Merging Super Resolution and Attribute Learning for Low-Resolution Person Attribute Recognition

RAMIN ABBASZADI AND NAZLI IKIZLER-CINBIS
Department of Computer Engineering, Hacettepe University, 06800 Ankara, Turkey
Corresponding author: Ramin Abbaszadi (n10241560@cs.hacettepe.edu.tr)

ABSTRACT In video surveillance, visual person attributes such as gender, backpack, type of clothing are crucial for person search or re-identification. For detecting and retrieving these attributes with high accuracy, the availability of high-quality videos is a necessity in general. However, this cannot be guaranteed in general surveillance videos or images; beside improving hardware technology, improving inference algorithms on low-resolution data is valuable. This paper attempts to propose two solutions in this direction: designing a combined neural network architecture from existing architectures, and a novel combination approach toward re-identification on low-resolution videos. The proposed architecture introduces a combined Neural Network architecture, called SRMAR, that jointly trains Super Resolution and Multi Attribute Recognition models for more effective recognition. Experiments on two benchmark datasets demonstrate the effectiveness and applicability of the proposed neural network architecture for low-resolution multi-attribute recognition. Furthermore, a higher-level linear combination scheme that optimally combines the proposed SRMAR architecture and multi-attribute recognition network is presented, yielding superior results in low-resolution person attribute recognition.

INDEX TERMS Attribute recognition, linear programming, linear combination, neural networks, parametric inference, super resolution.

I. INTRODUCTION
Person attribute recognition (PAR) concentrates on recognizing personal characteristics such as gender, age, hair, height, or clothing style, tools at hand, etc., especially from surveillance videos [1]. PAR has a variety of applications in the real world, including security applications, market behavior recognition, traffic management, population management, and more [2], [3]. Like any other tasks in computer vision, the PAR task suffers from limited computational resources and discrepancies in the image quality. Challenges include varying illumination conditions, different viewpoints, complex and varying appearance, and low-resolution (LR) in imagery. Furthermore, due to the nature of the problem, attribute recognition requires an analysis of more delicate details for achieving better results. However, when the imagery is in low resolution, the loss of information makes the recognition quite challenging.

Challenges due to illumination, viewpoint differences and varying appearance can be handled through large proper datasets to train Convolutional Neural Networks (CNNs) in a supervised fashion. Similarly, to address the issues of LR, recently, CNN architectures are being developed and applied with satisfactory results [4]. Here, instead of using neighborhood information of pixels, CNNs are designed to estimate the pixel value directly. These CNN-based models fall under the category of super-resolution (SR) [5], [6]. SR models can upsample LR images by learning from many samples, enabling the model to use local information and understand the scene [7].

Another property of PAR tasks is that there are usually many different attributes, e.g., age, height, hair, depending on the purpose of the application. Recently, this task has also been addressed using CNN architectures, specifically using so-called multi-attribute recognition (MAR) networks. This is currently a very active subject area, and it demonstrates a great potential for understanding the contents of the scene [8].

This work explores the PAR task in LR images as a multi-class recognition problem. To this end, we propose an end-to-end deep learning model that merges the power of SR models and MAR models in a joint architecture. We call this end-to-end CNN architecture as Super-Resolution Multi-Attribute...
FIGURE 1. Most real-world surveillance cases demand recognition of personal attributes from distant, which convinces the need to process these attributes in low resolution. Our proposed hybrid network combines the power of a SR network with multi-attribute recognition to provide better recognition for personal attributes in LR images.

Recognition (SRMAR) model. The idea is to combine the powers of SR networks and MAR models into a single one, which will be trained in an end-to-end fashion. The final architecture is achieved via the combination of two state-of-the-art architectures that have been proposed for SR and MAR tasks. We further propose a linear combination of the SRMAR network and the MAR network to boost the recognition performance. Our experiments verify that combining models is better for the low-resolution person attribute recognition.

The overall idea is demonstrated in Fig 1. To the best of our knowledge, this work is the first study that proposes an end-to-end learning model for MAR problems for LR images. We establish the corresponding analysis on two benchmark datasets: Market-1501 [9] and DukeMMTC-reID [10]. Experimental results demonstrate that the proposed SRMAR network increases recognition accuracy on LR images.

To sum up, the main contributions in this article are: i) A joint CNN architecture called SRMAR that utilizes an SR network trained to reconstruct LR images and a MAR network, ii) A novel linear combination scheme to combine the SRMAR model with the MAR model iii) improvements on the state-of-the-art PAR methods even on LR images and iv) extensive experiments the proposed models on two benchmark datasets that are widely used for PAR, and results demonstrate that using a joint network for SR and MAR tasks improve the recognition accuracy.

