Semi-Supervised Wide-Angle Portraits Correction by Multi-Scale Transformer

Fushun Zhu [1*] Shan Zhao [2*] Peng Wang [2*] Hao Wang [2] Hua Yan [1†] Shuaicheng Liu [3,2†]
[1] Sichuan University [2] Megvii Technology
[3] University of Electronic Science and Technology of China

Abstract

We propose a semi-supervised network for wide-angle portraits correction. Wide-angle images often suffer from skew and distortion affected by perspective distortion, especially noticeable at the face regions. Previous deep learning based approaches need the ground-truth correction flow maps for training guidance. However, such labels are expensive, which can only be obtained manually. In this work, we design a semi-supervised scheme and build a high-quality unlabeled dataset with rich scenarios, allowing us to simultaneously use labeled and unlabeled data to improve performance. Specifically, our semi-supervised scheme takes advantage of the consistency mechanism, with several novel components such as direction and range consistency (DRC) and regression consistency (RC). Furthermore, different from the existing methods, we propose the Multi-Scale Swin-Unet (MS-Unet) based on the multi-scale swin transformer block (MSTB), which can simultaneously learn short-distance and long-distance information to avoid artifacts. Extensive experiments demonstrate that the proposed method is superior to the state-of-the-art methods and other representative baselines. The source code and dataset are available at https://github.com/megvii-research/Portraits_Correction

1. Introduction

In recent years, a growing number of smartphones have been equipped with wide-angle cameras, which take wide-angle images with rich contents. However, a wider FOV camera often causes severe perspective distortions, which bends straight edges on buildings, and distorts faces, as shown in Fig. 1(a). Therefore, an ideal intelligent algorithm is required to correct the distortion image. After correction, the faces will look more natural while the curved lines in the background are also corrected, as shown in Fig. 1(b).

The traditional undistortion methods apply perspective projection using calibrated camera parameters, which correctly warp the lines at the background to straight [3, 8, 24]. Nevertheless, faces on the image are stretched unnaturally due to incorrect projection as a plane. Compared to perspective projection, the mercator and stereographic projections [28] can preserve the shape of faces locally, but they also bend linear structures in the background [4].

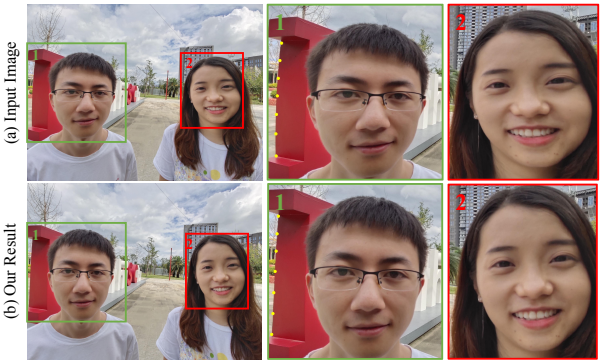


Figure 1. An example of our method. (a) the original wide-angle image with curved lines and distorted faces. (b) result by the proposed semi-supervised method, both lines and faces are corrected.

It is obvious that facial regions and background need two different types of projections for the wide-angle image correction. Carroll *et al.* [4] presented a content-preserving approach that finds an optimal mapping solution according to the user-specified lines. Recently, Shih *et al.* [26] designed an optimization problem to create a mesh that adapts stereographic projection on facial regions regionally and applies perspective projection on background, enabling a smooth transition between portraits and background by solving the optimization problem automatically. However, the method [26] sometimes causes distorted architectures nearby corrected faces. In addition, it requires portraits segmentation mask and camera parameters as additional inputs.

Tan *et al.* [29] proposed the first fully-supervised CNN-based method for wide-angle image correction, which consists of a line correction network and a portraits correction network. Tan's method obtained satisfactory results with the distorted image as input. However, there still exists disadvantages in their work. First, it needs many training pho-

*Equal contribution. †Corresponding authors.

tos under rich scenarios, and each face in the photo must be manually undistorted by specific tools. Meanwhile, errors may occur in manual annotation, causing uneven annotation quality or introducing dirty data. Therefore, the whole data preparation procedure is complex and expensive, making it unrealistic to improve performance by enlarging the training dataset. Second, Tan's method also creates artifacts in some cases because it does not use long-range semantic information for local variations of faces.

To address the above problems, we attempt to leverage a novel semi-supervised strategy, aiming to reduce the cost of preparing an expensive manual corrected dataset. Specifically, we adopt the semi-supervised strategy, containing direction and range consistency (DRC) and regression consistency (RC), to make full use of both labeled and unlabeled data by introducing a surrogate task (segmentation). Besides, compared with Tan *et al.* [29], we develop a novel network based on the multi-scale swin transformer block (MSTB), dubbed as Multi-Scale Swin-Unet (MS-Unet) which is better suitable for portraits correction. In particular, we also collect more than $5,000$ unlabeled distortion images from different phones and scenes to train MS-Unet by the semi-supervised strategy. Experimental results show that our approach can correct distortions in wide-angle portraits with superior performance than previous methods, and it only needs a small amount of manually labeled data. In summary, our main contributions are:

- We propose the first semi-supervised learning strategy for wide-angle portraits correction, which dramatically reduces the requirement of labeled training data.
- We develop a novel transformer-based network called MS-Unet, based on MSTB, to fully utilize both local-scale and long-range semantic information interaction for wide-angle portraits correction.
- We provide a high-quality unlabeled dataset that can be used to train semi-supervised wide-angle portraits correction algorithms.

