



# GUV-Net for high fidelity shoeprint generation

Muhammad Hassan<sup>1,5</sup> · Yan Wang<sup>1</sup> · Wei Pang<sup>2</sup> · Di Wang<sup>3</sup> · Daixi Li<sup>4</sup> · You Zhou<sup>1,5</sup> · Dong Xu<sup>6</sup>

Received: 9 June 2021 / Accepted: 24 September 2021 / Published online: 21 October 2021  
© The Author(s) 2021

## Abstract

Shoeprints contain valuable information for tracing evidence in forensic scenes, and they need to be generated into cleaned, sharp, and high-fidelity images. Most of the acquired shoeprints are found with low quality and/or in distorted forms. The high-fidelity shoeprint generation is of great significance in forensic science. A wide range of deep learning models has been suggested for super-resolution, being either generalized approaches or application specific. Considering the crucial challenges in shoeprint based processing and lacking specific algorithms, we proposed a deep learning based GUV-Net model for high-fidelity shoeprint generation. GUV-Net imitates learning features from VAE, U-Net, and GAN network models with special treatment of absent ground truth shoeprints. GUV-Net encodes efficient probabilistic distributions in the latent space and decodes variants of samples together with passed key features. GUV-Net forwards the learned samples to a refinement-unit proceeded to the generation of high-fidelity output. The refinement-unit receives low-level features from the decoding module at distinct levels. Furthermore, the refinement process is made more efficient by inverse-encoded in high dimensional space through a parallel inverse encoding network. The objective functions at different levels enable the model to efficiently optimize the parameters by mapping a low quality image to a high-fidelity one by maintaining salient features which are important to forensics. Finally, the performance of the proposed model is evaluated against state-of-the-art super-resolution network models.

**Keywords** Shoeprint · Super-resolution · Forensics · GAN · VAE · U-Net · Infusion

✉ You Zhou  
zyou@jlu.edu.cn

✉ Dong Xu  
xudong@missouri.edu

Muhammad Hassan  
hassan2117@mails.jlu.edu.cn; mhassandev@gmail.com

Yan Wang  
wy6868@jlu.edu.cn

Wei Pang  
w.pang@hw.ac.uk

Di Wang  
WangDi@ntu.edu.sg

Daixi Li  
lidaixi@everspry.com

<sup>1</sup> College of Computer Science and Technology, Jilin University, Changchun, China

<sup>2</sup> School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK

<sup>3</sup> Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly, Nanyang Technological University, Singapore, Singapore

## Introduction

Shoeprints are often found in crime scenes with poor quality images, and they have a critical role in podiatry investigations. Examining shoeprints becomes more challenging with limited available data, poor information contents, unavailability of ground truth and higher quality images, partial and incomplete prints, and most importantly, the lack of domain-specific processing algorithms [2,10,46,69,79,104]. Considering the aforementioned challenges, a higher resolution (HR) shoeprints with reduced noise and enhanced quality have great importance for forensic purposes. High quality (HR) shoeprints generation from their lower resolution (LR) counterparts offer high density, close looks,

<sup>4</sup> Everspray Science and Technology Company Ltd., Dalian, China

<sup>5</sup> Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Changchun, China

<sup>6</sup> Department of Electrical Engineering and Computer Science, Bond Life Sciences Center, University of Missouri, Columbia, MO, USA

and detailed observations and therefore, they can facilitate the analysis of the original prints acquired in the crime scene. However, generating high quality shoeprints via latest deep learning (DL) algorithms need special treatment of LR images where the corresponding HR images are unavailable or insufficient. Recently, deep learning (DL) provides a wide range of algorithms to the generation of high fidelity images with the availability of HR-LR image pairs [15–17,27,41,47,48,74,76,92]. Moreover, the realistic scenarios lack perfect shoeprints to train deep learning models for HR image generation. To better address the aforementioned challenges, we propose an end-to-end deep learning model for generating high-fidelity shoeprints having no HR images.<sup>1</sup> To the best of our knowledge, our proposed model is first-of-its-kind to recover HR shoeprint images from their LR ones.

Super-resolution (SR) is the generation of HR images from their LR counterparts, and it has been studied from using early convolution networks [15,16] to employ the recent Generative Adversarial Networks (GAN) based models [50]. SR learning strategies can be divided into interpolation-based [43,100], learning-based [15,97,98], and reconstruction-based [80]. SR includes the generation of single-LR or multi-LR images [61,99]. Single-image-super-resolution (SISR) [101] is an ill-posed reverse problem in which a high-fidelity image is generated from its lower LR image, and SISR is also attempted in the this study. For HR image generation via SISR, the input LR image can be passed through a model featured with either pre-upsampling [15,16,41], post-upsampling [17,76], progressive upsampling [47,48,92], or hourglass [27,52] learning approaches. The pre-upsampling methods are simpler as they only require prior interpolation, but they may circumvent checkerboard artifacts, amplify noise and blurriness, and cause expensive computation [42,77,84,85]. Similarly, post-upsampling methods can reduce the computational complexity by extracting the features at lower dimensional space, but the upsampling process in single step increases the learning difficulty for large scaling factors (i.e.,  $4\times$ ,  $8\times$ ) [50,53,89]. The progressive upsampling reduces the learning difficulty and improves performance, but it may cause training instability due to complicated model structure [47,48,92]. Similarly, the iterative up-and-down methods can better mine the relationship between the given LR and generated HR image pairs with high quality reconstruction results [27,52,93]. The iterative up-and-down approaches adapt the deconvolution (transpose) operation instead of upsampling to overcome high computational complexity and maintain accuracy [17]. In addition, there are some algorithms specialized to sports and medical images [26,33,38], surveillance and security sys-

tems [66,103], faces [87,96], scenes [81], arts [54], and so on. Most of these areas provide both high and low quality images for training DL models.

To the best of our knowledge, the proposed model (GUV-Net, see Fig. 1) is the first attempt to generate high-fidelity shoeprints from their low quality images with no original ground truths (HRs) in the database [28]. Thus, we interpolates (bicubic-interpolation) the original shoeprint images to HR images, which enables the model to learn the generation of good quality images while training. GUV-Net downsamples and upsamples a given LR image and enable the model to pass on low-level features to high level (high-resolution, refinement) learning. GUV-Net utilizes all possible learning phases to generate high-fidelity images. Overall, the model compresses the given input into a latent space, learns probabilistic distributions, generates distinct sample variants in a controlled way at the decoding side, passes the key features from encoder to decoder and then from decoder to refinement-phase (high spatial phase). Finally, the model discriminates high quality generated images following adversarial and other objective functions. GUV-Net borrows features from the GAN, VAE, and U-Net architectures and fuses them in an efficient way.

In addition to wide applications of GAN model [105], GAN has also been used for SISR generation. GAN based SISR generation is often visually pleasing but the generated HR images may contain fake details and textures, which deviate from the ground truths [89]. The undesirable generation in SR is caused by the inversion (taking LR from HR and generating back into HR), in which the critical spatial information may not be kept faithfully at the low dimensional space (latent space) to recover back image both at pixel and semantic levels [106]. GANs face the problem of complex distributions of images and depends on extensive high-fidelity data, which may make the convergence hard, the model difficult to optimize hyper-parameters, unstable to train, and sometimes GAN may produce visually absurd outputs [4,29]. Despite the widespread use of GAN for high resolution images, the generated samples often do not fully capture the diversity of true distributions [12,67], which make inadequate the solely GAN based model to the generation of high fidelity shoeprints.

On the other hand, variational autoencoder (VAE) follows the maximum likelihood principle with an encoder–decoder structure, compresses input to latent space, which can be more efficiently optimized compared to pixel space [67]. The inherent mathematical formulation in VAE makes it relatively cheap and stable to train [29]. The negative log-likelihood in the VAE objective function enables VAE to generalize well to unseen data and cover all modes of data so that mode collapse as observed in GAN can be avoided [67,88]. In contrast to GAN for SR applications, VAE based models are more in control during training and generating

<sup>1</sup> Note the interpolated HR images are merely used for the optimization of refinement-phase in terms of inverse-encoding.

samples (i.e., beneficial for shoeprint generation with respect to forensics) but may output blurry results [18,49,90]. In VAE, the encoding distribution regularizes and matches the latent space for LR and HR images and ensures the generative process to recover the missing information [35,67]. In addition to the GAN and VAE models, our model also infuses the U-Net-like [70] features to carry out the shoeprint super-resolution image generation from the challenging poor quality data. The introduction of U-Net enables the model to pass on spatial features from the compressing module to the decompressing module via skip-connections to maintain contextual features. In GUV-Net, similar skip-connections are further extended to pass low level features into high dimensional space (i.e., to refinement phase).

Based on the aforementioned information, the intended model addresses the critical challenges faced in the generation of high-fidelity shoeprints from their poor quality images with no ground-truths (HR images). Thus, combining the good features of GAN, VAE, and U-Net models may be a reliable way to generate the desired output. For this purpose, GAN and VAE are infused by overcoming the downsides of GAN (i.e., training and convergence instability [49], sensitive to hyper-parameters [55], mode collapse [4,59]) and VAE (i.e., blurriness [18,49,90] and over smoothness [49] of the generated images) for high-fidelity image generation [29,32,40,49,72]. Hence, the high quality images can be generated through training VAE in an adversarial manner [29] by imposing a discriminator in the data space [49]. The infusion of VAE into GAN brings training stability and optimization in the manifold of latent-space structure [71,72]. The sampling representation at the latent space of VAE can be utilized as a generator in GAN [9,91]. Similarly, multi-scale structure similarity (MS-SSIM) and  $L_1$  norm have been attempted to overcome the blurriness in VAE [90]. For latent space optimization, the posterior and prior distributions can be discriminated in order to generate a high quality images [57].

