. The effects of different pose extractors.

	cGAN [8]	TP-GAN [5]	PDD
FER(RAF-DB)	55.50	71.71	**74.47**

Table 4 shows the effects of different L_1 loss used in PDD. Obviously, stricter constraints work best. We conjecture that adding a series of loss constraints make the PDD reconstruct more accurate faces, thus obtaining more effective facial representation.

Table 4. Ablation study of the impact of integrating different losses in face transformation of PDD on the RAF-DB dataset.

Different L_1 loss used in PDD	FER Accuracy(%)
$\|\|s - D(\vec{F}_f, \vec{F}_p)\|\|_1$	73.57
$\|\|\hat{s} - D(\vec{F}_f, \vec{F}_p)\|\|_1$	73.50
$\|\|s - D(\vec{F}_f, \vec{F}_p)\|\|_1 + \|\|\hat{s} - D(\vec{F}_f, \vec{F}_p)\|\|_1$	73.37
$\|\|s - D(\vec{F}_f, \vec{F}_p)\|\|_1 + \|\|s - D(\vec{F}_f, \vec{F}_p)\|\|_1$	73.83
$\|\|\hat{s} - D(\vec{F}_f, \vec{F}_p)\|\|_1 + \|\|s - D(\vec{F}_f, \vec{F}_p)\|\|_1$	73.31
$\|\|s - D(\vec{F}_f, \vec{F}_p)\|\|_1 + \|\|\hat{s} - D(\vec{F}_f, \vec{F}_p)\|\|_1 + \|\|s - D(\vec{F}_f, \vec{F}_p)\|\|_1$	74.22
$\|\|s - D(\vec{F}_f, \vec{F}_p)\|\|_1 + \|\|\hat{s} - D(\vec{F}_f, \vec{F}_p)\|\|_1 + \|\|s - D(\vec{F}_f, \vec{F}_p)\|\|_1 + \|\|\hat{s} - D(\vec{F}_f, \vec{F}_p)\|\|_1$	**74.47**

1.4. Transfer Learning

We evaluate transfer learning performance on RAF-DB and LFW with fine-tuning. As shown in Table. 5, our PCL outperforms supervised learning. The result can be demonstrated our method is practical and advantageous.

Table 5. Comparison of transfer learning performance.

Method	FER(RAF-DB)	Face recognition (LFW)
Fully-Supervised	80.41	81.14
FaceCycle	80.48	77.96
SimCLR	82.30	78.20
Ours	**82.43**	**83.09**

2. Visualization

2.1. Visualization of Different Features

Fig. 2 visualized the pose-related features $\vec{F}_p$ and pose-unrelated facial features $\vec{F}_f$ in a 2D feature space by using the t-SNE [9] on RAF-DB, demonstrating that our method can effectively separate pose-related features from pose-unrelated facial features.

In Fig. 3, we visualized the SimCLR [2], the face-aware features $\vec{F}_s$, and pose-unrelated facial features $\vec{F}_f$ in a 2D feature space by using the t-SNE on the RAF-DB dataset, as well as the pose-related features $\vec{F}_p$ on the multi-view BU-3DFE dataset. From the Fig. 3(a), the facial features learned by SimCLR cannot distinguish two categories on the feature sphere space (*e.g.*, Happy and neutral) well. As shown in Fig. 3(b) and Fig. 3(c), the feature distances of different expression categories on $\vec{F}_s$ are larger than those of $\vec{F}_f$ on the feature sphere space for contrast learning, indicating that the face-aware features $\vec{F}_s$ containing pose information has better distinguishability than $\vec{F}_f$ and SimCLR. Fig. 3(d) shows the distribution of the learned pose-related features $\vec{F}_p$, indicating that $g_p(\cdot)$ in our PCL can learn effectively the detailed pose information.