2. Related Works

2.1. Wide-Angle Portraits Correction

Early wide-angle portraits correction methods always relied on traditional algorithms [4, 40]. Tehrani *et al.* [31, 32] presented methods to remove faces distortions and preserve background features during this process, but their solutions require user assistance. Shih *et al.* [26] proposed a mesh-based algorithm that can strike a balance between straight lines and faces correction effects automatically. Nevertheless, it requires the camera parameters and portraits segmentation as inputs. Recently, Tan *et al.* [29] proposed a two-stage deep neural network to complete wide-angle portraits correction with only an image as input. However, this fully-supervised method is limited to the number of labeled data that requires high-cost manual screening and processing. Fortunately, our method greatly reduces the limitation of the amount of labeled training dataset and learns the correction flow maps from distortion image to usual image.

2.2. Deep Semi-Supervised Learning

Deep semi-supervised learning provides a practical and effective approach to fully utilizing the mixture dataset containing labeled and unlabeled images. It has been widely used in image classification [13,36,37], semantic segmentation [1, 35, 38], machine translation [6, 9, 12], crowd counting [21, 23], text classification [15, 17, 18], text segmentation [30, 34] and so on. These works have proved that the semi-supervised method can promote the accuracy. Therefore, we introduce the semi-supervised strategy into the portraits correction domain and make a beautiful breakthrough.

2.3. Visual Transformer

The proposal of transformer [33] has been widely used in natural language processing (NLP). Inspired by their outstanding achievements, researchers have gradually applied transformers to the computer vision field recently [11, 16]. More impressively, Liu *et al.* [22] proposed an excellent hierarchical transformer structure called Swin Transformer, which is established upon shifted window partitioning mechanism. It has advanced performance on various vision tasks, including image classification, object detection, and semantic segmentation. Hu *et al.* [2] also devised a U-shaped transformer block called Swin-Unet, which focused on medical image segmentation and achieved surpassing results. Based on these works, we propose a new transformer network that can meet the need for long-distance semantic information of wide-angle portraits correction.

3. Method

Fig. 2 shows a pipeline of the proposed method. We devise a novel semi-supervised scheme to solve the problem of limited training data by utilizing both labeled and unlabeled data. As shown, we assume a single distortion image as input. Then, we get the correction flow maps and the segmentation mask as intermediate outputs. The correction flow maps are used to project the distortion image into a correction image. The segmentation mask is the bridge between labeled and unlabeled data.

3.1. Semi-supervised Learning Algorithm

As shown in Fig. 2, in our problem settings, we have a set of unlabeled images noted as $U = (I^u_{th})$ and a set of labeled images $L = (I^l_{th}, F_{th})$, where F_{th} represents the labels. We mix these images and adopt them to train the correction network through the semi-supervised method composed of DRC and RC, which are described in detail below.

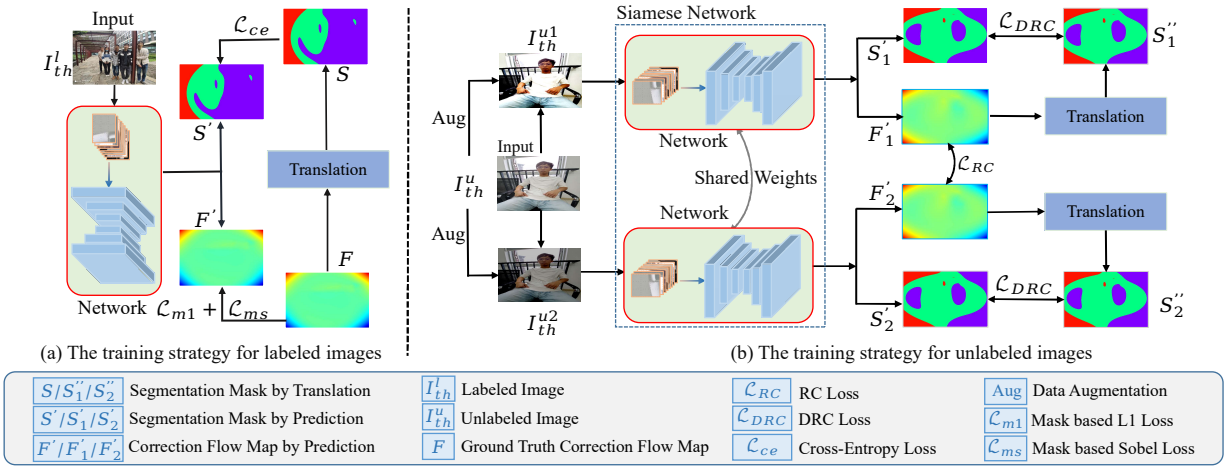


Figure 2. The pipeline of semi-supervised wide-angle portraits correction framework with the surrogate task (segmentation). (a) The network training strategy by utilizing the labeled images. (b) Utilize the unlabeled images to train our network. The training strategy consists of direction and range consistency (DRC), regression consistency (RC). For an unlabeled image I^u_{th}, when it is sent to the siamese network, the estimated segmentation mask and the correction flow map are utilized to compute the DRC loss $\mathcal{L}_{DRC}$ and RC loss $\mathcal{L}_{RC}$.

3.1.1 Direction and Range Consistency (DRC)

Many existing methods have proved that the estimation accuracy can be further improved by introducing approximate surrogate tasks [10, 23]. Inspired by the success, we attempt to present a surrogate task (segmentation) into the network that is different from the existing fully-supervised wide-angle portraits correction method [29]. In particular, the segmentation mask from the surrogate task can assist the network to construct a novel direction and range consistency learning strategy, which is helpful to improve the accuracy of wide-angle portraits correction.