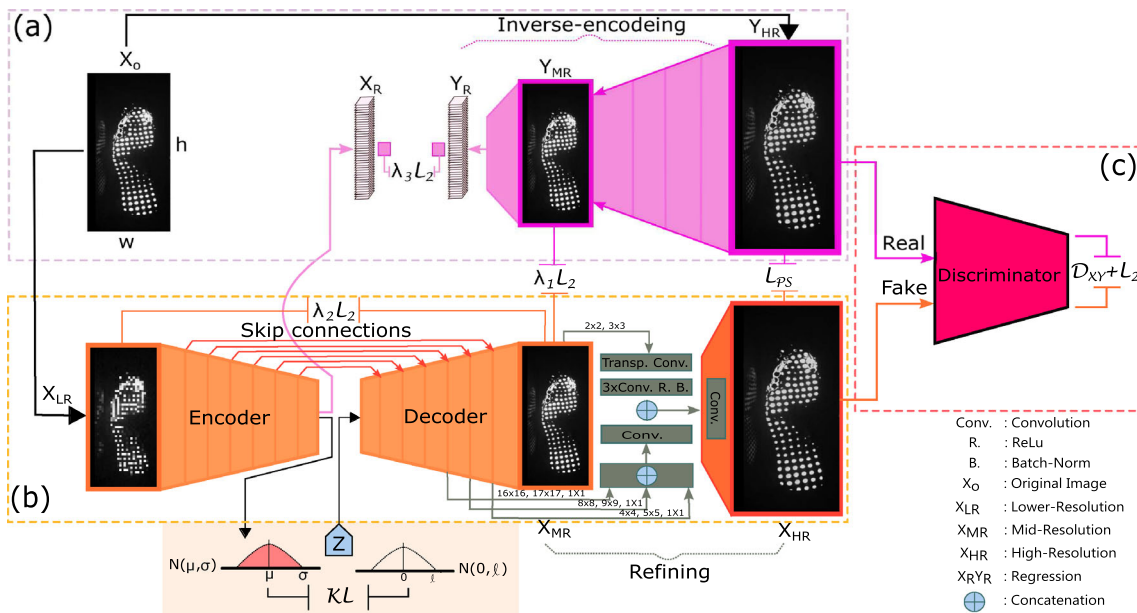
However, the infusion of GAN and VAE still has the following issues: poor scalability in high dimensions, limitations in scaling to high dimensions, expensive evaluation, variational inferencing, lack of distributions matching both in latent and visible spaces, and limited improvement in the quality of generated images [71]. Both VAE and GAN models for high fidelity image generation may synthesize unnecessary patterns [7], which lose the importance for producing high quality shoeprints with respect to forensics. To avoid undesirable patterns and texture in HR shoeprints, the infusion of U-Net into the generative models may control and guide high fidelity image generation [20]. As U-Net facilitates the preservation of spatial information [75], thus the U-Net equipped models control the learned features to generate the desired image against LR image. Considering the pros-and-cons of GAN, VAE, and U-Net architectures, the proposed GUV-Net model infuses the features of GAN,

VAE, and U-Net into a single DL model (Fig. 1), which is trained in an end-to-end fashion, and thus high quality shoeprints can be generated from their lower-quality counterparts with no perfect/ground truth images. The model optimizes the distribution in latent space and synthesize the guided shape image generation by avoiding unnecessary patterns in shoeprints. Similarly, the model overcomes the blurriness result by conditioning the decoder part with the reduced version (inverse-encoding unit) of interpolated HR input ( $Y_{HR}$ ). With the infusion of U-Net, the model can better control and preserve the spatial information to output the desired variant against LR image. In addition, the discriminative property borrowed from GAN also overcomes the deficiencies (blurriness) caused by the reconstruction-loss ( $L_2$ ) and generates high quality shoeprints with better perceptual ability. The model maintains training stability to generate HR ( $X_{HR}$ ) images learning from nice latent manifold structure together with skip spatial information. The generated  $X_{HR}$  images (fake) are then put in the adversarial training against the corresponding interpolated ones ( $Y_{HR}$ ) (Fig. 1). Overall, the model generates the coarse image ( $X_{MR}$ ) in the encoder–decoder structure up to the same level as  $X_{LR}$  with dimension  $h = h/n$  and  $w = w/n$  (where  $h, w$  are the height and width of  $X_{LR}$  obtained from  $X_o$ ), and the high-fidelity output ( $X_{HR}$ ) is obtained in the refinement unit. Different loss terms are adopted at distinct levels to facilitate the generation of high resolution shoeprints.

There are a wide range of specialized and generalized applications of SR; however, to the best of our knowledge, this study is the first attempt to address the challenges in the super-resolution tasks for shoeprint images. The rest of the paper is organized as follows: “Literature study” provides related work in shoeprint based processing and infusion models of GAN, VAE, and U-Net. “Methodology” presents the methodology for GUV-Net structure and training, and “Results evaluation and analysis” analyzes the derived results, which are followed by the conclusion and future directions in “Conclusion and future directions”.

## Literature study

Shoeprint images have been studied in many areas, including forensic podiatry [45,86,104], biological traits examination and investigation [62], gender prediction [5,8,63], and body morphological studies [95]. A number of operations underlying shoeprints have been performed including retrieving, recognition, pattern matching via different approaches, and these operations have been performed by many approaches, including manual [3], semi-automated [2,24], automated [1,68], and machine learning (ML) (in particular DL methods) [15,97,104]. Among these, DL methods have shown promising results in shoeprint related operations [14,22,56,104].



**Fig. 1** GU\_V\_Net receives  $X_{LR}$  (height =  $h/n$ , width =  $w/n$ ) shoeprint image obtained from original shoeprint  $X_o$  (height =  $h$ , width =  $w$ ) and generates an  $X_{HR}$  (height =  $h \times n$ , width =  $w \times n$ ). The model trains with generative loss-terms including  $L_{PS}$ ,  $\lambda_i L_2$ , where  $i = 1, 2, 3$ , and KL-divergence. Similarly, the discriminative loss-term composed of  $D_{XY}$  and  $L_2$ , where the details can be seen in the Section of loss-function. **a)** Prior to training, both  $X_{LR}$  and  $Y_{HR}$  images are generated through interpolation methods. **b)** The network receives  $X_{HR}$  as input, extracts features and encodes into latent representation (Encoder), learns probabilistic distribution (VAE), and passes (skip-connections)

spatial features into  $X_{MR}$  together with decoded layer. The model samples from the learned latent space and further optimize the weights at mid-level ( $X_{MR}$ ) against the interpolated image ( $Y_{MR}$ ) as ground truth. The  $X_{MR}$  version is further processed for refinement process together with passed features from the decoder at distinct levels. From the refinement stage, the model generates high-fidelity ( $X_{HR}$ ) images by optimizing the weights against  $Y_{HR}$  in high dimensional space. The generative module of GU\_V\_Net only be used for evaluation purposes. **c)** Discriminator resolves the real image ( $Y_{HR}$ ) obtained by interpolation and fake image ( $X_{HR}$ ) output by the generator

Regarding shoeprint enhancement, there are some conventional approaches [21]; however, DL algorithms for generating super resolution version of the low quality shoeprints are lacking.

A variety of models for Super resolution-SR have been proposed, starting from the early convolution neural networks [15,16] to the latest GAN based networks [50]. To generate high-fidelity HR images, the given HR images are downsampled via interpolation into LR images and then mapped back to the HR ones using different learning approaches [15–17,41,47,48,76,92]. Some of these approaches first upsample (pre-upsampling) the LR images to the HR space, and then CNN learns in the HR space to refine the coarse images [42,77,84,85]. Such methods are simple but may amplify checkerboard artifacts, noise, and blurriness. To reduce the complexity, feature extraction can be performed at lower dimensional space (post-upsampling) and then upscaled to HR space either with interpolation or with transpose-convolution learning [17,50,53,89]. To reduce the complexity in HR space, a progressive learning based strategy have been adopted by the use of cascade CNNs to generate HR images at smaller factors [47,48,92]. However, more guidance and advanced training are needed for a com-

plicated model designing to avoid instability during training. To better mine the mapping between LR–HR pairs of images, iterative up and down sampling methods are applied with effective learning to provide high quality images [27,52,93]. Such methods mostly adopt the encoder–decoder like structure to mine the non-linear relationship between the LR–HR image space. Moreover, the aforementioned methods are trained with the availability of HR versions. However, in our case, the available original shoeprint images are not in good quality and vulnerable to noise and distortion. Bear in mind the above challenges and network architecture designs, our model (GU\_V-Net) adopts feature extraction and learns both at lower dimensional space and latent space (i.e., at encoder–decoder structure), and at high dimensional space through post upscaling via transpose convolution. Thus, GU\_V-Net infuses the features from variational autoencoder into a U-Net-like structure and append with adversarial as in GAN.

A wide range of deep learning models has been proposed to address the challenges in SR [43,61,99,100] adopting either GAN [29,32,40,49,72] or VAE [20] for the generative purpose. However, SR generation with GAN [4,11,67] or VAE [18,49,90] in isolated forms faces challenges such as training in stability, sensitivity to the nature of datasets and

the low quality of output result. To compensate the downsides of generative models, several attempts have been made to infuse GAN with VAE [4,29,49,55,59,90], and VAE with U-Net [20] for output high-fidelity images. However, the existing infused forms of GAN and VAE still have limitations in terms of training complexity and generating high quality images [71], and lacking of approaches with no ground truths (HR images). Furthermore, VAE is also infused into U-Net structure for guided shape and controlled image generation with desired variants of the queried image [20]. For images like shoeprints which are vulnerable to noise and hard to find the cleaned and good quality ground truths, the infusion of GAN, VAE, and U-Net architectures with their positive aspects is expected to more effectively generate high resolution output retaining the original patterns and textures. For this purpose, the proposed model adopts the generative features from VAE with a control structure and passing. Moreover, the adoption of VAE enables the optimal and controlled image synthesis at latent space and stabilizes the HR shoeprint generation as vital to forensics. The adoption of adversarial learning from GAN structure encourages the model to synthesis a high quality realistic version of the given shoeprint.