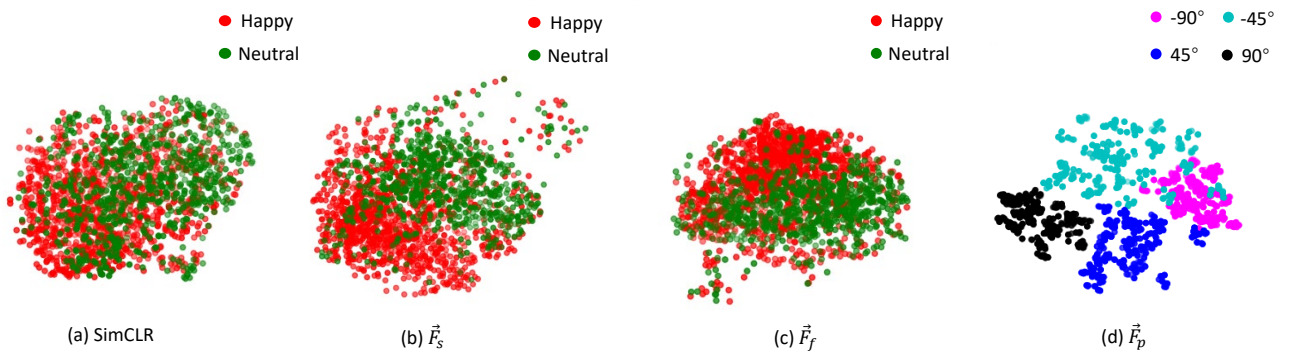

Figure 3. The features learned by SimCLR and our PCL respectively in t-SNE feature visualization. (a) The facial feature learned by SimCLR for FER (RAF-DB). (b) $\vec{F}_s$ extracted from backbone B for FER (RAF-DB). (c) $\vec{F}_f$ extracted from $g_f(\cdot)$ for FER (RAF-DB). (d) $\vec{F}_p$ extracted from $g_p(\cdot)$ for pose estimation (BU-3DFE).

2.2. Visualization of Reconstructed Faces

Fig. 4 visualized several reconstructed faces with pose-related and pose-unrelated facial features disentangled by our method. As shown in Fig. 4(b) and (c), our PCL can reconstructed the same faces with different poses according to varied pose-related features and the same pose-unrelated facial features. Fig. 4(d) shows the reconstructed frontal faces with the pose-unrelated facial features $\vec{F}_f$ from the image s. Additionally, as shown in Fig. 4(e) and (f), we used pose-related features from the image s and its pose-flipped image $\hat{s}$. We can observe that the generated images only include varied pose information with few face patterns.

2.3. Prediction Examples on Different Downstream Tasks

Fig. 5 shows the prediction results of our method on the three downstream tasks, *i.e.*, facial expression recognition, AU detection and facial recognition. From the results, one can see that the PCL can accurately predict the results on different tasks, even for hard samples.

3. Detailed Network Structures

3.1. Backbone and Subnets

Fig. 6 presents the network architecture of the backbone B and its two separating subnets in our PCL. Referring to FaceCycle [1], we adopt a shallow backbone that consists of ten convolutional blocks, two channel attention blocks, and two residual basic blocks. The channel attention module was inspired from self-attention [10], and we just used it to compute relations between channels rather than spatial pixels. The subnet contains four layers, *i.e.*, two 3×3 convolutional layer and two leakyReLU as activation function. The size of input is 4096 and the size of output is 2048. Additionally, we adopt leakyReLU with leakage 0.1 as the activation function in the backbone B and subnet.

3.2. Reconstruction Network

Inspired by [6], we used the 6-layer CNN as our reconstruction network D. Fig. 7 presents the detailed network architecture of the D. D contains five generator blocks and an output block. The generator block consists of ReLU, bilinear upsampling, 3×3 conv, and batch-norm. The only difference between the output block and the generator block is that the last layer of the output block is a tanh function.