This design is mainly motivated by four aspects: 1) The portraits correction flow maps represent the offset and direction of each pixel for correcting distortion images. By introducing the segmentation task to flow maps, the network pays more attention to learning the direction change of each pixel. It is conducive to the network to understand the portraits correction better. 2) If we generate a binary mask, the network will pay more attention to the guiding role of direction but ignore the importance of regional consistency. Thus the multi-category mask is generated by multiple thresholds to supervise the segmentation task. In the segmentation mask, the pixels classified into the same category represent that their values change within the same threshold range. In other words, the segmentation mask is also helpful to guide the network to learn the information of regional consistency, so that the correction flow map predicted by the network will also become smoother. 3) As shown in Fig. 2, the predicted correction flow map can also be converted to the segmentation mask. Hence the loss function can be constructed between portraits correction and segmentation, making it feasible to introduce unlabeled data for our semi-supervised scheme. 4) Meanwhile, the segmentation mask can be generated without extra costs. It is conducive for the transformation between the flow map and segmentation mask so. In addition, the unlabeled data can be fully utilized when conducting our DRC learning strategy.

Semantic segmentation and portraits correction have similar characteristics, making it possible to learn the consistency between them. When the semantic segmentation is deployed as the surrogate task in this paper, it predicts whether the flow map value $F(i,j)$ meets the given direction and range. We judge the offset by the threshold $\delta \in \mathbb{N}^+$, the pixels whose offset is in the range $(-\infty, \text{-}\delta]$ or $[\delta, +\infty)$ keep negative or positive directions, and the offset in the range $(\text{-}\delta, \delta)$ are merged into one set which indicates slight movement. The prediction target of the segmentation task is defined as follows:

$$S(i,j) = \begin{cases} 0, & if \quad F(i,j) \leqslant -\delta \\ 1, & if \quad -\delta < F(i,j) < \delta \\ 2, & if \quad F(i,j) \geqslant \delta \end{cases}, \tag{1}$$

where $S(i,j)$ denotes the segmentation mask, (i,j) is the pixel position of mask or flow map, and δ represents the predefined threshold is set to 5 in our experiments.

The proposed DRC learning strategy is shown in Fig. 2. As mentioned above, by introducing the surrogate task, the network can vigorously supervise direction and regional consistency. For labeled data, the ground truth of the correction flow map is used for training the portraits correction task. Meanwhile, it also converts into a multi-classification mask, which is utilized for training the surrogate task. As for the unlabeled data, no ground truth is available. Nevertheless, the predicted correction flow map can also generate

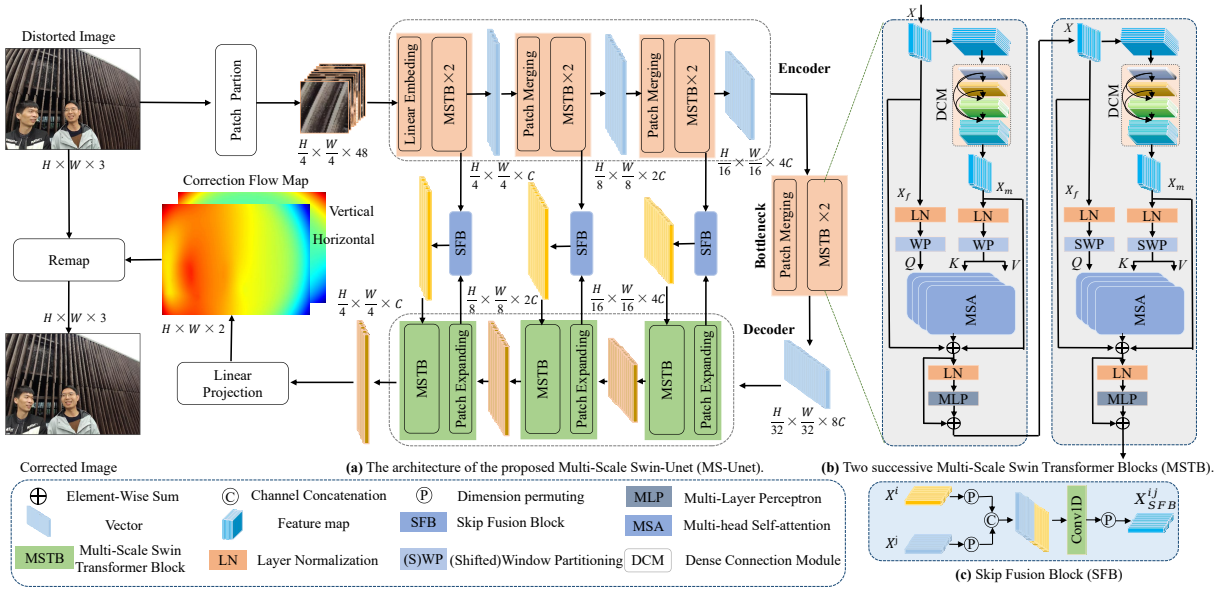


Figure 3. (a) The overview of our proposed Multi-Scale Swin-Unet (MS-Unet). The network mainly consists of encoder, decoder, bottleneck and skip fusion blocks (SFB). (b) The architecture of two successive MSTBs. The primary difference between them is the windowing configurations (window partition and shifted window partition). (c) The detailed architecture of SFB.

the segmentation mask through the multiple thresholds. The unlabeled data is still allowed to train the network by DRC loss, and the details will be given in Section 3.3.

3.1.2 Regression Consistency (RC)

Besides the DRC, we also introduce the regression consistency (RC) to improve the network robustness. Fig.2 illustrates the details of RC. Specifically, we can obtain two different images I_{th}^{u1} and I_{th}^{u2} with various augmentation methods (e.g., noise, smoothing, and sharpening), from an unlabeled image I_{th}^{u}. Many previous works have stated that an image with different perturbations can obtain similar predictions through a robust network. Therefore, we expand the MS-Unet into a shared-weight siamese structure. The unlabeled images I_{th}^{u1} and I_{th}^{u2} are respectively fed into the two networks, and a consistent loss is established between their outputs. The detailed loss implementation of RC will be given in Section 3.3.