## Methodology

SR needs both low and high resolution images to train an end-to-end deep learning model. In most cases, the LR images are obtained from their HR counterparts through different degradation methods including interpolation, noise, blur, and so on [89]. The benchmark datasets provide both LR and HR pairs of images while some only provide HR images [93]. The state-of-the-art (SOTA) models then use the desired scaling factor to downsample the HR to LR images [17,89,93]. However, the unavailability of HR images makes SISR problem more challenging. Moreover, the HR image generation from dataset susceptible to noise which becomes more crucial to model from their LR image. Sometimes, the generation of HR from their LR counterparts with noise and distortion may raise unnecessary regions in HR images (shoeprints) which lose their critical role in the fields such as forensics. Regarding the aforementioned challenges, our proposed model (GUV-Net) provides a deep learning based SR model specialized to shoeprint images. In this section, we describe our network-architecture in details, objective function, and training the understudied models. The network architecture and training details are presented in the following section.

### Network architecture

Our GUV-Net architecture is mainly divided into three units including inverse-encoding or preparation of ground truth

samples (Fig. 1a), generation of fake images via the main network (Fig. 1b), and adversarial learning (Fig. 1c).

### Sample preparation

In normal circumstances, acquiring shoeprint images with high resolution are rare and challenging, especially scanning shoe outsole while stepping on a scanning-machine. Thus, the collected huge amount of dataset is lack of high resolution-HR images and their LR counterparts [28]. Therefore, the original shoeprint ( $X_o$ , with *height* =  $h$  and *width* =  $w$ ) is simultaneously downsampled ( $X_{LR}$ ) and upsampled ( $Y_{HR}$ ) to generate both high-resolution and low-resolution images, respectively. The isolated downsample and upsample may not reflect the real distribution in realistic environment. The overall downsampling and upsampling can be formulated as follows:

$$\begin{aligned} X_{LR} &= \Psi\{X_o; (h/n, w/n)\} \\ Y_{HR} &= \Upsilon\{X_o; (h \times n, w \times n)\} \\ Y_{MR} &= \zeta\{Y_{HR}; \varphi\} \\ Y_R &= \zeta_R\{Y_{MR}; \varphi_R\} \\ &\text{where } X_o \in \mathbb{R}_o^{\text{HWC}}, \end{aligned} \quad (1)$$

where  $\Psi$  and  $\Upsilon$  are the inter-area and bi-cubic interpolations used for LR and HR shoeprint images generation, respectively.  $Y_{MR}$  is obtained at training time by the convolution operation  $\zeta$  over the learning parameters  $\varphi$ . The network further convolves over  $Y_{MR}$  using learning parameter  $\zeta_R$  with linear activation and ends with a regression  $Y_R$ . GUV-Net receives  $X_{LR}$  and compares the middle phase ( $X_{MR}$ ) of the main model with  $Y_{HR}$ , which enables the model to learn progressively at distinct levels. Similarly, during training, the interpolated high-level real image  $Y_{HR}$  convolves to a middle-level mapping ( $Y_{MR}$ ) to optimize the mainstream learning of  $X_{HR}$  generation.  $Y_{HR}$  further contributes to discriminative learning in a similar way as in GAN.

### Variational inferecing and skip-connections

GUV-Net imitates and infuses three popular deep learning structures including VAE, U-Net, and GAN models. The aim of infusing VAE in GUV-Net is to extract features at the compressed level and generate multiple corresponding output among which the closed one will be chosen. The generation via VAE at lower dimensional space reduces computational complexity and learning ability at detail level. Furthermore, the sampling generation of a fake HR shoeprint from the latent space enables the model to avoid the generation of unnecessary regions caused by the inclusion of GAN. Similarly, the skip-connections  $S_f$  incorporated from U-Net enables the model to maintain the salient spatial features

passing from encoder  $e\{X_{LR}; \phi\}$  to decoder  $d\{e \odot S_f; \vartheta\}$ . Both VAE and U-Net share the same encoding and decoding structure in a harmonized way. The network structure of GUV-Net adopts U-Net to maintain necessary key patterns to facilitate the generation of high-fidelity images, where VAE empowers the model to utilize probabilistic distributions to sample and generate variants of shoeprints.

**Encoding:** The infused form of VAE and U-Net receives  $X_{LR}$  (Eq. 1) and maps to a compressed form ( $z_i \in \mathbb{R}^k$  where  $k$  is the dimension of latent space) at the bottleneck (Fig. 1a). The input passes through five blocks where the first three blocks are residual based and the remaining two blocks are stacked convolutional layers. Each block ends with batch normalization (BN) [37] and ReLu [60] to normalize zero-centered around ( $\mu$ ) with standard deviation ( $\sigma$ ) by regularizing *internal covariate shift* and providing a stable learning environment to subsequent deep layers and latent space optimization. All the convolution layers of the block have the same kernel window-size ( $3 \times 3$ , with more focus on local features), non-linear operations (ReLu), strides of one (1), and the same padding [30,36,83]. The size of the channels reduces by halves in each subsequent deep layer where the loss of information can be compensated by increasing the number of filters [78]. The encoding layer capable of the network model to learn at different levels of  $X_{LR}$  representation, from various dimensional spaces to a number of filters, and finally generates latent variable ( $z_i$ ).

**Bottleneck layer:** The encoder part ( $e\{X_{LR}; \phi\}$ ) generates probabilistic distributions (posterior) over the latent space ( $z_i$ ) and then forwards the latent sampling ( $P_i$ ) to the decoder  $d\{e \odot S_f; \vartheta\}$  in order to generate back the images ( $X_{MR}$ ) (Fig. 1). In parallel, the encoder also maps the input ( $X_{LR}$ ) to a linear regression value  $X_R$  which further compares against  $Y_R$ . To utilize the decoder as a generative part,  $e\{X_{LR}; \phi\}$  maps  $X_{LR}$  to posterior distribution  $P(z|X_{LR})$  in the latent space  $z$ , as shown below:

$$z_i \sim \mathcal{P}_i(z|X_{LR}). \quad (2)$$

The drawn sample ( $z_i$ ) from the distribution  $P_i(z|X_{LR})$  maps into the same shape of decoder ( $d\{e \odot S_f; \vartheta\}$ ) for the generation operations. The distribution regularizes both locally ( $\mu$ ) and globally ( $\sigma$ ), where  $\mu$  and  $\sigma$  are the mean and variance for the sample in the given distribution. The sampling process needs to back-propagate the error through the network which is made possible by the parameterize-trick [44]. The decoder then generates multiple outputs corresponding to the same input from a distribution around the center rather than from a fixed point. The latent space needs to be smoothly interpolated between the distributions (i.e., via KL-divergence) for

the samples, so as to generate images with restored information facilitating further the high-fidelity samples.

**Decoding layer:** On the decoding side, a random sample ( $z_i$ , for  $i = 1, 2, \dots, n$ ) is generated from a probabilistic distribution  $P(z|X_{LR})$  and then is projected to  $X_{MR}$ , as shown below:

$$X_{MR} = d\{[z_i \odot R_i] \odot S_f(X_{LR}); \vartheta\}, \quad (3)$$

Here,  $X_{MR}$  is the reconstruction map corresponding to  $z_i$  with adjustable weights ( $\odot R_i$ ), which are regularized by the objective function (see loss-function, “Objective function”) and finally merged with the contextual skipped features ( $\odot S_f$ ). Besides KL-divergence, a reconstruction loss term  $\lambda_2 L_2$  is also included between  $X_{MR}$  and  $Y_{MR}$ .

**Skip layers:** As featured by U-Net, GUV-Net avoids the loss of salient features [82] by passing the contextual information ( $S_f(X_{LR})$ ) between  $e\{X_{LR}; \phi\}$  to  $d\{z_i; \vartheta\}$  in order to generate  $X_{MR}$  with learned detail-level features. At each encoding block, prior to down-sampling, the salient features are retained and passed to the corresponding decoding layer. These extracted features are merged ( $\odot$ ) channel-wise (i.e., axis = 3), and then passed to the next layer followed by the transpose convolution layers [102]. The transpose layer expands the features’ window-size by avoiding to memorize, and learns useful knowledge as necessary for generating high resolution images. The skip-connections are merged as element-wise-sum at the deeper decoding layer, and we keep the number-of-features fixed and refined, which is very important for pixel-wise prediction and generation [58]. However, the lower level merged skip-connections enable variant features, preserving detailed information, and better gradient propagation across the network [31,51]. Overall, the skip-connections control and avoid the loss of spatial information which needs to be retained in HR images required in forensics.

## Refinement and high-fidelity image generation

In the proposed study, during sample preparation, the original image ( $X_o$ ) interpolated to lower ( $X_{LR}$ ) and higher dimensional spaces ( $Y_{HR}$ ) which are then upsampled from  $X_{LR}$  to  $X_{HR}$  while training with balanced computational complexity and progressive learning at distinct scaling factors. GUV-Net mainly prioritizes the generation of high-fidelity shoeprints from their lower resolution noised versions, where the feature learning complexity is dealt at lower dimensional space. The output ( $X_{MR}$ ) from the infused structure borrowed from U-Net with VAE further maps to high resolution version  $X_{HR}$ , as shown below:

$$X_{HR} = C_{in}[\{(X_{MR}; \Phi); L_{\mathcal{P}\mathcal{S}}\}; D_{disc}] \quad (4)$$

$C_{in}$  denotes the learning of  $\Phi$  given  $X_{MR}$ . The terms  $\Phi$  and  $D_{disc}$  denote the learning parameters and discriminative learning at high spatial space, respectively.  $L_{\mathcal{PS}}$  is composed of pixel-wise difference ( $\mathcal{P}$ ) and structure similarity ( $\mathcal{S}$ ) based optimization. Thus, the learning process of high dimensional space is optimized by the content and perceptual objective function followed by the discriminator structure. The inclusion of  $(1 - \lambda \cdot SSIM, \text{ see loss-function 3.2})$  [62] at higher-level tunes the network parameters and bring contextual similarities between given and generated images.