The acronyms used throughout the paper is listed in Table 1 for ease of reference.

II. RELATED WORK
This section briefly reviews the multi-attribute recognition (MAR) and super-resolution (SR) techniques.

A. MULTI-ATTRIBUTE RECOGNITION
One of the primary approaches towards leveraging semantically defined mid-level attributes is carried out by [1]. Low-level colors and texture features are obtained from six equitized images, which horizontally slice each person’s image for attribute detection. Each slice consisted of 8 color channels (YCbCr, HSV, and RGB) and 21 filters for textures. To detect attributes, a Support Vector Machine classifier is used. Deng et al. [11] evaluate several techniques such as intersection kernel-based Support Vector Machine, Markov random field with Gaussian kernel, and random forest. More recently, Li et al. [12] propose two models based on deep architectures to address the drawbacks of hand-crafted features and ignoring the relationship between attributes. Their single attribute recognizing model (DeepSAR) is designed for recognizing each attribute individually. The second model [12] is the deep learning framework to exploit the relationship between attributes.

Jianqing et al. [13], propose a multi-label ConvNet model for predicting multiple attributes in conjunction. Li et al. [8] try to get to the best performance of large-scale person


re-identification via paying attention to the connection and correlations between attribute labels and ID labels. Their algorithm has two stages. The first is PAR framework, and the other is attribute re-weighting module. With CNN in the PAR framework, the pedestrian image feature representations are extracted. In the first stage, losses of attributes are computed; those scores (loss results) after concatenating are sent to the second stage. Finally, the latest output of this MAR module is conjugated with the global features.

Shi et al. [14], propose a network with two modules: coarse and fine alignment modules. The first module uses a part detector to locate the body parts and form the candidate attributes; then, in the second module, these attributes are aggregated together via a bilinear-pooling layer. Wu et al. [15] propose a parallel model, which consists of intra-attention and inter-attention parts to learn the relationships of images or attributes.

B. SUPER RESOLUTION

One of the pioneering works on SR was [16], then the term SR appeared later around the year 1990 [5]. One of the recent approaches in single-image SR in the pre-deep learning era we can refer to is [7] based on image statistics. First, a sparse representation for each image patch of the LR input image was searched, and then they generated HR output using the coefficients of this representation. Deep learning techniques such as CNN-based SR methods have yielded outstanding results. Wang et al. [4] show that the conventional sparse coding structure’s domain expertise could be combined with deep learning’s critical ingredients to achieve further improved results. Dong et al. [17] use a three-layer deep fully convolutional network and use bicubic interpolation as an upscaling algorithm. With single-shot SR, they propose an end-to-end CNN-based mapping from the LR images to the High-Resolution (HR) images directly. Their proposed SR CNN. Later, they propose a fast SR convolution neural network [18] based on SR CNN [17]. fast SR convolution neural network has fewer parameters and lower computation costs than its predecessor, and these properties enable it to work on a regular CPU in real-time.

Generative adversarial networks (GANs) have been found to have an essential role in SR research [19]–[23]. Ledig et al. [24], focus on single image SR and present a GAN for super image resolution. In the GAN for SR, a conceptual loss function is used that includes an adversarial and content loss together. Yamanaka et al. [25] use deep CNN models and introduce an efficient and faster single-shot SR model.

For better performance, they used deeper CNN architectures, where they extend a parallelized CNN architecture. This architecture consists of two smaller parts, the feature extraction network, and the re-construction network. Lim et al. [26] develop an enhanced and improved deep SR network (EDSR) that is optimized by dropping out redundant modules in conventional residual networks. Haris et al. [27] propose Deep Back-Projection Networks (DBPN), which works via employing up and down-sampling layers iteratively and implementing an error feedback mechanism for projection errors.

III. PROPOSED METHOD

The overall pipeline of the proposed method end-to-end framework is illustrated in Fig.2. Our framework consists of joint training of two main parts: i) SR framework and ii) attribute recognition framework. In the following of this article, we discuss the details of these two main components.

A. SUPER RESOLUTION NETWORK

The first part of the proposed network is the SR component. Note that we aim to learn a joint network, where the SR part help the recognition of attributes in LR. In this context, any end-to-end trainable SR model can be integrated into our proposed framework. To this end, we utilize two recent state-of-the-art SR models: EDSR [26] and DBPN [27]. The reason for selecting these two is that they are state-of-the-art, covering two different areas of the spectrum in terms of network size, and we can reproduce exactly the same results that stated in the corresponding papers.

