

SOFT-INTROVAE FOR CONTINUOUS LATENT SPACE IMAGE SUPER-RESOLUTION

Zhi-Song Liu, Zijia Wang, Zhen Jia

Dell Research

ABSTRACT

Continuous image super-resolution (SR) recently receives a lot of attention from researchers, for its practical and flexible image scaling for various displays. Local implicit image representation is one of the methods that can map the coordinates and 2D features for latent space interpolation. Inspired by Variational AutoEncoder, we propose a Soft-introVAE for continuous latent space image super-resolution (SVAE-SR). A novel latent space adversarial training is achieved for photo-realistic image restoration. To further improve the quality, a positional encoding scheme is used to extend the original pixel coordinates by aggregating frequency information over the pixel areas. We show the effectiveness of the proposed SVAE-SR through quantitative and qualitative comparisons, and further, illustrate its generalization in denoising and real-image super-resolution.

Index Terms— Introspective Variational AutoEncoder, super-resolution, latent space

1. INTRODUCTION

Image super-resolution (SR) aims to enlarge the low-resolution (LR) images to the larger desired high-resolution (HR) images. It is widely used in digital display, broadcast and data compression/restoration. With various display devices and data resolutions, a flexible arbitrary image SR model can adjust different needs to produce image/videos with best visual experiences. Most existing state-of-the-art SR methods [1, 2, 3, 4, 5, 6] either focus on fixed super-resolution solutions (one model for one dedicated upsampling scenario), or integral upsampling scales ($2\times$, $4\times$ or $8\times$). They result in costly training efforts and imperfect image resolutions.

Instead, continuous image super-resolution [7, 8] provides arbitrary image/video scaling with photo-realistic visual quality. The goal is to discover the hidden latent space where the missing pixels can be estimated by the continuous feature representation. The advantage is that it can adjust size-varied display devices and reduce many training efforts when applying out-of-distribution super-resolution tasks.

However, continuous image super-resolution tends to generate over-smooth images and is sensitive to noise. In order to produce clean photo-realistic super-resolution images,

we propose Soft-introVAE for continuous latent space image super-resolution (SVAE-SR). The novelty is to use an autoencoder to discover the continuous image distribution space, where we condition the continuous LR features for supervision. The reused encoder works adversarially against the decoder for discriminating between real and generated samples. A soft threshold function is utilized to replace the hard margin in the evidence lower bound (ELBO). Furthermore, the positional encoding is embedded as the frequency expansion to introduce more pixels for prediction. To sum up, our key claims are 1) Soft-introVAE for arbitrary image super-resolution that can measure the adversarial conditional distribution of SR and HR images for reconstruction, and 2) a positional encoding scheme is modified to involve more neighborhood pixels for estimation.

2. RELATED WORKS

Implicit neural representation. The idea of implicit neural representation is to use multi-layer perceptron (MLP) to learn pixels or other signals from coordinates. It has been widely used in 3D shape modeling [9, 10], surface reconstruction [11, 12], novel view rendering [13, 14, 15, 16] and so on. For instance, Mildenhall et al. [13] propose to map the camera poses to the pixel values via a multi-layer MLP network. They use multiple-view images to optimize the network for implicit feature representation. Instead of using a voxel or point cloud, the implicit neural representation can 1) capture the fine details of scenes for photo-realistic reconstruction, and 2) also emit complex 3D representation as a small number of differentiable network parameters.

Image super-resolution. The goal of image super-resolution is to enlarge LR images to the desirable resolutions with high visual quality. Depending on the metrics of evaluation, it can categorize into distortion-based SR [1, 2, 3, 4, 17, 18] and perception-based SR [5, 6, 19]. For the former one, using deeper neural networks with sophisticated designs usually leads to lower pixel distortions. For example, EDSR [2] proposes a network using more convolutional kernels and layers for optimization. RDN [4] proposes a residual dense network to allow feature sharing. Most recently, attention [20] is also used in image SR [17, 18] to involve more neighborhood pixels for estimation. For perception-based SR, GAN [21] and VAE [22] are two major architectures used for photo-realistic