The encoder–decoder structure passes  $X_{MR}$  and skip-connections to the refinement unit (RU) (see refinement unit in Fig. 1c). The network upscales  $X_{MR}$  to HR space via transpose convolution ( $2 \times 2, 3 \times 3$ ). The learning and refinement at higher dimensional space via transpose convolutional operations [17,27,58,89,102] reduce computational complexity, noise amplification, and blurriness [41]. Moreover, the transpose-convolution operations bring new information while training the projection of  $X_{LR}$  to  $X_{HR}$  image space [17,58]. In RU,  $X_{MR}$  further convolves through a stacked of convolutional operations ( $3 \times \text{Conv. R.B}$ ) where each is followed by batch-normalization (B.N) and rectified linear unit (ReLU). Similarly, the parameters through skip-connections are upscaled through variant strides ( $16 \times 16, 8 \times 8, 4 \times 4$ ) and filter-size ( $17 \times 17, 9 \times 9, 5 \times 5$ ). All the skip-connections are projected to the same filter dimensionality with the use of  $1 \times 1$  convolutions which further merged element-wise-sum along the third-dimension. The outcome from merged operation is proceeded to a convolutional layer and then merged in the RU. This process continues to a convolution layer with filter size  $5 \times 5$  followed by mapping to HR image space with a single feature map (filter size  $7 \times 7$ ) activated via tangent function. To avoid the checkerboard-like pattern and over-smoothness in the high quality images [23,76], GUV-Net uses the objective loss term ( $L_{\mathcal{PS}}$ ) to tune the mainstream generative model.

In normal circumstances, the ground truth HR images are available, which enable the model to fine-tune the network parameters for an optimal HR space. However, in our case, the ground truth HR images ( $Y_{MR}$ ) are acquired by bi-cubic interpolation from the original low quality images ( $X_0$ ) which may enable the model to only remember the mapping to the interpolated ones. To avoid such phenomena and allow the network model to learn, a parallel compressing operation performed from higher dimensional ( $Y_{HR}$ ) to lower dimensional ( $Y_{MR}$ ). For a regression output, ( $Y_{MR}$ ) space is followed to  $Y_R$  which further enable parameter fine-tuning of mainstream network at level  $X_{MR}$ .

### Adversarial inferencing

To avoid the blurriness in VAE [25] infused with U-Net, and bring sharpness and better quality in the generated

images [65], GUV-Net learns jointly generative and inference networks in an adversarial manner (Fig. 1c). Adversarial learning plays a min–max game to distinguish the original and fake (generated or synthetic) images. Similarly, GUV-Net brings the inferencing features to reason at the latent space and generates high-quality samples [49]. VAE infused with U-Net together with high-fidelity section (IU) is trained as a generative model and tries to fool discriminator for reaching better level of quality. In our case, the generator maps  $X_{LR}$  to  $X_{MR}$  proceeded by  $X_{HR}$  and the discriminator distinguishes  $Y_{HR}$  and  $X_{HR}$  as real and fake, respectively. The min-max game of learning in GAN can be formulated as follows:

$$V(D, G) = \min_G \max_D (D_{XY}, G_X), \tag{5}$$

Similarly, the generative ( $G_X$ ) and discriminative ( $D_{XY}$ ) operations can be illustrated in mathematical forms as follows:

$$G_X = G\{\underbrace{e(X_{LR}; \phi) \rightarrow d(X_{MR}; \vartheta)}_{\text{Feature extraction}} \rightarrow \underbrace{C_{in}(X_{HR}; \Phi)}_{\text{Refining Unit}}\}$$

$$G_X = G(X_{LR}, X_{HR}; \omega) \text{ where } \omega = \{\phi, \vartheta, \Phi\} \tag{6}$$

The discriminator performs binary classification by assigning probability 1 to  $Y \sim P(Y_{HR})$  and 0 to  $X \sim P(X_{HR})$ . Hence, the discriminator can be optimized as follows:

$$D_{XY} = \frac{P(Y_{HR})}{P(Y_{HR}) + P(X_{HR})} \tag{7}$$

The discriminator plays a vital role in the abstract reconstruction error in the circumstances where VAE is infused in the network model. The discriminator part measures the sample similarity [49] at both element and feature levels. In addition, the discriminator is made stronger to distinguish between real and fake images by including  $L_2$  loss term.

## Objective function

### Discriminative loss function

The objective function for GUV-Net can be mainly divided into discriminative and generative objective functions. The discriminative objective function (DoF) is a combination of sigmoid cross-entropy ( $\mathcal{D}_{XY}$ ) and regression loss ( $L_2$ ) functions which can be formulated as whole in the following equation.

$$\text{DoF} = D_{XY} + L_2, \tag{8}$$

where  $D_{XY}$  and  $L_2$  are the cross-entropy and mean square error (MSE) losses between the real and fake images.  $D_{XY}$  in

Eq. (5) as a loss term for the discriminator seeks to maximize the log probabilities of real and inverse probability for fake images [25].

$$D_{XY} = \frac{1}{n} \sum_{k=1}^n [\log(D(Y_{HR}^k)) + \log(1 - D(G(X_{LR}^k)))] \tag{9}$$

where  $n$  denotes the number of batches. Similarly, the regression loss term  $L_2$  can be illustrated as follows:

$$L_2 = \frac{1}{n} \left[ \sum_{i=1}^n \{X_{HR} - Y_{HR}\}^2 \right]$$

**Generative loss function**

Similarly, the generative objective function (GoF) is composed of multiple loss terms, which is the accumulated sum with more weightage given to generative part ( $G_X$ ) of GAN for adversarial learning, as shown below:

$$GoF = G_X + \mathcal{KL} + \lambda_{1,2,3}L_2 + L_{ps} \tag{10}$$

In Eq. (10), the adversarial term adopted from the GAN model seeks to minimize the inverse probability.

$$G_X = \frac{1}{n} \sum_{k=1}^n [\log(1 - D(G(X_{LR})))]. \tag{11}$$

$G_X$  in GoF encourages the generator to produce samples that being predicted fake by the discriminator with low probability [55]. The generative loss term takes part in min-max game to distinguish real and fake images to produce a realistic high-fidelity image [25,50,73].

Moreover,  $\mathcal{KL}$  in Eq. (10) is the KL-divergence, and it computes the log difference between the probability of data in actual distribution  $P(X_{LR})$  and that of the approximating distribution  $Q(X_{LR})$ . Thus, in the VAE part of GUV-Net, the inference model ( $Q_\phi(z|X_{LR})$ ) approximates the posterior (true) distribution ( $P_\theta(z|X_{LR})$ ) in terms of KL-divergence to minimize the gap [44]:

$$\mathcal{KL}(Q_\phi(z|X_{LR})||P_\theta(z|X_{LR})) = \mathbb{E}_{q_\phi} \left[ \log \frac{q_\phi(z|x)}{p_\theta(z|x)} \right]. \tag{12}$$

Specifically in our case, KL-divergence measures the difference between the distribution  $\mathcal{N}(\mu_i, \sigma_i)$  of inference model with mean  $\mu_i$  and variance  $\sigma_i$ , and standard normal distribution  $\mathcal{N}(0, I)$  with mean 0 and unit variance  $I$ . After the Bayesian inference simplification [13,19], KL-divergence can be rewritten as follows:

$$\mathcal{KL}(\mathcal{N}(\mu, \sigma)||\mathcal{N}(0, I)) = \frac{1}{2} \sum_{i=1}^l (\sigma_i^2 + \mu_i^2 - 1 - \exp(\sigma_i^2)), \tag{13}$$

and by choosing  $I = 1$ , Eq. (13) becomes:

$$\mathcal{KL}(\mathcal{N}(\mu, \sigma)||\mathcal{N}(0, 1)) = \frac{1}{2} \sum_{i=1}^l (\sigma_i + \mu_i^2 - 1 - \exp(\sigma_i)). \tag{14}$$

Similarly,  $\mathcal{L}_2$  is the pixel-wise loss to efficiently evaluate noisy images while training [6]. The subscript of  $\lambda$  in Eq. (10) is 3, representing the three versions of  $\mathcal{L}_2$  for pixel-wise difference. For  $\lambda_1$ , the difference between  $X_{LR}$  and  $X_{MR}$  can be illustrated as follows:

$$\lambda_1 \mathcal{L}_2 = \lambda_1 \cdot \frac{1}{r \cdot c} \left[ \sum_{r=0}^{h-1} \sum_{c=0}^{w-1} \{X_{LR}(r, c) - X_{MR}(r, c)\}^2 \right], \tag{15}$$

where  $r$  and  $c$  denote the row and column indexes, respectively.  $X_{LR}(r, c)$  and  $X_{MR}(r, c)$  denote the corresponding pixel positions in the input ( $X$ ) and projected ( $Y$ ) images, respectively. While  $h$  and  $w$  are the height and width of both images, respectively. Similarly, for  $\lambda_2$ , the element-wise difference between  $X_{HR}$  and  $Y_{HR}$  can be formulated by re-writing Eq. (15) as follows:

$$\lambda_2 \mathcal{L}_2 = \lambda_2 \cdot \frac{1}{r \cdot c} \left[ \sum_{r=0}^{h-1} \sum_{c=0}^{w-1} \{X_{HR}(r, c) - Y_{HR}(r, c)\}^2 \right]. \tag{16}$$

For  $\lambda_3$ , the loss term can be formulated as follows:

$$\lambda_3 \mathcal{L}_2 = \lambda_3 \cdot \frac{1}{n} \left[ \sum_{i=1}^n \{X_R^i - Y_R^i\}^2 \right], \tag{17}$$