References

[1] Jia-Ren Chang, Yong-Sheng Chen, and Wei-Chen Chiu. Learning facial representations from the cycle-consistency of face. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9680–9689, 2021. 2, 3

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan

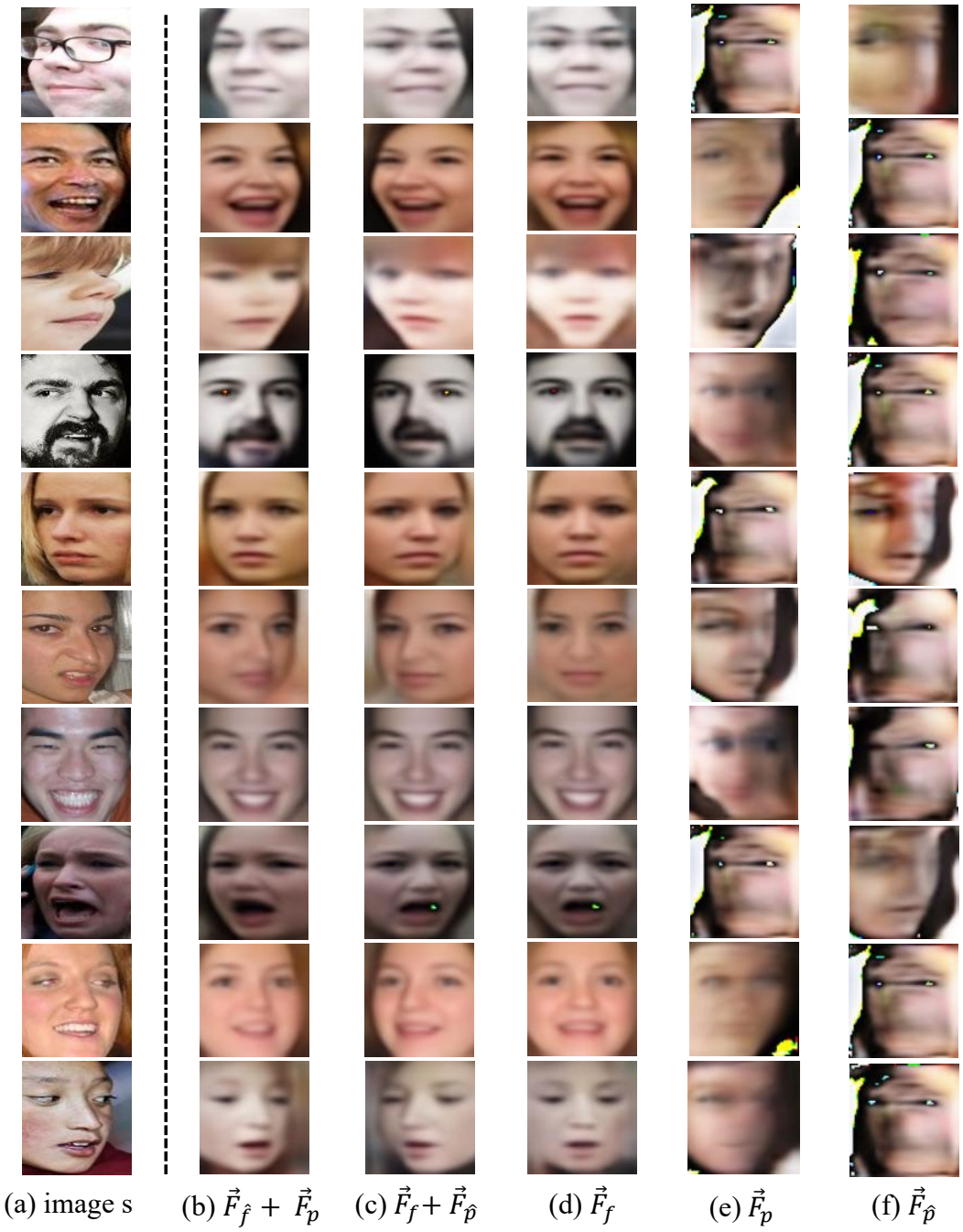

Figure 4. The reconstructed faces with disentangled pose-unrelated facial and pose-related features. (a) Source image s, (b)-(f) the reconstructed faces with different features. $\vec{F}_f$: pose-unrelated facial feature from s, $\vec{F}_p$: pose-related feature from s, $\vec{F}_{\hat{f}}$: pose-unrelated facial feature from pose-flipped $\hat{s}$, $\vec{F}_{\hat{p}}$: pose-related feature from pose-flipped $\hat{s}$.