3.2. Multi-Scale Swin-Unet (MS-Unet)

3.2.1 Architecture Overview

Although the semi-supervised scheme can significantly boost the performance, the portraits correction depends on a superior network. Motivated by the success of vision transformers [2,7,11,16], we develop the MS-Unet, derived from Swin-Unet [2], for the wide-angle portraits correction task. As shown in Fig. 3(a), our proposed MS-Unet can be divided into four major parts: encoder, decoder, bottleneck and skip fusion blocks.

Overall, there are two primary differences between MS-Unet and Swin-Unet. First, as the core unit of Swin-Unet, the swin transformer block ignores the importance of local-scale information, which leads to some objects (e.g., faces with different sizes) being distorted after correction. Second, directly employing the skip connection may not be the optimal scheme for hierarchical features fusion owing to their difference. To alleviate these issues, we leverage the MSTB as the basic unit of our MS-Unet to integrate local-scale and long-range information. Furthermore, the simple yet efficient SFB is designed to replace the skip connection.

3.2.2 Multi-Scale Swin-Transformer Block (MSTB)

Similar to EMSA of RestT [39], we develop the dense connection module (DCM) into the MSTB for local multi-scale information extraction. In Fig.3 (b), two successive MSTBs are presented. Each MSTB contains the DCM, layernorm (LN), multi-head self-attention (MSA), skip connection, and multi-layer perceptron (MLP). The window partitioning (WP) and shifted window partitioning (SWP) are used in two successive MSTBs.

When the features $X \in \mathbb{R}^{C\times H\times W}$ with a height of H and a width of W are fed into the MSTB, they will pass through two parallel branches for computing the input of MSA (the query Q, key K, and value V). In the left branch, X are split into non-overlapping windows with a size of $h \times w$ by (S)WP. The features are flattened and reshaped as $X_f \in \mathbb{R}^{N\times C}$, where $N = h \times w$. Then a full con-

nection layer is applied to obtain query $Q \in \mathbb{R}^{N\times d}$, where $d = C/k$ and k is the head number. In the right branch, the features X are first utilized to extract local-scale information by DCM. Inspired by [14], the DCM consists of two 1×1 layers, and three 3×3 depthwise separable convolution layers with different dilation rates $D = (1, 2, 3)$. To be specific, the 1×1 convolution layers are employed to change the feature dimension. Each 3×3 depthwise separable convolution layer will receive the features from all preceding layers $(i.e, x_0, ..., x_{r-1})$ as input:

$$x_r = C_r([x_0, ..., x_{r-1}]), \tag{2}$$

where C_r denotes the concatenation operation. Then we apply the same operations like the left branch on these features from the DCM to generate $X_m \in \mathbb{R}^{N\times C}$. The key $K \in \mathbb{R}^{N\times d}$, and value $V \in \mathbb{R}^{N\times d}$ are obtained through X_m. Afterward, the MSA can be calculated as follows:

$$MSA(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}} + B)V, \tag{3}$$

where $B \in \mathbb{R}^{N\times N}$ refers to learnable relative position bias.

3.2.3 Skip Fusion Block (SFB)

As mentioned above, the crucial difference among features from different hierarchical stages will be ignored when directly adopting skip connection. Hence, we designed the simple yet efficient skip fusion block (SFB) to replace the skip connection. As shown in Fig. 3 (c), before the features $X^i \in \mathbb{R}^{N\times C}$, $X^j \in \mathbb{R}^{N\times C}$ (from the i^{th} stage of encoder, and the j^{th} stage of decoder) are sent to the next stage of decoder, they pass through the SFB to form new features X^{ij}_{SFB} with dimension $\mathbb{R}^{N\times C}$. The whole calculation process is defined as follows:

$$X^{ij}_{SFB} = D(CON(C[D(X^i), D(X^j)])), \tag{4}$$

where $D(\cdot)$ is dimension permuting, $C[\cdot]$ refers to the concatenation, and $CON(\cdot)$ is 1D convolution layer.

3.3. Loss Function

In practice, the MS-Unet is optimized by adopting the supervised losses on the labeled data L, and the semi-supervised losses on the unlabeled data U.

3.3.1 Supervised Loss

Our constructed supervised loss $\mathcal{L}_s$ is composed of three parts, including mask-based $L1$ loss $\mathcal{L}_{m1}$, mask-based sobel loss $\mathcal{L}_{ms}$, and the cross-entropy loss $\mathcal{L}_{ce}$. The detailed definitions are described as follows:

a) $\mathcal{L}_{m1}$ *Loss*: In our method, we introduce the weighted mask, which uses the weight value of the portraits area to be greater than that of the background so that the network will pay more attention to the distorted portraits. Eq. 5 gives the definition of this loss.

$$\mathcal{L}_{m1} = \left| F^{'} - F \right| M, \tag{5}$$

where F and $F^{'}$ represent the ground truth and estimated flow maps, respectively, M denotes the weighted mask.

b) $\mathcal{L}_{ms}$ *Loss*: In portraits correction, the object edges directly affect the overall visual effects of a correction image. Therefore, we introduce the sobel loss, which can be expressed as follows:

$$\mathcal{L}_{ms} = \left[\left| G_x(F^{'}) - G_x(F) \right| + \left| G_y(F^{'}) - G_y(F) \right| \right] M, \tag{6}$$

where the G_x and G_y mean the sobel operator in horizontal and vertical direction, respectively.

c) $\mathcal{L}_{ce}$ *Loss*: To supervise the mask generated from the segmentation task, we convert the ground truth flow map into a mask label and deploy the cross-entropy loss. The loss function is defined as follows:

$$\mathcal{L}_{ce} = Slog(S^{'}) + (1 - S)log(1 - S^{'}), \tag{7}$$

where the S is the ground truth mask converted from the flow map, and the $S^{'}$ refer to the estimated mask. To sum up, the training loss for a labeled image is:

$$\mathcal{L}_s = \mathcal{L}_{m1} + \lambda_1 \mathcal{L}_{ms} + \lambda_2 \mathcal{L}_{ce}, \tag{8}$$

where λ_1 and λ_2 are hyper-parameters of $\mathcal{L}_{ms}$ loss and $\mathcal{L}_{ce}$ loss respectively, both being set to 10 in our experiments.