EDSR SR model [26] extends existing SR networks like [24], [28] that use residual networks by removing additional modules/layers like batch normalization and ReLU. With the elimination of batch normalization layers, the authors show that the performance increases considerably, and the required memory is reduced by 40% during training. This reduction helps in improving the model by expanding the size; as a result, EDSR yields better performance. For upsampling factors \( \times 3 \) and \( \times 4 \), the training model is initialized with a pre-trained \( \times 2 \) network. This expedites the training and increases the final performance.

The second SR model that we have used is the DBPN [27] model. DBPN [27] has two stages, the reciprocally connected up and down-sampling stage and the error feedback stage. Generally, feed-forward architectures that act as a one-way mapping only map rich representations of the input to output space. Such an approach is not successful in mapping LR images because of the limited features available. To solve

| Acronyms used in the paper. |
|-----------------------------|
| PAR | Person Attribute Recognition | SR | Super-Resolution |
| MAR | Multi-Attribute Recognition | LR | Low-Resolution |
| DBPN | Deep Back-Projection Network | CNN | Convolutional Neural Network |
| EDSR | Enhanced and improved Deep SR network | SRMAR-D | SRMAR network uses DBPN as backbone |
| SRMAR | Super-Resolution Multi-Attribute Recognition | SRMAR-E | SRMAR network uses EDSR as backbone |
this problem, DBPN [27] model generates HR features during up-sampling and, during down-sampling, these features are projected back to LR space. The second stage, error feedback stage, has a mechanism from the up to down-scaling steps that positively influences the training process to achieve a better reconstruction.

Both of these SR networks were trained following to the original papers [26], [27] for EDSR and DBPN models respectively, using [29] loss function on the outputs of SR network:

$$L_{SR}(\hat{I}, I) = \frac{1}{hwc} \sum_{ijk} (\hat{I}_{ijk} - I_{ijk})^2 + \epsilon^2$$  (1)

We combine the $L_{SR}$ with dice loss function:

$$DC(\hat{I}, I) = 1 - \frac{2 \sum_{i,j,k} \hat{I}_{i,j,k} I_{i,j,k}}{\sum_{i,j,k} \hat{I}_{i,j,k} + \sum_{i,j,k} I_{i,j,k}}$$

and get the following loss function:

$$L_{CSR}(\hat{I}, I) = DC(\hat{I}, I) + L_{SR}(\hat{I}, I)$$  (2)

which is then used to train the SRMAR model.

B. THE ATTRIBUTE NETWORK

The second part of the proposed model is a network for learning attributes. LR images that are up-scaled by the SR component are fed into the attribute recognition network for predicting the corresponding attributes. For the MAR task, we adopt the recent network proposed by Lin et al. [8]. This network aims for person re-identification and pedestrian attribute recognition at the same time. We adopt the person attribute recognition part of their model, which is trained just on the attribute data set using ResNet-50 [30] as backbone. This backbone is followed by the attribute recognition, which includes $M$ (number of attributes) Fully Connected layers followed by a Softmax layer [8]. The binary cross-entropy:

$$L_{MA}(\hat{Y}, Y) = -\sum_i Y_i \log(\hat{Y}_i) + \frac{\gamma}{2} \sum_j ||w_j||_2$$  (3)

is used as the loss function in training where $\hat{Y}$ is the predicted output and $Y$ is the ground truth label, $\gamma$ is regularization factor set to 0.02, and $w_j$ represents the weights in the convolution layer $j$.

C. THE SUPER-RESOLUTION MULTI-ATTRIBUTE RECOGNITION NETWORK

For the combined SRMAR neural network, the training process can be handled in different ways. Designing loss functions and optimizers are essential. We have experimented with different loss functions, data augmentation, and other hyper-parameter optimization techniques. Our idea is to design weighted loss based on SR and MAR network loss functions. This helped to reduce the overfitting effect during the training, however, it requires a subtle selection of the regulation parameter $\beta$;

$$L = \beta L_{CSR} + L_{MA}$$  (4)

Based on results, during the training we decided to set $\beta = 0.00005$ which is a trade off between the accuracy and over-fitting of the network. Merging SR and MAR networks leads to a large network that should be trained with care.