for photorealistic and identity preserving frontal view synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[6] A Sophia Koepke, Olivia Wiles, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. In *British Machine Vision Conference (BMVC)*, page 302, 2018. 3

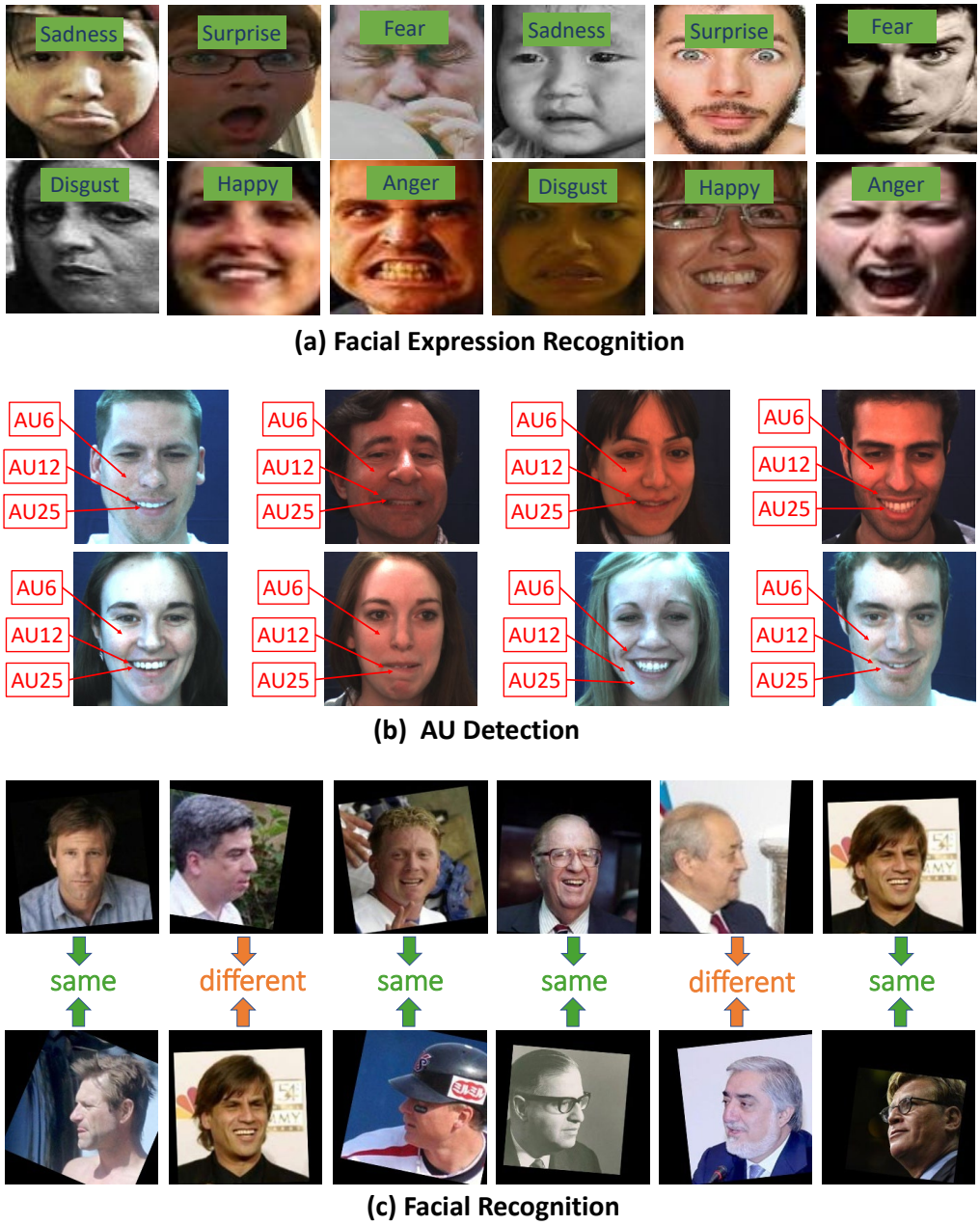

Figure 5. Examples of self-supervision prediction on different downstream tasks. (a) Facial expression recognition task. (b) AU detection task. (c) Facial recognition task.