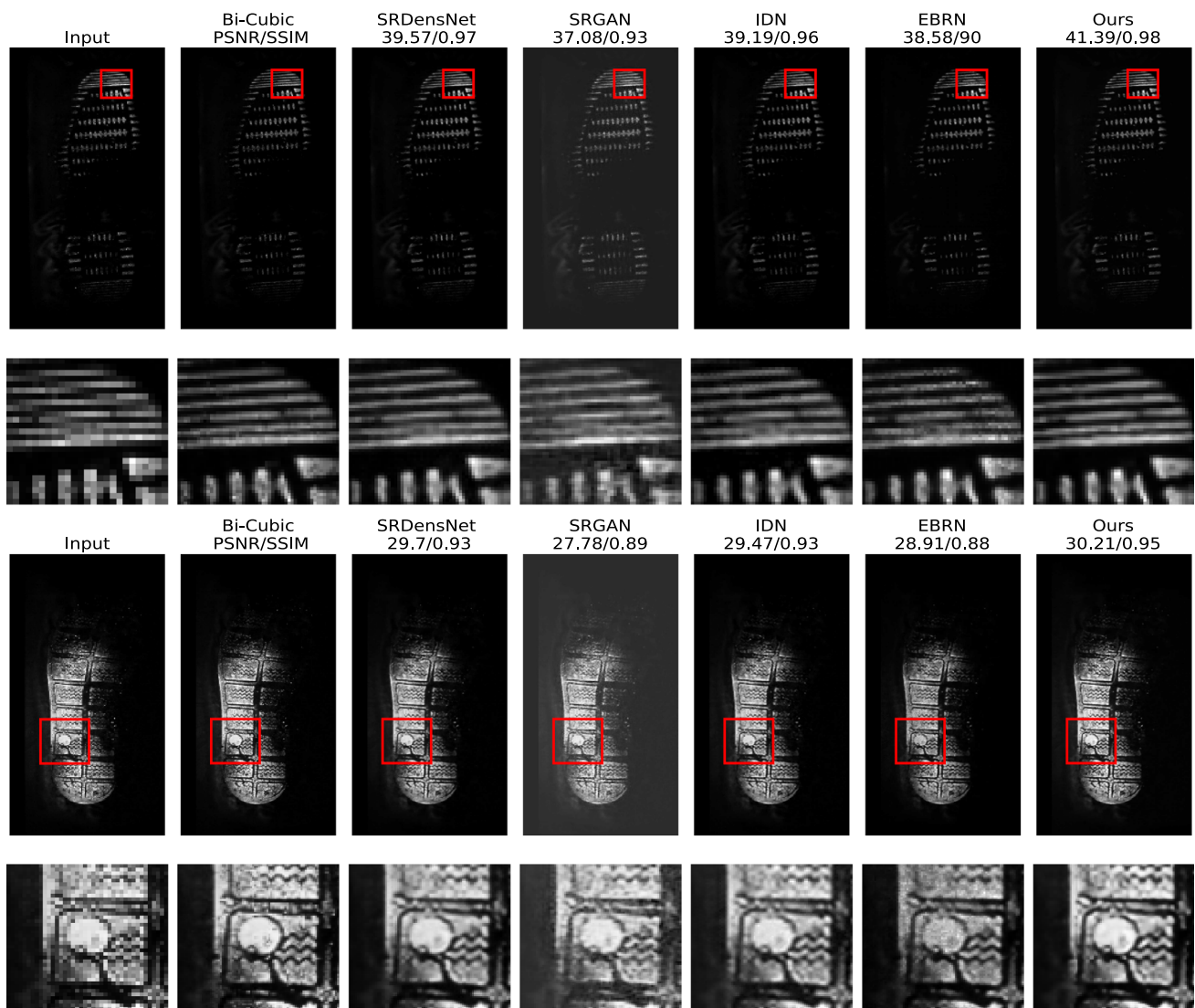
where  $X_R$  and  $Y_R$  are the regression values computed from the encoder and inverse-encoding parts corresponding to sample ( $i$ ), respectively.

Furthermore, the refinement process at a higher spatial level is optimized by following both pixel-wise loss ( $\mathcal{L}_2$ ) and structure similarity (SSIM) [6].  $\mathcal{L}_2$  favors higher peak-to-single-noise-ratio (PSNR) while SSIM improves the perceptual quality in the generated HR images [18,39]. The final term  $\mathcal{L}_{ps}$  in Eq. (10) can be rewritten as follows:

$$\begin{aligned} \mathcal{L}_{ps} &= \mathcal{L}_p + (1 - \mathcal{L}_s), \\ \mathcal{L}_{ps} &= \lambda_2 \mathcal{L}_2 + (1 - \mathcal{L}_s). \end{aligned} \tag{18}$$

The HR shoeprints should maintain the original structure in terms of forensic applications; hence, the structure similarity





**Fig. 2** Two shoeprint samples are shown with each has two rows. The generated shoeprints (first and third rows) by the SOTA models and GUV-Net with highlighted regions (second and fourth rows). Each network receives the input shoeprint and generates the corresponding HR

(upscaled by  $\times 2$ ) images. Bi-Cubic interpolated shoeprints are used as ground truths (for details see “Result” section). Similarly, the corresponding PSNR and SSIM metric values are also displayed

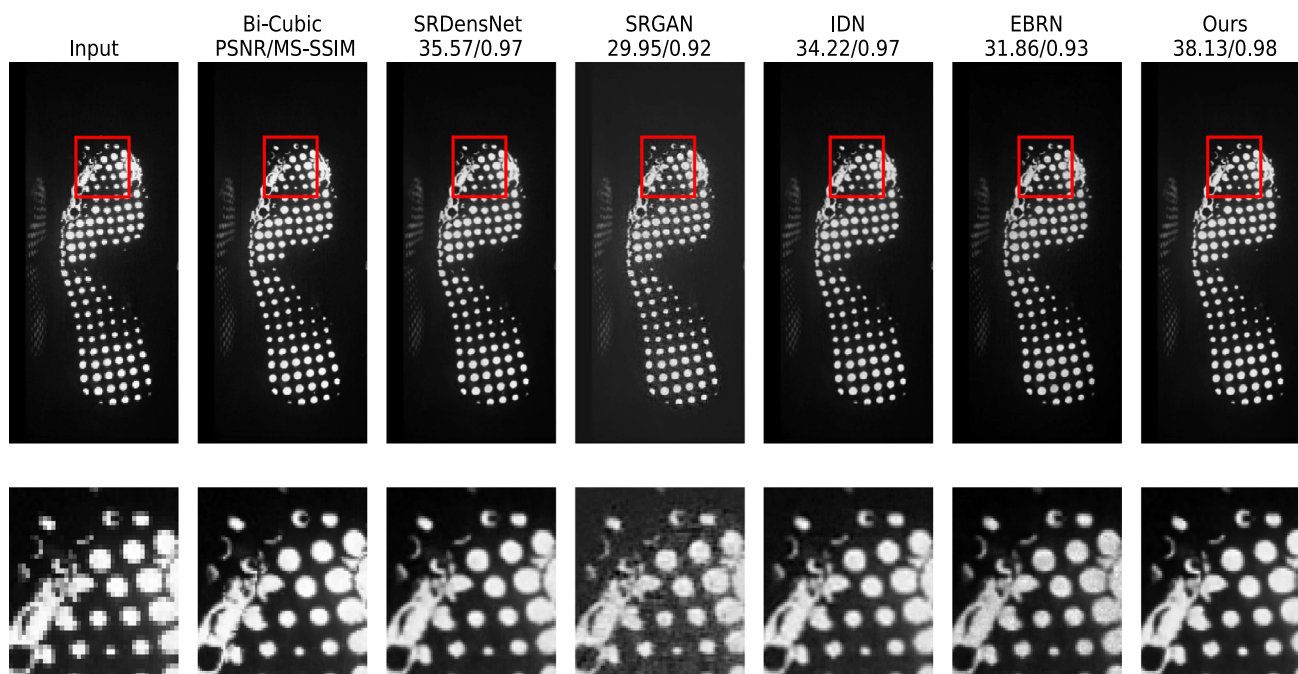
(SSIM) [6] index as an objective function is also included. It also quantifies the perceptual quality of the degraded images. By including SSIM as a loss term, GUV-Net penalizes the learning parameters at high dimensional spaces between  $X_{HR}$  and  $Y_{HR}$ . SSIM focuses mainly on three properties of the given images as shown in the following illustration.

$$\mathcal{L}_s(X_{HR}, Y_{HR}) = [\mathbb{L}(X, Y)] \cdot [\mathbb{C}(X, Y)] \cdot [\mathbb{S}(X, Y)], \quad (19)$$

where  $\mathbb{L}$ ,  $\mathbb{C}$ , and  $\mathbb{S}$  denote the luminous, contrast, and structure differences between  $X$  and  $Y$ . SSIM enables GUV-Net to generate high quality and visually pleasant images having similarity in structure with their LR images.

### Model training

To assess the performance of GUV-Net, some SOTA models included SRDensNet [89], SRGAN [50], IDN [34], and EBRN [64] are also trained on the same training (84,000 images) and testing (4000 images) datasets [28], as well as with fine-tuned hyper-parameters corresponding to the current problem of HR shoeprints. All models, including GUV-Net, are trained for super-resolution with upscaling factors  $\times 2$  and  $\times 4$ . ADAM is used for the optimization of GUV-Net with an initial learning rate  $10^{-4}$ , with learning decay  $10^{-1}$  after every 20 epochs,  $\beta_1 = 0.9$ , and  $\beta_2 = 0.999$ ,



**Fig. 3** SR of shoeprint together with highlighted regions generated with upscaling factor  $\times 2$ . Each image contains PSNR and MS-SSIM score. The higher the score, the better the quality of the images is

which are also applied to the understudied SOTA models [34,50,64,89].

## Results evaluation and analysis

We compare the performance of GUV-Net against SOTA models through both subjective and objective evaluation methods. Some random images with their zoomed regions of both input and generated SR images are shown. Similarly, the corresponding values for PSNR, SSIM Fig. 2, and MS-SSIM Fig. 3 are also calculated.