3.3.2 Semi-Supervised Loss

For an unlabeled image, we construct the unsupervised loss $\mathcal{L}_u$ based on the surrogate (segmentation) task and flow map task, which is used to guide the prediction consistency of the network. Specifically, The unsupervised loss contains two parts: the loss of DRC $\mathcal{L}_{DRC}$ and RC $\mathcal{L}_{RC}$.

$$\begin{aligned} \mathcal{L}_u &= \mathcal{L}_{RC} + \mathcal{L}_{DRC} \\ &= [\mathcal{L}_{m1}(F^{'}_1, F^{'}_2) + \lambda_2 \mathcal{L}_{ms}(F^{'}_1, F^{'}_2)] + \\ &\quad [\mathcal{L}_{ce}(S^{'}_1, S^{''}_1) + \mathcal{L}_{ce}(S^{'}_2, S^{''}_2)], \end{aligned} \tag{9}$$

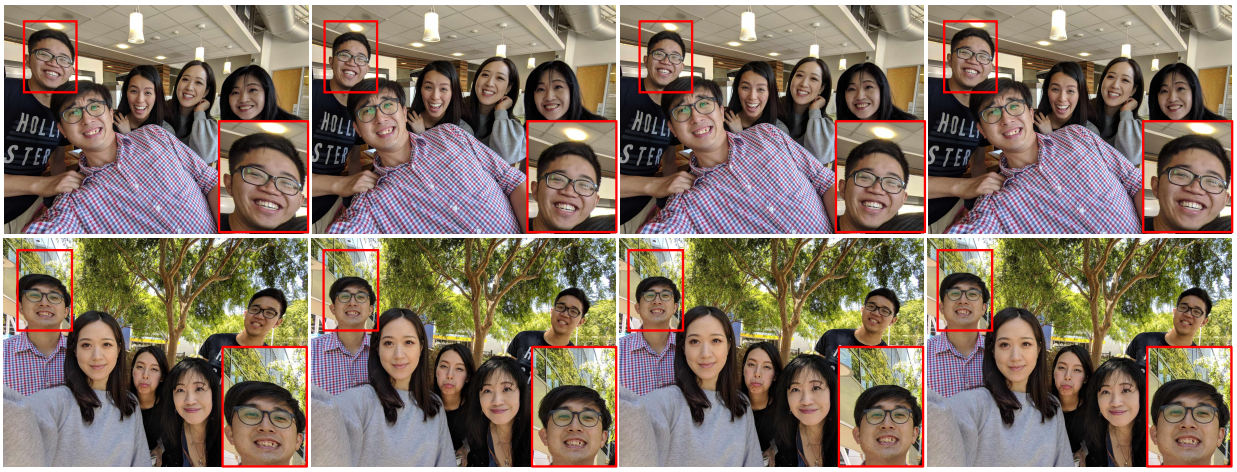
where the $F^{'}_1$ and $F^{'}_2$ are the estimated flow maps from both the branches of the siamese network, $S^{'}_1$ and $S^{'}_2$ refer to the segmentation mask converted from $F^{'}_1$ and $F^{'}_2$, while the $S^{''}_1$ and $S^{''}_2$ indicate the output of the siamese network.

4. Experiments

4.1. Implementation Details

4.1.1 Datasets

Following the existing method [29], we conduct extensive experiments on the wide-angle dataset [29], captured



(a) Projection Image (b) Shih's Result (c) Tan's Result (d) Our Result

Figure 4. Qualitative results of different correction methods. Notice the coordination of lines and face area marked with red boxes.

with 5 different smartphones. The training dataset contains over $5,000$ images and 129 in the testing dataset. Many kinds of labels are provided for each image in the dataset, containing the face mask, correction flow maps, and landmarks. In addition, we collected more than $5,000$ images by 4 different smartphones (including Samsung Note 10, Xiaomi 11, vivo X23 and vivo iQOO) as the unlabeled set.

4.1.2 Training Details

We train the MS-Unet via a two-step scheme. Similar to [29], the input size is set as 512×384. Before the semi-supervised strategy starts, we need to train a correction flow map predictor, which can provide the pseudo labels for the surrogate task when both labeled and unlabeled images are utilized. We found that the predictor can achieve good results with only 200 epochs. Then, we introduce the surrogate task (segmentation) to the network, which can enhance the learning ability of the network. Based on the surrogate task, the semi-supervised method can further improve the network's performance. In this stage, the total training epoch is set to $1,000$. Notably, the supervised loss is utilized for labeled images, while the unsupervised loss is for unlabeled images. To compute the loss conveniently, each batch only contains the labeled images or unlabeled images. For both steps, we use Adam to optimize our model with an initial learning rate of 1×10^{-4}, and the weight decay of 2×10^{-4}. We trained the MS-Unet (8.82 GFLOPs, Parameters: 8.79M) using 4 Geforce RTX 2080Ti, and tested it with only one GPU, which can run at around 40 FPS.