D. LINEAR COMBINATION OF MODELS

We also investigate the effect of combining the SRMAR network with the MAR network. The overall architecture for this combination scheme is demonstrated in Fig.3. The idea is to combine the predictive power of the SRMAR model and the MAR model, since our preliminary experiments indicate
that some attributes are predicted better by SRMAR network, whereas some others are predicted by MAR network.

The simplest approach to achieve this is defining equal weights (i.e. $w_k = 1/k$, $k = 2, 3, \ldots$) for combining models, which is simply taking the average of the two outputs. Here we are trying to use networks output statistics on the in-sample data to maximize the accuracy of the linear combination. Obviously one can put other accuracy measures in calculations, but that restricts the problem and makes it difficult to handle with linear optimization techniques. Using the reformulation, the problem reduces to a linear programming problem. This type of modeling falls in the subject of so-called data envelopment analysis which appears in different areas of science addressing similar problems [31].

More formally, let $\mu_{i,p}$ and $\mu_{i,n}$ for $i = 1, 2$ be the mean of the outputs of the SRMAR network ($i=1$) and MAR network ($i=2$) on positive and negative ground truth data respectively. $\sigma_{i,p}, \sigma_{i,n}, i = 1, 2$ represents the corresponding standard deviations for SRMAR and MA networks. Let $\psi_1$ and $\psi_2$ be the outputs of SRMAR and MAR networks respectively with values in $[-1, 1]$, taking weighted average of the two as an output as stated in the following,

$$y = w_1 \psi_1 + w_2 \psi_2$$  \hspace{1cm} (5)

The goal is to calculate $w_1$ and $w_2$ based on statistics of the outputs of the two networks so that $y > 0$ indicates possessing the attribute and $y \leq 0$ otherwise. $w_1$ and $w_2$ need to hold some relations. Lets denote networks output by $\psi^p_1$ and $\psi^p_2$ to indicate that the input possesses the attribute (denoting by “p”) and does not (denoting by “n”) respectively, then the direct reformulation of the $y > 0$ (corresponding with the first) and $y \leq 0$ (corresponding with the second) leads to:

$$\psi^p_1 w_1 + \psi^p_2 w_2 > 0,$$
$$\psi^n_1 w_1 + \psi^n_2 w_2 \leq 0,$$

while maximizing the number of correct predictions. We can calculate the confidence interval of the normalized coefficients as follows: $\psi^p_1 \in (\mu_{1,p} - \epsilon_1, \mu_{1,p} + \epsilon_1)$, $\psi^p_2 \in (\mu_{2,p} - \epsilon_2, \mu_{2,p} + \epsilon_2)$, $\psi^n_1 \in (\mu_{1,n} - \epsilon_1, \mu_{1,n} + \epsilon_1)$ and $\psi^n_2 \in (\mu_{2,n} - \epsilon_2, \mu_{2,n} + \epsilon_2)$. Here $\epsilon_i = Z_{0.1}^{0.5} \sqrt{\frac{N}{Z_{0.1}^2}} Z_{0.01}^i$ for $i = 1, 2$ is the normalization of outputs and $Z_{0.1}$ stands for normal distribution with mean 0 and variance 1 and $N$ is number of samples (here number of in-sample data), $\sigma$ stands for the corresponding distribution standard deviation.

In our case we set $W_i = \{0, 0.01 \times j, j = 1, 2, \ldots, 100\}$ for $i = 1, 2$ and therefore get finite feasible region $W_1 \times W_2$ by which we calculate the target function by selecting the pair ($w_1, w_2$) that lead to the highest accuracy.As the size of training set is large, to solve the resulted integer programming we use the python wrapper of the integer programming solver SCIP [32]. The proposed architecture of this linear combination strategy is demonstrated in Fig.3.

Note that the proposed method will perform at least as better as simple average weights as the solution $w_1 = 1/2$, $w_2 = 1/2$ is already inside the feasible region of the linear programming model. As demonstrated in the
experiments section, this combination strategy have led to better results compared to using a single model.

IV. EXPERIMENTS

A. DATASETS

We carry out our experimental evaluation on two widely used benchmark datasets Market-1501 [9], and DukeMTMC-reID [10]. Market-1501 [9] is one of the largest person re-ID datasets, containing 32,668 images and 3,368 query images. 751 identities (19,732) are used for training and 750 identities (13,328 images) are for testing [8]. There are 27 attributes provided in the dataset. Following [8], we work over 11 attributes where some are taken as the average of the related attributes. The attributes were annotated in the identity level. In this dataset, background and junk images are not considered during training or testing since they do not have the corresponding attribute labels. The second dataset is DukeMTMC-reID dataset [10]. It contains 702 identities (16,522 images) for training and 702 identities (19,889 images) for testing. 23 attributes are covered in this dataset.