### Bi-cubic interpolated shoeprints as ground truths

For computing PSNR, SSIM, and MS-SSIM, we used bi-cubic interpolated shoeprint as ground truth. The interpolated version is generated with a new dimension ( $h \times n, w \times n$ ) from that of original shoeprint images ( $h, w$ ) for training purpose. The original shoeprints have a variety of sizes and dimensions. To bring them into the same dimensional structure, a variational scaling factor ( $\eta$ ) is used. Recall Eq. (1),  $Y_{HR}$  can be rewritten as follows:

$$Y_{HR} = \mathcal{Y}\{X_o, (h \times \eta n, w \times \eta n)\},$$

where  $\eta$  can be a fractional or multiple numbers chosen according to the dimension of  $Y_{HR}$ . Therefore, the bi-cubic interpolated shoeprints are not the versions created from  $X_{LR}$

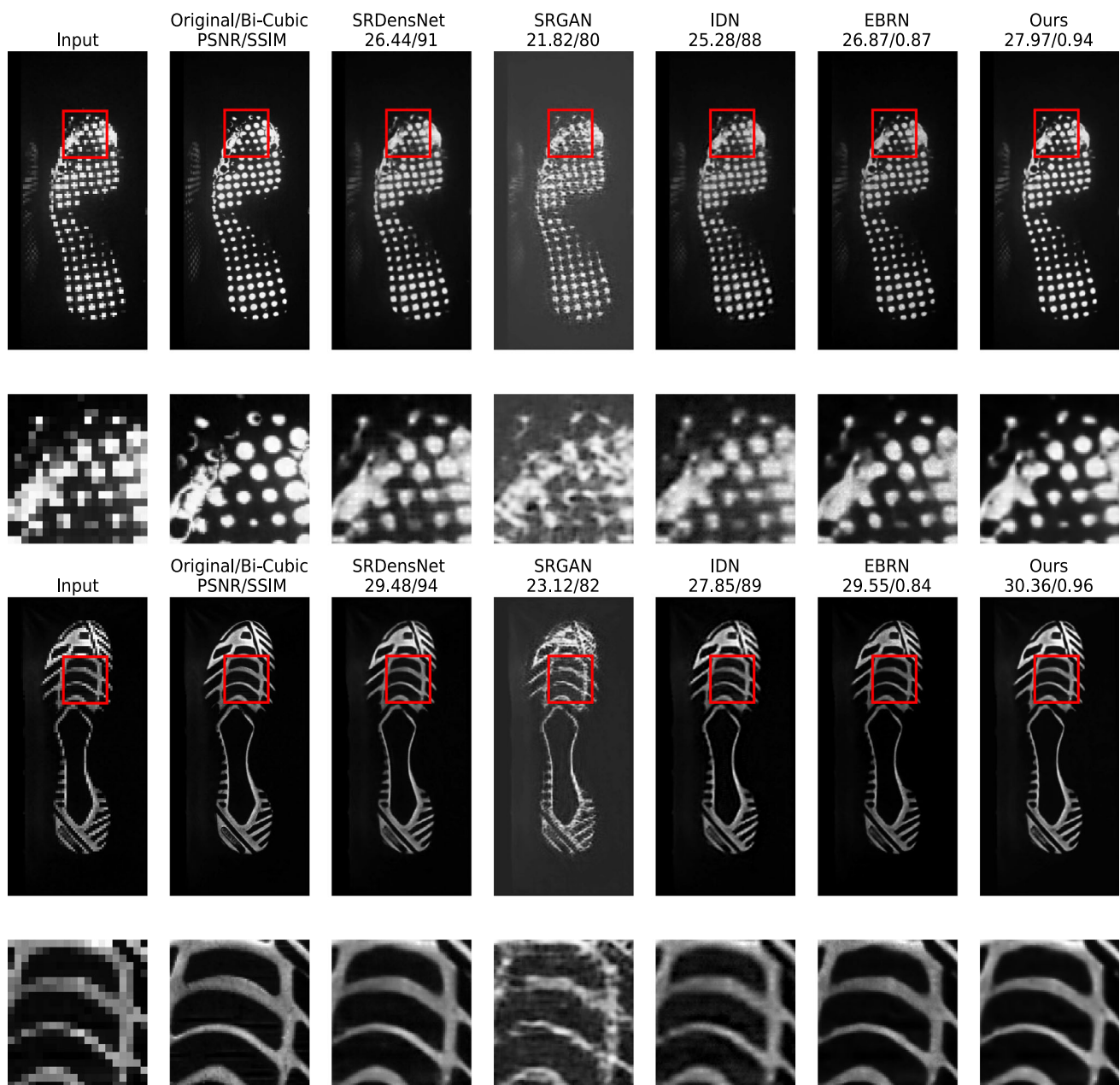
images; hence, these upscaled images can be observed with good quality. However, the trained models including GUV-Net receive  $X_{LR}$  during training and evaluation. Regarding the aforementioned reason and absent ground truths, bi-cubic interpolated shoeprints are used as a baseline to carry out both subjective and empirical evaluations.

### Performance evaluation

Based on the SR generation with upscaling factor  $\times 2$ , GUV-Net results higher PSNR and SSIM values against the trained SOTA models (Fig. 2). Similarly, in the given visualization, the patterns in cropped regions of shoeprints generated by GUV-Net can be observed more clearly as compared to SOTA models. This reflects the learning specialization to shoeprint by producing a better result in terms of noise and structure similarities with that of ground truths.

To better assess the quality of SR images, we also used another highly adapted evaluation metric (MS-SSIM) based on the assumption of human visual system [94]. MS-SSIM follows multi-scale processes for multi-stage sub-sampling operations to subjectively compare the given images. For this purpose, the generated images by SOTA models and GUV-Net models are displayed in Fig. 3, where GUV-Net results significant MS-SSIM score.

Furthermore, the generated results with upscaling factor  $\times 4$  can also be observed together with highlighted regions (Fig. 4). The SR images generated by GUV-Net keep sufficient original patterns and texture from the input image



**Fig. 4** Two shoeprint samples are selected from the testing dataset. Each sample is shown in generated form (first row) and zoomed-in region (second row). The visualization shows the input and generated SR images by SOTA and GUV-Net models for the upscaling factor  $\times 4$

( $X_{LR}$ ), as well as reduce the noise level by producing high PSNR score compared to SOTA. GUV-Net also keeps the low level features empowered by the direct connections from the decoder part into the refinement module (see Fig. 1b). Moreover, GUV-Net outperforms the SOTA models in terms of the empirical evaluations using PSNR, SSIM, and MS-SSIM as tabulated the averaged values in Table 1. In addition, SRDensNet performs best at second position following dense structure which is followed by GUV-Net for passing information and refinement unit.

## Ablation study

### Exclusion of variational inferencing

GUV-Net generates samples through the decoding layer by inferencing in the latent space optimized by the  $KL$ -divergence. To know the importance of features borrowed from VAE architecture, we re-designed GUV-Net by excluding the inferencing unit at the bottle-neck of encoder–decoder structure. We trained the network with the same network-parameters by solely adding autoencoder instead of VAE. The

**Table 1** Average PSNR, SSIM, and M-SSIM values for the testing dataset

Model name	PSNR	SSIM	MS-SSIM
$\times 2$			
SRDensNet	34.02	0.96	0.97
SRGAN	30.56	0.90	0.93
IDAN	33.40	0.95	0.96
EBGRN	32.71	0.90	0.96
GUV-Net (ours)	<b>35.79</b>	<b>0.97</b>	<b>0.98</b>
$\times 4$			
SRDensNet	34.02	0.94	0.98
SRGAN	30.56	0.82	0.94
IDAN	33.40	0.89	0.98
EBGRN	32.71	0.84	0.98
GUV-Net (ours)	<b>35.79</b>	<b>0.96</b>	<b>0.99</b>

The statistical results of the table is divided into two sections, corresponding to upsampling  $\times 2$  and  $\times 4$ . The quantitative (i.e., PSNR, SSIM, MS-SSIM) values for the the SOTA models including with our proposed one are shown both for upsampling  $\times 2$  and  $\times 4$ . The significant results are shown in bold in terms of high fidelity shoeprint generation

modified version of GUV-Net performs satisfactory result for the scaling factor  $\times 2$  in terms of PSNR and SSIM; however, it shows poor result at high scaling factors (i.e.,  $\times 4$ ,  $\times 8$ ). The model convergence was negatively affected after the training reached to 10 epochs and produced blurry results.

### Refining association with skip-connections

Similarly, we remove the skip-connections between the decoding part and the refinement unit to better observed the contribution of passing information from various spatial levels. The model performance retains normal in terms of perception quality and SSIM but shows low PSNR value for the scaling factor  $\times 4$  and above. Thus, the skip-connections not only pass the low level features from distinct levels but also take part in refining the high dimensional space. The existing of these connections show more importance where the original and low quality images are often found in distorted forms.

### Conclusion and future directions

In this study, we proposed GUV-Net for SISR specialized in shoeprint generation. GUV-Net possesses features of the three popular network structures: GAN, VAE, and U-Net, which effectively addresses the crucial challenges in shoeprint generation. The main challenges addressed by GUV-Net is the unavailability of ground truths and the generation of SR shoeprints from their naturally distorted versions. The model is trained and tuned following multiple loss terms

in an efficient way. To the best of our knowledge, this is the first model to attempt super-resolution image generation, which is of great importance in forensic investigation by maintaining the key patterns and textures of LR images. Moreover, the model efficiently retains the salient features and patterns from the LR ( $X_{LR}$ ) to HR ( $X_{HR}$ ) version. GUV-Net outperforms the SOTA models in terms of subjective (Figs. 3, 4) and objective (Table 1) evaluations.

The unavailability of HR shoeprints arises multiple questions regarding training and evaluation of GUV-Net. The SR image quality with upscaling factors (UF)  $\times 2$ ,  $\times 4$ , and  $\times 8$  generated by GUV-Net sustains to an acceptable level; however, GUV-Net including SOTA show poor result for higher upscaling factor (i.e.,  $UF \times 8 >$ ). In the future, GUV-Net can be extended to higher upscaling factors (i.e.,  $\times 16$ ,  $\times 32$ ,  $\times 64$ ). For this purpose, the depth of the encoder–decoder structure can be deepen together with skip-connections between the decoder and refinement modules to get an improved version.