4.1.3 Evaluation Metrics

We use the same evaluation metrics (LineAcc and ShapeAcc) as [29] to evaluate the performance of our method. More specifically, LineAcc is used to evaluate the curvature variation of the marked lines and defined as follows:

$$LS = 1 - \frac{1}{n}\sum_{i=0}^{n-1}\left(\frac{y_{d_i} - y_{d_{i-1}}}{x_{d_i} - x_{d_{i-1}}} - \frac{y_{g_0} - y_{g_n}}{x_{g_0} - x_{g_n}}\right), \quad (10)$$

where LS denotes the similarity between slope of these two lines, n is the number of uniformly sampled points in each line. (x_{g_i}, y_{g_i}) and (x_{d_i}, y_{d_i}) indicate the coordinate of the corresponding point in the reference and distortion image.

ShapeAcc aims to evaluate the face similarity between the correction image and the reference image. Based on face landmarks, the ShapeAcc is described as follows:

$$FC = \frac{1}{n}\sum_{i=0}^{n-1} \|L_{g_i}\| \, \|L_{d_i}\| \, cos\theta, \quad (11)$$

where FC is the similarity between the corrected and target face, n is the number of fixed sampled points in each face. L_g and L_d are the corresponding face landmarks in the correction image and the reference image.

4.2. Ablation Study

In order to verify the influence of different factors on our proposed method, we conducted some ablation experiments on Tan's dataset [29] and our unlabeled dataset. Notably, the network structure, the semi-supervised strategy, and the number of unlabeled samples are all considered below.

4.2.1 Effect of the Correction Network

We explore that how the proposed modules affect the network performance using fully-supervised method. Specifically, we utilize the Swin-Unet as our baseline, and the performance of three different networks is evaluated. 1) Baseline: directly employ the Swin-UNet; 2) Baseline+MSTB:

based on 1), the MSTB is considered to replace the swin transformer block; 3) Baseline+MSTB+SFB (MS-Unet): the SFB is added to fuse the hierarchical features, and Table 1 presents the results. We can observe that the performance boosts significantly with the addition of each module from the table. Oddly, when both MSTB and SFB are added to the network, the full MS-Unet can achieve the best LineAcc (66.825) and ShapeAcc (97.491). These experiments demonstrate that MSTB indeed promotes the network to extract more complementary information, which boosts the correction ability dramatically. Meanwhile, SFB provides a better feature fusion strategy than skip connections.

In addition, we compared MS-Unet and Tan's method under the same conditions, and the experiment shows that the accuracy of MS-Unet is slightly higher than Tan's network (LineAcc: 66.784, ShapeAcc: 97.490).

Table 1. Ablations on the structure of proposed MS-Unet.

Index	Baseline	MSTB	SFB	LineAcc	ShapeAcc
1)	✓	-	-	66.514	97.460
2)	✓	✓	-	66.763	97.487
3)	✓	✓	✓	**66.825**	**97.491**

4.2.2 Effect of the Semi-Supervised Strategy

Several experiments are conducted to evaluate the impact of our proposed semi-supervised scheme. In practice, we first utilize the fully-supervised method to train our MS-Unet with only Tan's dataset [29]. The training result is regarded as the baseline for comparison, and we present it in the first row shown in Table 2. Then we add the surrogate task to the network and train the two-task MS-Unet. The second row in Table 2 reports the results of the two-task MS-Unet. Compared with the baseline, it shows a slight improvement after adding the surrogate task (LineAcc: $66.825 \rightarrow 66.871$, ShapeAcc: $97.491 \rightarrow 97.493$). The result indicates that introducing a surrogate task plays a guiding role in networking training to a certain extent. Afterward, both labeled (Tan's dataset) and unlabeled data are deployed to accomplish the experiments about semi-supervised strategy. The DRC is conducted based on the segmentation task, and the third row in Table 2 lists the comparison result. Compared with the two-task MS-Unet, adding DRC can further improve the estimation accuracy of the correction flow maps, especially the LineAcc (from 66.871 to 67.154). Besides, the effect of RC is also evaluated, and the result is presented in the fourth row of Table 2. The result also outperforms the single-task MS-UNet, which is only trained by the fully-supervised scheme. The MS-Unet attains the best result (LineAcc: 67.209, ShapeAcc: 97.500) when DRC and RC are employed during the semi-supervised training. These experiment results prove that our proposed semi-supervised strategy can assist the consistency learning between portraits correction task and surrogate and improve the correction performance.

Table 2. Performance comparison of different semi-supervised strategies. 'Seg' indicates a segmentation task without direction and range consistency. 'DRC' refers to the direction and range consistency, and 'RC' refers to the Regression Consistency.

	Baseline	Seg	DRC	RC	LineAcc	ShapAcc
1)	✓	-	-	-	66.825	97.491
2)	✓	✓	-	-	66.871	97.493
3)	✓	✓	✓	-	67.154	97.494
4)	✓	-	-	✓	66.848	97.497
5)	✓	✓	✓	✓	**67.209**	**97.500**