The aforementioned datasets are used in this paper because we can reproduce the results of a state-of-the-art model that already trained on these datasets [8], and furthermore, these datasets are known to be among the most referred datasets in multi-attribute recognition research.

B. TRAINING DETAILS

In case of multi-attribute recognition, feature extraction part is followed by M small sub-nets, each constructed by a convolutional layer, a Pooling layer followed by a fully-connected layer, and finally, there is a Softmax layer to get classification probability. We use the original backbone network (ResNet-50) [30], which was pre-trained on ImageNet. In the training stage of the MAR network, the loss function is as stated in the 3. For SR network which has a kind of encoder-decoder architecture, Eq.1 is taken as the loss function. In training SR and MAR, Adam optimizer is used as the optimizer. When training SRMAR model, in the SR part, all layers except the last 15 layers were frozen to not letting them to be trained on the multi-attribute data. Our preliminary experiments indicate that freezing more layers leads to poor results while freezing fewer layers causes overfitting. The last 15 layers is selected based on the outcome of the different trials. During training SRMAR network, DiffGrad optimizer [33] with a cyclic learning schedule is used to optimize combined loss function (Equation 4). The maximum size of images in selected datasets is 64 \times 128; therefore, we consider this as the reference size. Paying attention to the input image sizes of SR models EDSR (2x, 3x, and 4x) and DBPN (2x, 4x, and 8x), the images were downsampled with bi-cubic interpolation into four sizes: 32 \times 64, 21 \times 42, 16 \times 32, and 8 \times 16 respect to SR model input sizes. The batch size is taken as 32 in all experiments and the networks were trained for 40-60 epochs. For training, the number of epoch to 60 epoch, batch size to 32, with initial learning rate to 0.001, with a learning scheduler multiplies the learning rate by 0.1 every five epochs.

C. EXPERIMENTAL RESULTS

The SRMAR model is evaluated via extensive experiments over the two benchmark datasets. There are two versions, SRMAR-E uses EDSR [26] SR model within the joint network, whereas SRMAR-D uses DBPN [27] SR model. The proposed SRMAR model is compared with the MAR [8] model that is applied to the same resolution images. Several resolutions such as (32 \times 64), (21 \times 42) and (16 \times 32 or 8 \times 16) are evaluated in the experiments. The number of total attributes is 30 and 23 for the Market-1501 [9] and the DukeMTMC-reID [10] datasets, respectively. For the sake of representation, we summarized them into 11 (Market-1501) and 10 (DukeMTMC-reID) attributes by averaging similar attributes that belong to one category (such as upper body colors or lower body colors). The rightmost column represents the average accuracy for each method processing over the presented image resolution.

In Table 2, the results of the SRMAR-E method over the DukeMTMC-reID dataset is shown. According to Table 2, for all the resolutions, the SRMAR-E model improves the recognition performance of the MAR model significantly. For 16 \times 32 resolution input size, the original model without any SR component achieves an accuracy of 68.78\%, whereas the proposed SRMAR-E model achieves 83.22\% accuracy. The accuracy is even more improved to 84.37\% when the MAR network is combined with the SRMAR network using the proposed linear combination strategy (Combined-E). Similarly, for the input size 21 \times 42, the proposed SRMAR-E model improves the accuracy of the MAR [8] model from 74.98\% to 82.14\%, and the linear combination of the two models (Combined-D) achieve an accuracy of 85.58\%. For the 32 \times 64 input size, even though the improvements are not that drastic, still the accuracy improve from 77.21 \% to 82.89\%.

Table 3 presents similar experiments using the SRMAR-D (that utilizes the DBPN model as the SR component). We observe that the SRMAR model improves over MAR [8] from 77.21\% to 79.71\% for large input, from 74.98\% to 79.91\% for medium input, and from 65.45\% to 80.16\% for small input. Like the SRMAR-E, we observe that the relative improvement in accuracy for small size input (8 \times 16) is more than the other two input sizes. Also, we observe that the linear combination of two models is again effective in increasing the overall accuracy even further.