[7] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019. 1

[8] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[9] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 2

[10] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 3

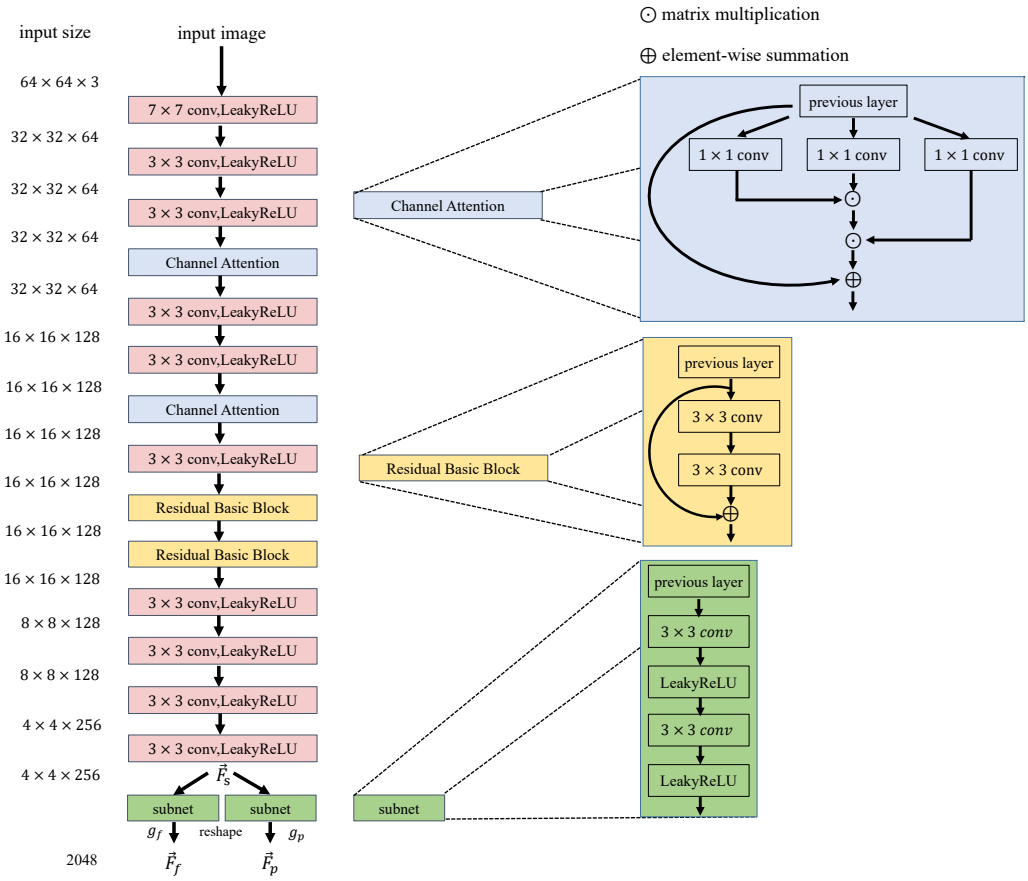

Figure 6. The network architecture of the backbone B and subnets.

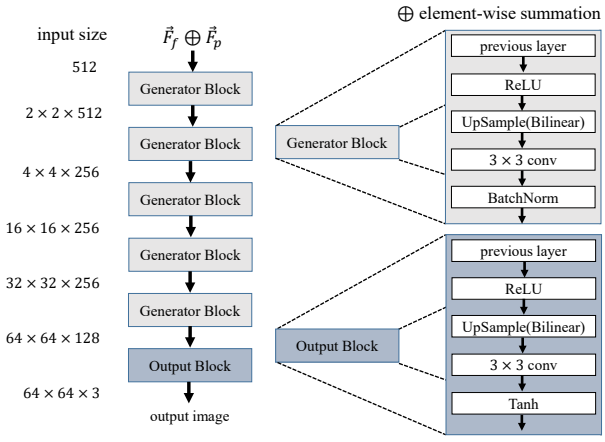

Figure 7. The network architecture of the reconstruction network D.