4.2.3 Effect of the Number of Unlabeled Samples

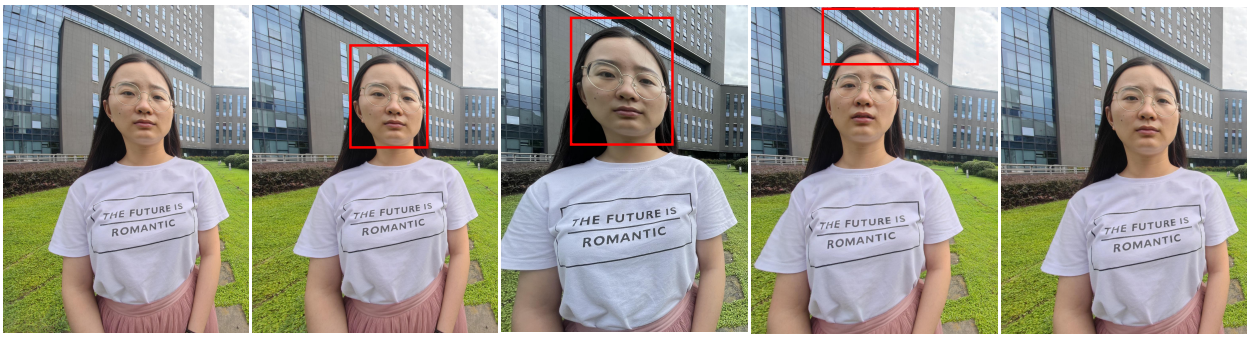
We examine the influence of the number of unlabeled images on network performance through Tan's dataset [29] and our unlabeled data. We change the number of unlabeled images from 0 to $5,000$ while the number of labeled images is fixed. The results are listed in Table 3, where it shows our MS-Unet trained with semi-supervised strategy obtains consistent superior performance compared with that using the fully-supervised scheme. Meanwhile, we can draw that in a specific range, the performance of the MS-Unet will improve as the amount of unlabeled images increases.

Table 3. The impact of the number of unlabeled images.

Index	numbers	LineAcc	ShapeAcc
1)	0	66.871	97.493
2)	1000	66.929	97.493
3)	2000	66.999	97.494
4)	3000	67.105	97.497
5)	4000	67.155	97.496
6)	5000	**67.209**	**97.500**

4.3. Comparison with Other Methods

We also compare our method with previous state-of-the-art methods on Tan's and Google's test sets. Table 4 illustrates that our method obtains the highest metric results in both two test sets. The visual comparisons in Fig. 4 also confirm the results. Note that the projection image can correct lines but make faces distorted seriously. Shih [26] and Tan [29] try to seek the optimal trade-off between the faces and the background. Unfortunately, several bent structures still exist in the background, and a few faces are still distorted. From our results, the faces are more natural than other methods, and the corrected lines in the background are satisfactory. Generally, both quantitative and qualitative results verify the superior performance of our method.



(a) Distortion Image (b) Projection Image (c) iPhone12's Result (d) Xiaomi11 ultra's Result (e) Our Result

Figure 5. Visual comparison between our method and some wide-angle portraits correction methods from smartphones.

Table 4. The results of our proposed method and the two classic methods on two wide-angle portraits correction test sets.

Method	Tan's test set		Google's test set	
	LineAcc	ShapeAcc	LineAcc	ShapeAcc
Shih [26]	66.143	97.253	61.551	97.464
Tan [29]	66.784	97.490	64.650	97.499
Ours	**67.209**	**97.500**	**66.098**	**97.512**

Fig. 5 depicts the results of our method and other famous portraits correction algorithms from smartphones (i.e., Xiaomi 11 ultra, and iPhone 12). We can observe that serious stretching of portraits appears in iPhone 12. Although Xiaomi 11 ultra improves over the distortion image, there is still slight deformation on the face and curved lines in the background. Our method shows better results, as the face is natural while correcting the lines in the background.

Only a few wide-angle portraits correction works employ the deep learning methods due to its challenge. Based on the correction flow maps, the correction task is regarded as a pixel-level regression problem, which is closely related to some other tasks, such as crowd counting [10, 21] and semantic segmentation [2, 5]. Hence, we introduce some efficient networks from these fields to predict the correction flow maps. All the networks are trained by the fully-supervised scheme, and Table 5 shows the results. Notably, our proposed MS-Unet surpasses all the methods. The primary reason is that the CNN-based networks focus on learning local-scale information while the transformer-based networks concentrate on long-range information. For wide-angle portraits, the long-range information can ensure the corrected image generally looks more natural, and the face corrected by local information is more authentic. Therefore, combining both advantages, the MS-Unet will capture multi-scale information for more accurate estimation.

Finally, our proposed semi-supervised strategy is used to train these networks. Different from the original network architecture, the surrogate task is added to the network during the training process. And all the semi-supervised results are listed in Table 5. We can observe that the accuracy of these networks assisted by both labeled and unlabeled data is improved compared with the conventional fully-supervised scheme. The experimental results also demonstrate the generalization ability of our semi-supervised method.

Table 5. The effectiveness evaluation of the proposed semi-supervised scheme on different networks.

Method	Fully-Supervised		Semi-Supervised	
	LineAcc	ShapeAcc	LineAcc	ShapeAcc
RefineNet [20]	66.348	97.449	66.569	97.455
UNet [25]	65.246	97.473	66.534	97.475
CSRNet [19]	65.967	97.469	66.236	97.471
Deeplab v3+ [5]	66.200	97.482	66.565	97.487
Swin-Unet [2]	66.514	97.460	66.859	97.469
HRNet [27]	66.748	97.477	66.805	97.491
Ours	**66.825**	**97.491**	**67.209**	**97.500**

5. Conclusion

In this paper, we develop a novel semi-supervised wide-angle portraits correction method using a multi-scale transformer. By combining DRC and RC in our semi-supervised manner, we can solve the limitations of labeled data and fully utilize unlabeled data. In addition, four kinds of smartphones are adopted to collect unlabeled data. Furthermore, we especially propose the MS-Unet, built upon the MSTB, to capture both local-scale and long-range information for improving artifacts around portraits. Extensive experimental results show that our proposed method is much better than the existing advanced methods and can be popularized in the application of wide-angle portraits correction.

Acknowledgement This work was supported by the National Natural Science Foundation of China (NSFC) under grants No.61872067 and No.62172032.