For the DukeMTMC-reID dataset, we observe a similar trend in our experiments. Table 4 and Table 5 presents the corresponding results. In Table 4, the results of the SRMAR-E model are presented. The overall accuracy is increased from
TABLE 2. Accuracy results on Market-1501 dataset using EDSR SR model.

| model | img size | gender | age | hair.l | slv.l | ll.cloth | tl.cloth | b.pack | h.bag | bag | co.up | co.low | mean |
|-------|----------|--------|-----|--------|-------|----------|----------|--------|-------|-----|-------|--------|------|
| MAR [8] | 16 × 32 | 56.12 | 74.27 | 63.11 | 80.34 | 62.23 | 53.51 | 64.05 | 86.27 | 55.57 | 77.63 | 73.62 | 86.78 |
| 32 × 64 | 56.12 | 75.54 | 63.93 | 93.52 | 67.52 | 88.00 | 74.91 | 90.57 | 71.54 | 85.12 | 82.57 | 77.21 |

TABLE 3. Accuracy results on Market-1501 dataset using DBPN SR network.

| model | img size | gender | age | hair.l | slv.l | ll.cloth | tl.cloth | b.pack | h.bag | bag | co.up | co.low | mean |
|-------|----------|--------|-----|--------|-------|----------|----------|--------|-------|-----|-------|--------|------|
| MAR [8] | 8 × 16 | 54.20 | 71.2 | 62.52 | 78.91 | 62.12 | 53.01 | 47.80 | 73.7 | 64.06 | 76.31 | 78.61 | 65.45 |
| 32 × 64 | 56.12 | 75.54 | 63.93 | 93.52 | 67.52 | 88.00 | 74.91 | 90.57 | 71.54 | 85.12 | 82.57 | 77.21 |

TABLE 4. Accuracy results on DukeMTMC-reID dataset using EDSR SR model.

| model | img size | w.hat | boots | l.up | b.pack | h.bag | bag | co.shoes | co.up | co.low | mean |
|-------|----------|-------|-------|------|--------|-------|-----|----------|-------|--------|------|
| MAR [8] | 16 × 32 | 53.33 | 75.77 | 76.24 | 84.32 | 51.90 | 75.78 | 82.91 | 78.21 | 80.14 | 76.75 | 71.71 |
| 32 × 64 | 58.81 | 75.01 | 77.41 | 87.74 | 54.86 | 92.84 | 83.53 | 88.00 | 90.51 | 82.27 | 79.09 |

TABLE 5. Accuracy results on DukeMTMC-reID dataset using DBPN SR model.

| model | img size | w.hat | boots | l.up | b.pack | h.bag | bag | co.shoes | co.up | co.low | mean |
|-------|----------|-------|-------|------|--------|-------|-----|----------|-------|--------|------|
| MAR [8] | 8 × 16 | 42.37 | 55.31 | 76.14 | 75.90 | 54.91 | 73.48 | 79.21 | 56.27 | 80.07 | 76.54 | 68.02 |
| 16 × 32 | 53.33 | 75.77 | 76.24 | 84.32 | 51.90 | 75.78 | 82.91 | 78.21 | 80.14 | 76.75 | 71.71 |
| 32 × 64 | 58.81 | 75.01 | 77.41 | 87.74 | 54.86 | 92.84 | 83.53 | 88.00 | 90.51 | 82.27 | 79.09 |

79.09% to 83.05% for 32 × 64 sized input, from 77.05% to 81.94% for 21 × 42 sized input, and from 71.71% to 81.57% for input sizes 16 × 32. Again, the improvements are quite significant, especially when used with lower resolution images. As shown in Table 4, the proposed linear combination of MAR [8] and SRMAR-E models are also effective, offering a performance improvement more than 1% over SRMAR-E model.
Table 5 shows the performance of the proposed SRMAR-D model, and the following improvements are achieved over the reference MAR model [8]: for 32 × 64 size input, the overall accuracy has increased from 79.09% to 79.34%, for 16 × 32 sized input, the accuracy has increased from 77.05% to 79.40%, and for the 8 × 16 input from 68.02% to 78.98%. The same pattern is observed in the value of improvements w.r.t sizes; the smaller size, the better improvement.