Similarly, in our future work, the modified version of GUV-Net should focus more on noise and blur control in SR generation. For the improved version of GUV-Net, the training and convergence rates can also be considered which has been given less emphasize due to the challenge in SR shoeprint generation with no ground truths. Moreover, SR shoeprint generation needs special attention to study network models for no-reference HR images. The SISR of shoeprint image generation through GUV-Net using a fusion strategy can be extended to other vision tasks.

**Acknowledgements** This research is supported by the National Natural Science Foundation of China (Grants nos. 61772227, 61972174, 61972175, 62072212), Science and Technology Development Foundation of Jilin Province (nos. 20180201045GX, 20200201300JC, 20200401083GX, 20200201163JC), and as the Paul K. and Diane Shumaker Endowment Fund to DX.

**Author Contributions** MH contributed to the theoretical development, experimental design, and prototype development. The author prepared the datasets for network training and evaluation, designed, and tuned the network model, trained, and validated the GUV-Net model. Furthermore, the author evaluated the findings through both subjective and objective evaluation metrics and analyzed against the real-world applications. Finally, proceeded the findings into documentation in the form of paper writing. YW made significant contributions to the network design, prototype development, parameters selection, and analysis. The author revised the draft and approved the final version for submission. WP contributed to model designing and development, and the interpretation of findings for analysis. The author also contributed to carry out the state-of-the-art studies and comparison. DW analyzed, interpreted the datasets, and revised the paper writing as proofreading. The author also directed to perform the qualitative and quantitative comparison and analysis against the state-of-the-art models. DL collected shoeprint samples from subjects using the given device. The author also interpreted the data associated with the current study. YZ supervised and directed the whole procedure, assisted in prototyping the developmental process, analyzed and interpreted the associated data in the study. The author performed proofreading and critical revision of the article. DX

supervised and directed the whole study, encouraged to carry out various experimental designs, analysis, result evaluations, and contributed to the final draft with proofreading and approval of the version to be submitted.

## Declarations

**Conflict of interest** As the corresponding author on behalf of all the authors, I declare that all the authors are aware of the submission and have no conflict of interest. The submitted paper contains original, unpublished results, and is not currently under consideration elsewhere.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Acevedo Mosqueda M, Acevedo Mosqueda M, Carreno Aguilera R, Martinez Zuñiga F, Pacheco Bautista D, Patiño Ortiz M, Yu W (2019) Computational intelligence for shoeprint recognition. *Fractals* 27(04):1950080
- Alexandre G (1996) Computerized classification of the shoeprints of burglars' soles. *Forensic Science International* 82(1):59–65
- AlGarni G, Hamiane M (2008) A novel technique for automatic shoeprint image retrieval. *Forensic Sci Int* 181(1–3):10–14
- Arjovsky M, Bottou L (2017) Towards principled methods for training generative adversarial networks. arXiv preprint [arXiv:1701.04862](https://arxiv.org/abs/1701.04862)
- Atamturk D (2010) Estimation of sex from the dimensions of foot, footprints, and shoe. *Anthropologischer Anzeiger*, pp 21–29
- Avci I, Sankur B, Sayood K (2002) Statistical evaluation of image quality measures. *J Electron Imaging* 11(2):206–223
- Bao J, Chen D, Wen F, Li H, Hua G (2017) Cvae-gan: fine-grained image generation through asymmetric training. In: Proceedings of the IEEE international conference on computer vision, pp 2745–2754
- Basu N, Bandyopadhyay SK (2017) Crime scene reconstruction—sex prediction from blood stained foot sole impressions. *Forensic Sci Int* 278:156–172
- Bhattacharyya A, Fritz M, Schiele B (2019) "best-of-many-samples" distribution matching. arXiv preprint [arXiv:1909.12598](https://arxiv.org/abs/1909.12598)
- Bodziak WJ (1999) *Footwear impression evidence: detection, recovery and examination*. CRC Press, Boca Raton
- Chan KC, Wang X, Xu X, Gu J, Loy CC (2020) Glean: generative latent bank for large-factor image super-resolution. arXiv preprint [arXiv:2012.00739](https://arxiv.org/abs/2012.00739)
- Chen Z, Tong Y (2017) Face super-resolution through wasserstein gans. arXiv preprint [arXiv:1705.02438](https://arxiv.org/abs/1705.02438)
- Chen Z, Wang R, Zhang Z, Wang H, Xu L (2019) Background-foreground interaction for moving object detection in dynamic scenes. *Inf Sci* 483:65–81
- Cui J, Zhao X, Liu N, Morgachev S, Li D (2019) Robust shoeprint retrieval method based on local-to-global feature matching for real crime scenes. *J Forensic Sci* 64(2):422–430
- Dong C, Loy CC, He K, Tang X (2014) Learning a deep convolutional network for image super-resolution. In: European conference on computer vision. Springer, pp 184–199
- Dong C, Loy CC, He K, Tang X (2015) Image super-resolution using deep convolutional networks. *IEEE Trans Pattern Anal Mach Intell* 38(2):295–307
- Dong C, Loy CC, Tang X (2016) Accelerating the super-resolution convolutional neural network. In: European conference on computer vision. Springer, pp 391–407
- Dosovitskiy A, Brox T (2016) Generating images with perceptual similarity metrics based on deep networks. arXiv preprint [arXiv:1602.02644](https://arxiv.org/abs/1602.02644)
- Duchi J (2007) Derivations for linear algebra and optimization. *Berkeley California* 3(1):2325–5870
- Esser P, Sutter E, Ommer B (2018) A variational u-net for conditional appearance and shape generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8857–8866
- Francis X, Sharifzadeh H, Newton A, Baghaei N, Varastehpour S (2019) Feature enhancement and denoising of a forensic shoeprint dataset for tracking wear-and-tear effects. In: 2019 IEEE international symposium on signal processing and information technology (ISSPIT). IEEE, pp 1–5
- Francis X, Sharifzadeh H, Newton A, Baghaei N, Varastehpour S (2019) Learning wear patterns on footwear outsoles using convolutional neural networks. In: 2019 18th IEEE international conference on trust, security and privacy in computing and communications/13th IEEE international conference on big data science and engineering (TrustCom/BigDataSE). IEEE, pp 450–457
- Gao H, Yuan H, Wang Z, Ji S (2019) Pixel transposed convolutional networks. *IEEE Trans Pattern Anal Mach Intell* 42(5):1218–1227
- Geradts Z, Keijzer J (1996) The image-database rebezo for shoeprints with developments on automatic classification of shoe outsole designs. *Forensic Sci Int* 82(1):21–31
- Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks. arXiv preprint [arXiv:1406.2661](https://arxiv.org/abs/1406.2661)
- Greenspan H (2009) Super-resolution in medical imaging. *Comput J* 52(1):43–63
- Haris M, Shakhnarovich G, Ukita N (2018) Deep back-projection networks for super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1664–1673
- Hassan M, Wang Y, Wang D, Li D, Liang Y, Zhou Y, Xu D (2021) Deep learning analysis and age prediction from shoeprints. *Forensic Sci Int* 327:110987. <https://doi.org/10.1016/j.forsciint.2021.110987>
- Heydari AA, Mehmood A (2020) SRVAE: super resolution using variational autoencoders. In: *Pattern Recognition and Tracking XXXI*, vol 11400. International Society for Optics and Photonics, pp 114000U
- Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint [arXiv:1704.04861](https://arxiv.org/abs/1704.04861)
- Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR).