References

[1] Yauhen Babakhin, Artsiom Sanakoyeu, and Hirotoshi Kitamura. Semi-supervised segmentation of salt bodies in seismic images using an ensemble of convolutional neural networks. In *German Conference on Pattern Recognition*, pages 218–231, 2019. 2

[2] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021. 2, 4, 8

[3] Robert Carroll, Aseem Agarwala, and Maneesh Agrawala. Image warps for artistic perspective manipulation. *ACM Trans. Graphics*, 29(4):1–9, 2010. 1

[4] Robert Carroll, Maneesh Agrawala, and Aseem Agarwala. Optimizing content-preserving projections for wide-angle images. *ACM Trans. Graph.*, 28(3):43, 2009. 1, 2

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. ECCV*, pages 801–818, 2018. 8

[6] Yong Cheng. Semi-supervised learning for neural machine translation. In *Joint training for neural machine translation*, pages 25–40. 2019. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[8] Song-Pei Du, Shi-Min Hu, and Ralph R Martin. Changing perspective in stereoscopic images. *IEEE Trans. on Visualization and Computer Graphics*, 19(8):1288–1297, 2013. 1

[9] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48, 2021. 2

[10] Junyu Gao, Qi Wang, and Xuelong Li. Pcc net: Perspective crowd counting via spatial convolutional network. *IEEE Trans. on Circuits and Systems for Video Technology*, 30(10):3486–3498, 2019. 3, 8

[11] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020. 2, 4

[12] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. *Proc. NeurIPS*, 29:820–828, 2016. 2

[13] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proc. CVPR*, pages 558–567, 2019. 2

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proc. CVPR*, pages 4700–4708, 2017. 5

[15] Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. Leveraging just a few keywords for fine-grained aspect detection through weakly supervised co-training. *arXiv preprint arXiv:1909.00415*, 2019. 2

[16] S. Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, F. Khan, and M. Shah. Transformers in vision: A survey. *ArXiv*, abs/2101.01169, 2021. 2, 4

[17] Ju-Hyoung Lee, Sang-Ki Ko, and Yo-Sub Han. Salnet: Semi-supervised few-shot text classification with attention-based lexicon construction. In *Proc. AAAI*, pages 13189–13197, 2021. 2

[18] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. *Proc. NeurIPS*, 32:10276–10286, 2019. 2

[19] Yuhong Li and Xiaofan Zhang. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proc. CVPR*, pages 1091–1100, 2018. 8

[20] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proc. CVPR*, pages 1925–1934, 2017. 8

[21] Yan Liu, Lingqiao Liu, Peng Wang, Pingping Zhang, and Yinjie Lei. Semi-supervised crowd counting via self-training on surrogate tasks. In *Proc. ECCV*, pages 242–259, 2020. 2, 8

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 2

[23] Yanda Meng, Hongrun Zhang, Yitian Zhao, Xiaoyun Yang, Xuesheng Qian, Xiaowei Huang, and Yalin Zheng. Spatial uncertainty-aware semi-supervised crowd counting. In *Proc. ICCV*, pages 15549–15559, 2021. 2, 3

[24] Darko Pavić, Volker Schönefeld, and Leif Kobbelt. Interactive image completion with perspective correction. *The Visual Computer*, 22(9-11):671–681, 2006. 1

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015. 8

[26] YiChang Shih, Wei-Sheng Lai, Chia-Kai Liang, and Chia-Kai Liang. Distortion-free wide-angle portraits on camera phones. *ACM Trans. Graphics*, 38(4):1–12, 2019. 1, 2, 7, 8

[27] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proc. CVPR*, pages 5693–5703, 2019. 8

[28] Benny Stale Svardal, Kjell Einar Olsen, and Odd Ragnar Andersen. Stereographic projection system, Apr. 15 2003. US Patent 6,547,396. 1

[29] Jing Tan, Shan Zhao, Pengfei Xiong, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical wide-angle portraits correction with deep structured models. In *Proc. CVPR*, pages 3498–3506, 2021. 1, 2, 3, 5, 6, 7, 8

[30] Youbao Tang and Xiangqian Wu. Scene text detection and segmentation based on cascaded convolution neural networks. *IEEE transactions on Image Processing*, 26(3):1509–1520, 2017. 2

[31] Mahdi Abbaspour Tehrani, Aditi Majumder, and Meenakshisundaram Gopi. Undistorting foreground objects in wide angle images. In *2013 IEEE International Symposium on Multimedia*, pages 46–52. IEEE, 2013. 2

[32] Mahdi Abbaspour Tehrani, Aditi Majumder, and M Gopi. Correcting perceived perspective distortions using object specific planar transformations. In *Proc. ICCP*, pages 1–10, 2016. 2

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, pages 5998–6008, 2017. 2

[34] Chuan Wang, Shan Zhao, Li Zhu, Kunming Luo, Yanwen Guo, Jue Wang, and Shuaicheng Liu. Semi-supervised pixel-level scene text segmentation by mutually guided network. *TIP*, 30:8212–8221, 2021. 2

[35] Huaxin Xiao, Yunchao Wei, Yu Liu, Maojun Zhang, and Jiashi Feng. Transferable semi-supervised semantic segmentation. In *Proc. AAAI*, 2018. 2

[36] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proc. CVPR*, pages 10687–10698, 2020. 2

[37] Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. Variational autoencoder for semi-supervised text classification. *Proc. AAAI*, 2017. 2

[38] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proc. CVPR*, pages 7151–7160, 2018. 2

[39] Qinglong Zhang and Yu-Bin Yang. Rest: An efficient transformer for visual recognition. *Advances in Neural Information Processing Systems*, 34, 2021. 4

[40] Denis Zorin and Alan H Barr. Correction of geometric perceptual distortions in pictures. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 257–264, 1995. 2