From the results in Table 2-5, the performance of individual attributes can further be evaluated. In the case of the Market-1501 dataset, Table 2, the attribute that has the lowest recognition rate is the gender attribute, whereas the long sleeve (l.slv) attribute seems to be the attribute with the highest recognition accuracy for the MAR model [8] in different resolutions. SRMAR-E model improves the recognition accuracy of gender attribute significantly. The performance improvement is also remarkable for the age, style of clothing (ll.clth), color of up clothing (c.up) and performance improvement is also remarkable for the age, recognition accuracy of gender attribute significantly. The highest recognition accuracy for the MAR model [8] in different resolutions. SRMAR-E model improves the recognition accuracy of gender attribute significantly. The performance improvement is also remarkable for the age, style of clothing (ll.clth), color of up clothing (c.up) and color of down clothing (c.down) attributes. For some of the resolutions, especially for 21 × 42 and 32 × 64 input sizes, for ”bag” and ”hair” attributes, there is a reduction in accuracy in Table 2. In such cases, using the linear combination of these models as proposed helps. As can be seen, the linear combination strategy resolves accuracy reduction of the SRMAR-E for ”bag” attribute from 69.51% to 73.76% for 32 × 64 sized input, from 62.23% to 74.21% for 21 × 42 sized input, and from 70.93% to 76.32% for 16 × 32 sized input.

Table 6 summaries the experimental results for 16 × 32 resolution input size over both datasets. As can be seen from this table, SRMAR-E model that uses the EDSR [26] SR model performs better than the SRMAR-D model that uses DBPN [27] SR model on both of the benchmark datasets. Moreover, the proposed linear combination or SRMAR and MAR models offer a notable increase in the accuracy of both SRMAR-(E, D). Our observations suggest that combining both SRMAR and MAR models in the proposed way offers the best recognition performances for the recognition of person attributes in LR images. Based on these results, linear combination is superior to using individual models.

In addition, we have shown that improving the person-attribute recognition accuracy on LR images can be achieved by the help of SR models, and further improvement can be obtained by combining the models.

**V. CONCLUSION**

Identifying personal attributes in LR images is a crucial task for the surveillance applications, however, this task is rarely addressed in the literature. In order to fill this gap, this paper evaluates the effects of the SR CNN architectures on improving MAR performance in LR images. To this end, the paper adapts one of the state-of-the-art models proposed for MAR and two different SR network architectures and then construct a combined architecture entitled SRMAR-(E,D). To the best of our knowledge, this research is the first combined learning model for multi-attribute recognition in low resolution images. We also propose a linear combination scheme to combine the proposed SRMAR network with the base multi-attribute recognition network.

The experiments are carried out in two benchmark datasets that have been proposed for person attribute recognition. The experimental results demonstrate that, for the input images in low resolution, the proposed end-to-end convolutional architecture is successful in improving the recognition performance of the base model for person attribute recognition. The performance improvement is remarkable especially for the small-sized inputs, whereas the joint learning of SR and MAR networks contributes to the recognition performance in both small and larger resolutions. The proposed linear combination scheme, which combines the base model and SRMAR model via statistical properties of the models performance on training set, successfully combines the recognition power of both models.

**REFERENCES**

[1] R. Layne, T. Hospedales, and S. Gong, “Person re-identification by attributes,” in Proc. Brit. Mach. Vis. Conf., vol. 2, 2012, p. 8.

[2] H. J. Galiyawala, M. S. Raval, and A. Laddha, “Person retrieval in surveillance videos using deep soft biometrics,” in Deep Biometrics. Cham, Switzerland: Springer, 2020, pp. 191–214.

[3] Y. Shi, Z. Wei, H. Ling, Z. Wang, J. Shen, and P. Li, “Person retrieval in surveillance videos via deep attribute mining and reasoning,” IEEE Trans. Multimedia, vol. 23, pp. 4376–4387, 2021.

[4] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, “Deep networks for image super-resolution with sparse prior,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 370–378.

[5] M. Irani and S. Peleg, “Super resolution from image sequences,” in Proc. 10th Int. Conf. Pattern Recognit., vol. 2, Jun. 1990, pp. 115–120.

[6] S. Anwar, S. Khan, and N. Barnes, “A deep journey into super-resolution: A survey,” ACM Comput. Surveys, vol. 53, no. 3, pp. 1–34, May 2021.

[7] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution via sparse representation,” IEEE Trans. Image Process., vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[8] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, “Improving person re-identification by attribute and identity learning,” Pattern Recognit., vol. 95, pp. 151–161, Nov. 2019.

[9] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1116–1124.

[10] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by GAN improve the person re-identification baseline in vitro,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3754–3762.

[11] Y. Deng, P. Luo, C. C. Loy, and X. Tang, “Pedestrian attribute recognition at far distance,” in Proc. 22nd ACM Int. Conf. Multimedia, Nov. 2014, pp. 789–792.

[12] D. Li, X. Chen, and K. Huang, “Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios,” in Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPRA), Nov. 2015, pp. 111–115.