- IEEE Computer Society, pp 2261–2269. <https://doi.org/10.1109/CVPR.2017.243>
32. Huang H, Li Z, He R, Sun Z, Tan T (2018) Introvae: introspective variational autoencoders for photographic image synthesis. arXiv preprint [arXiv:1807.06358](https://arxiv.org/abs/1807.06358)
  33. Huang Y, Shao L, Frangi AF (2017) Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). IEEE Computer Society, pp 578–5796
  34. Hui Z, Wang X, Gao X (2018) Fast and accurate single image super-resolution via information distillation network. In: 2018 IEEE/CVF conference on computer vision and pattern recognition. IEEE, pp 723–731
  35. Hyun S, Heo JP (2020) Varsr: variational super-resolution network for very low resolution images. Springer, Berlin
  36. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K (2016) Squeezenet: alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. arXiv preprint [arXiv:1602.07360](https://arxiv.org/abs/1602.07360)
  37. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. PMLR, pp 448–456
  38. Isaac JS, Kulkarni R (2015) Super resolution techniques for medical image processing. In: 2015 International conference on technologies for sustainable development (ICTSD). IEEE, pp 1–6
  39. Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. Springer, pp 694–711
  40. Khan SH, Hayat M, Barnes N (2018) Adversarial training of variational auto-encoders for high fidelity image generation. In: 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1312–1320
  41. Kim J, Lee JK, Lee KM (2016) Accurate image super-resolution using very deep convolutional networks. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 1646–1654
  42. Kim J, Lee JK, Lee KM (2016) Deeply-recursive convolutional network for image super-resolution. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 1637–1645
  43. Kim KI, Kwon Y (2008) Example-based learning for single-image super-resolution and jpeg artifact removal (Report No.TR-173). Max Planck Institute for Biological Cybernetics. [https://eprints.lancs.ac.uk/id/eprint/69842/1/Example\\_based.pdf](https://eprints.lancs.ac.uk/id/eprint/69842/1/Example_based.pdf)
  44. Kingma DP, Welling M (2019) An introduction to variational autoencoders. arXiv preprint [arXiv:1906.02691](https://arxiv.org/abs/1906.02691)
  45. Kong B, Supancic J, Ramanan D, Fowlkes C (2017) Cross-domain forensic shoeprint matching In: British machine vision conference (BMVC). London, UK, pp 1–5
  46. Kortylewski A (2017) Model-based image analysis for forensic shoe print recognition. Ph.D. thesis, University of Basel
  47. Lai WS, Huang JB, Ahuja N, Yang MH (2017) Deep laplacian pyramid networks for fast and accurate super-resolution. In: 30th IEEE conference on computer vision and pattern recognition, CVPR 2017. IEEE Computer Society, pp. 5835–5843. <https://doi.org/10.1109/CVPR.2017.618>
  48. Lai WS, Huang JB, Ahuja N, Yang MH (2018) Fast and accurate image super-resolution with deep Laplacian pyramid networks. IEEE Trans Pattern Anal Mach Intell 41(11):2599–2613
  49. Larsen ABL, Sønderby SK, Larochelle H, Winther O (2016) Autoencoding beyond pixels using a learned similarity metric. In: International conference on machine learning. PMLR, pp 1558–1566
  50. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). pp 105–144
  51. Lee CY, Xie S, Gallagher P, Zhang Z, Tu Z (2015) Deeply-supervised nets. In: Artificial intelligence and statistics. PMLR, pp 562–570
  52. Li Z, Yang J, Liu Z, Yang X, Jeon G, Wu W (2019) Feedback network for image super-resolution. In: 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp. 3862–3871. <https://doi.org/10.1109/CVPR.2019.00399>
  53. Lim B, Son S, Kim H, Nah S, Mu LK (2017) Enhanced deep residual networks for single image super-resolution. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW). pp 136–144
  54. Lu X, Yuan H, Yan P, Yuan Y, Li X (2012) Geometry constrained sparse coding for single image super-resolution. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp 1648–1655
  55. Lucic M, Kurach K, Michalski M, Gelly S, Bousquet O (2017) Are GANs created equal? A large-scale study. arXiv preprint [arXiv:1711.10337](https://arxiv.org/abs/1711.10337)
  56. Ma Z, Ding Y, Wen S, Xie J, Jin Y, Si Z, Wang H (2019) Shoeprint image retrieval with multi-part weighted CNN. IEEE Access 7:59728–59736
  57. Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B (2015) Adversarial autoencoders. arXiv preprint [arXiv:1511.05644](https://arxiv.org/abs/1511.05644)
  58. Mao XJ, Shen C, Yang YB (2016) Image restoration using convolutional auto-encoders with symmetric skip connections. arXiv preprint [arXiv:1606.08921](https://arxiv.org/abs/1606.08921)
  59. Metz L, Poole B, Pfau D, Sohl-Dickstein J (2016) Unrolled generative adversarial networks. arXiv preprint [arXiv:1611.02163](https://arxiv.org/abs/1611.02163)
  60. Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. In: Icm1
  61. Ni KS, Nguyen TQ (2007) Image superresolution using support vector regression. IEEE Trans Image Process 16(6):1596–1610
  62. Okubike EA, Nandi ME, Iheaza EC, Obun OC (2019) Stature prediction using shoe print dimensions of an adult Nigerian population. Arab J Forensic Sci Forensic Med (AJFSFM) 1(8):989–1003
  63. Ozden H, Balci Y, Demir C, Turgut A, Ertugrul M (2005) Stature and sex estimate using foot and shoe dimensions. Forensic Sci Int 147(2–3):181–184
  64. Qiu Y, Wang R, Tao D, Cheng J (2019) Embedded block residual network: a recursive restoration model for single-image super-resolution. In: 2019 IEEE/CVF international conference on computer vision (ICCV). pp 4180–4189
  65. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint [arXiv:1511.06434](https://arxiv.org/abs/1511.06434)
  66. Rasti P, Uiboupin T, Escalera S, Anbarjafari G (2016) Convolutional neural network super resolution for face recognition in surveillance monitoring. In: International conference on articulated motion and deformable objects. Springer, pp 175–184
  67. Razavi A, van den Oord A, Vinyals O (2019) Generating diverse high-resolution images with VQ-VAE. DGS@ICLR 2019 Workshop
  68. Rida I, Al-Maadeed N, Al-Maadeed S, Bakshi S (2018) A comprehensive overview of feature representation for biometric recognition. Multimed Tools Appl 79(7–8):4867–4890
  69. Rida I, Bakshi S, Proença H, Fei L, Nait-Ali A, Hadid A (2019) Forensic shoe-print identification: a brief survey. arXiv preprint [arXiv:1901.01431](https://arxiv.org/abs/1901.01431)
  70. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 234–241

71. Rosca M, Lakshminarayanan B, Mohamed S (2018) Distribution matching in variational inference. arXiv preprint [arXiv:1802.06847](https://arxiv.org/abs/1802.06847)
72. Rosca M, Lakshminarayanan B, Warde-Farley D, Mohamed S (2017) Variational approaches for auto-encoding generative adversarial networks. arXiv preprint [arXiv:1706.04987](https://arxiv.org/abs/1706.04987)
73. Sajjadi MS, Scholkopf B, Hirsch M (2017) Enhancenet: single image super-resolution through automated texture synthesis. In: Proceedings of the IEEE international conference on computer vision. pp. 4491–4500
74. Shamsolmoali P, Li X, Wang R (2019) Single image resolution enhancement by efficient dilated densely connected residual network. *Signal Process Image Commun* 79:13–23
75. Shamsolmoali P, Zareapoor M, Wang R, Zhou H, Yang J (2019) A novel deep structure u-net for sea-land segmentation in remote sensing images. *IEEE J Sel Top Appl Earth Obs Remote Sens* 12(9):3219–3232
76. Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R, Rueckert D, Wang Z (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). pp 1874–1883
77. Shocher A, Cohen N, Irani M (2018) Zero-shot super resolution using deep internal learning. In: 2018 IEEE/CVF conference on computer vision and pattern recognition. pp 3118–3126
78. Simonyan K, Zisserman A Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
79. Srihari SN (2011) Analysis of footwear impression evidence, final technical report, award number: 2007-dn-bx-k135, awarded to research foundation of the State University of New York. US Department of Justice Report
80. Sun J, Xu Z, Shum HY (2008) Image super-resolution using gradient profile prior. In: 2008 IEEE conference on computer vision and pattern recognition. IEEE, pp 1–8
81. Sun L, Hays J (2012) Super-resolution from internet-scale scene matching. In: 2012 IEEE international conference on computational photography (ICCP). IEEE, pp 1–12
82. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: CVPR. pp 1–9
83. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: 2016 IEEE conference on computer vision and pattern recognition. pp 2818–2826
84. Tai Y, Yang J, Liu X (2017) Image super-resolution via deep recursive residual network. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). pp 2790–2798
85. Tai Y, Yang J, Liu X, Xu C (2017) Memnet: a persistent memory network for image restoration. In: 2017 IEEE international conference on computer vision (ICCV). IEEE Computer Society, pp 4549–4557
86. Tang Y, Srihari SN, Kasiviswanathan H, Corso JJ (2010) Footwear print retrieval system for real crime scene marks. In: International workshop on computational forensics. Springer, pp 88–100
87. Tappen MF, Liu, C (2012) A bayesian approach to alignment-based image hallucination. In: European conference on computer vision. Springer, pp 236–249
88. Theis L, Oord A, Bethge M (2015) A note on the evaluation of generative models. arXiv preprint [arXiv:1511.01844](https://arxiv.org/abs/1511.01844)
89. Tong T, Li G, Liu X, Gao Q (2017) Image super-resolution using dense skip connections. In: 2017 IEEE international conference on computer vision (ICCV). IEEE, pp 4809–4817
90. Tsunekawa S, Inoue K, Yoshioka M (2018) Image up-sampling for super resolution with generative adversarial network. In: Australasian joint conference on artificial intelligence. Springer, pp 258–270
91. Wan C, Probst T, Van Gool L, Yao A (2017) Crossing nets: combining gans and vaes with a shared latent space for hand pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 680–689
92. Wang Y, Perazzi F, McWilliams B, Sorkine-Hornung A, Sorkine-Hornung O, Schroers C (2018) A fully progressive approach to single-image super-resolution. In: 2018 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW). IEEE Computer Society, pp. 977–97709
93. Wang Z, Chen J, Hoi SC (2020) Deep learning for image super-resolution: a survey. *IEEE Trans Pattern Anal Mach Intell* 43(10):3365–3387. <https://doi.org/10.1109/TPAMI.2020.2982166>. Accessed 1 Oct 2021
94. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. In: The thirty-seventh asilomar conference on signals, systems & computers, 2003, vol 2. IEEE, pp 1398–1402
95. Xiao R, Shi P (2008) Computerized matching of shoeprints based on sole pattern. In: International workshop on computational forensics. Springer, pp 96–104
96. Yang CY, Liu S, Yang MH (2013) Structured face hallucination. In: 2013 IEEE conference on computer vision and pattern recognition. pp 1099–1106
97. Yang J, Lin Z, Cohen S (2013) Fast image super-resolution based on in-place example regression. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1059–1066
98. Yang J, Wang Z, Lin Z, Cohen S, Huang T (2012) Coupled dictionary training for image super-resolution. *IEEE Trans Image Process* 21(8):3467–3478
99. Yang J, Wright J, Huang T, Ma Y (2008) Image super-resolution as sparse representation of raw image patches. In: 2008 IEEE conference on computer vision and pattern recognition. IEEE, pp 1–8
100. Yang J, Wright J, Huang TS, Ma Y (2010) Image super-resolution via sparse representation. *IEEE Trans Image Process* 19(11):2861–2873
101. Yang W, Zhang X, Tian Y, Wang W, Xue JH, Liao Q (2019) Deep learning for single image super-resolution: a brief review. *IEEE Trans Multimed* 21(12):3106–3121
102. Zeiler MD, Krishnan D, Taylor GW, Fergus R (2010) Deconvolutional networks. In: 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, pp 2528–2535
103. Zhang L, Zhang H, Shen H, Li P (2010) A super-resolution reconstruction algorithm for surveillance images. *Signal Process* 90(3):848–859
104. Zhang Y, Fu H, Dellandréa E, Chen L (2017) Adapting convolutional neural networks on the shoeprint retrieval for forensic use. In: Chinese conference on biometric recognition. Springer, pp 520–527
105. Zheng H, Wang R, Ji W, Zong M, Wong WK, Lai Z, Lv H (2020) Discriminative deep multi-task learning for facial expression recognition. *Inf Sci* 533:60–71
106. Zhu J, Shen Y, Zhao D, Zhou B (2020) In-domain gan inversion for real image editing. In: European conference on computer vision. Springer, pp 592–608