[13] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li, “Multi-label CNN based pedestrian attribute learning for soft biometrics,” in Proc. Int. Conf. Biometrics (ICB), May 2015, pp. 535–540.
[14] Y. Shi, H. Ling, L. Wu, J. Shen, and P. Li, “Learning refined attribute-aligned network with attribute selection for person re-identification,” *Neurocomputing*, vol. 402, pp. 124–133, Aug. 2020.

[15] J. Wu, H. Liu, J. Jiang, M. Qi, B. Ren, X. Li, and Y. Wang, “Person attribute recognition by sequence contextual relation learning,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3398–3412, Oct. 2020.

[16] R. Tsai and T. Huang, “Multiframe image restoration and registration,” *Adv. Comput. Vis. Image Process.*, vol. 1, pp. 317–339, Jan. 1984.

[17] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.

[18] C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network,” in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 391–407.

[19] D. Zhou, R. Duan, L. Zhao, and X. Chai, “Single image super-resolution reconstruction based on multi-scale feature mapping adversarial network,” *Signal Process.*, vol. 166, Jan. 2020, Art. no. 107251.

[20] N. C. Rakotonirina and A. Rasoanaivo, “ESRGAN+: Further improving enhanced super-resolution generative adversarial network,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3637–3641.

[21] P. Shamsolmoali, M. Zareapoor, R. Wang, D. K. Jain, and J. Yang, “G-GANISR: Gradual generative adversarial network for image super resolution,” *Neurocomputing*, vol. 366, pp. 140–153, Nov. 2019.

[22] C. Fang, Y. Zhu, L. Liao, and X. Ling, “TSGAN: Real-world text image super-resolution based on adversarial learning and triplet attention,” *Neurocomputing*, vol. 455, pp. 88–96, Sep. 2021.

[23] J. He, J. Zheng, Y. Shen, Y. Guo, and H. Zhou, “Facial image synthesis and super-resolution with stacked generative adversarial network,” *Neurocomputing*, vol. 402, pp. 359–365, Aug. 2020.

[24] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[25] J. Yamanaka, S. Kuwashima, and T. Karita, “Fast and accurate image super resolution by deep CNN with skip connection and network in network,” in *Proc. NIPS*. Cham, Switzerland: Springer, 2017, pp. 217–225.

[26] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.

[27] M. Haris, G. Shakhnarovich, and N. Ukit, “Deep back-projection networks for super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.

[28] S. Nah, T. H. Kim, and K. M. Lee, “Deep multi-scale convolutional neural network for dynamic scene deblurring,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3883–3891.

[29] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, “Two deterministic half-quadratic regularization algorithms for computed imaging,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 2, Sep. 1994, pp. 168–172.

[30] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[31] K. C. Land, C. A. K. Lovell, and S. Thore, “Chance-constrained data envelopment analysis,” *Managerial Decis. Econ.*, vol. 14, no. 6, pp. 541–554, Nov. 1993.

[32] G. Garrard, R. Anderson, K. Bestuzheva, W.-K. Chen, L. Eifler, M. Gasse, and P. Gemander, “The SCIP optimization suite 7.0,” *Zuse Inst. Berlin, Berlin, Germany*, Tech. Rep. 20-10, Mar. 2020.

[33] S. R. Dubey, S. Chakraborty, S. K. Roy, S. Mukherjee, S. K. Singh, and B. B. Chaudhuri, “DiffGrad: An optimization method for convolutional neural networks,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4500–4511, Nov. 2020.

**RAMIN ABBASZADI** received the B.S. degree in computer science from the University of Tabriz, Iran, and the M.S. degree in information technology of double programming from the Azerbaijan University of Architecture and Construction, Baku, in 1998, and the University of Tehran, Iran, in 2009. He is currently pursuing the Ph.D. degree in computer engineering with Hacettepe University, Ankara, Turkey. He has worked as a Lecturer and Network Administrator at Urmia University for six years. He works as a Lead Software Engineer at the AI Team, Ekin Safe City Technologies, Ankara. His research interests include computer vision, deep learning, and machine learning.

**NAZLI IKIZLER-CINBIS** received the B.Sc., M.Sc., and Ph.D. degrees from the Department of Computer Engineering, Bilkent University, From 2005 to 2006, she was a Visiting Scholar at the University of Illinois at Urbana–Champaign (UIUC). She worked as a Postdoctoral Research Associate at Boston University, Boston, MA, USA. Since 2011, she has been a Faculty Member at Hacettepe University. Her research interests include computer vision and machine learning, specifically focusing on video processing, human action and interaction recognition, and zero-shot learning.