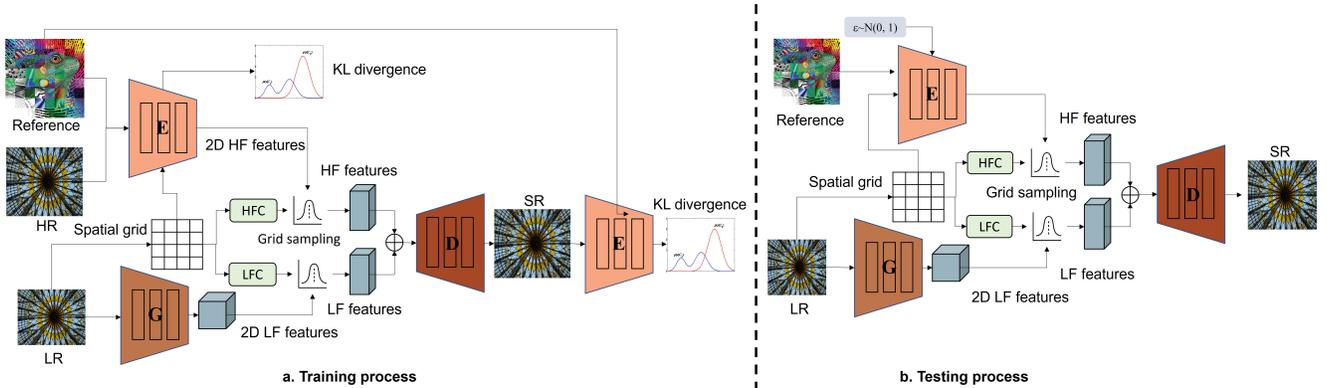


Fig. 1: SVAE-SR. (a) At training, it uses an encoder (E), a decoder (D) and the feature extractor (G) to learn the continuous latent space distribution. Given the reference image as the conditional feature, the encoder works as a discriminator to identify whether the SR distribution is close to the HR distribution. The decoder learns the local implicit image function to have arbitrary scales of upsampling. (b) At test time, SVAE-SR samples a random vector in the trained latent space and combines it with the reference features for high-frequency signal estimation. A final LIIF is trained to map the continuous features to the estimated RGB pixels.

reconstruction. ESRGAN [5] proposes a patch based GAN network to supervise the perceptual quality via a binary discriminator. SRVAE [6] achieves the goal using a conditional VAE to minimize the distribution divergence. However, all the methods are for fixed upsampling factors, like $2\times$ and $4\times$. To have an arbitrary super-resolution solution, LIIF [7] proposes to use the Local Implicit Image Function (LIIF) to explore continuous feature representation for SR. MetaSR [8] uses a Meta-Upscale Module to weight the upsampling filters for prediction dynamically.

3. METHOD

Here, we describe the proposed Soft-Intro Variational AutoEncoder for continuous latent space image Super-Resolution (SVAE-SR). SVAE-SR performs continuous super-resolution by projecting image features into the latent space, and it conditions the LR features with random coordinate space sampling and transfers the new features to the decoder for reconstruction. The encoder is reused as a discriminator to distinguish the distributions of HR (real) and SR (fake) images.

Overview. To learn the ground truth HR images \mathbf{Y} , SVAE-SR is a feed-forward network (shown in Figure 1) that reconstructs SR images \mathbf{Y}' from LR images \mathbf{X} and reference images \mathbf{R} . It consists of an encoder (E), a decoder (D) and a feature extractor (G). The encoder learns the latent vector of the joint HR-Reference distribution. The decoder uses the local implicit image function (LIIF) to transform the 2D features into pixel-coordinate pairs for 2D grid based interpolation. The proposed positional encoding scheme process the 2D grid by high-frequency coding (HFC) and low-frequency coding (LFC) to split the frequency bands. The LFC is with the LR features for low-frequency component reconstruction and the HFC is with the latent features for high-frequency reconstruction.

Soft-IntroVAE. Overall, SVAE-SR works as a soft-IntroVAE to explicitly extract the latent distribution for pixel reconstruction. Different from existing CNN/GAN based SR methods [1, 5], using VAE [6] has been proven to be useful for robust photo-realistic quality restoration for real image SR. Soft-IntroVAE [23] demonstrates its outstanding performance in image generation. By combining the advantages of VAE and GAN, it adversarially optimizes the encoder as an introspective discriminator. It fits the task of image SR that it can model the pixel correlations as multivariate Gaussian distributions $z \sim Q_{\omega}(z|\mathbf{X}) = N(z; \mu_i, \sigma_i^2 I)$. We further modify it to a conditional Soft-IntroVAE model that it uses a reference image as the condition to learn the missing high-frequency signal. Formally, we give the mathematical description of the proposed conditional Soft-IntroVAE as,

$$\begin{aligned}
 L_{E_{\phi}}(\mathbf{Y}|\mathbf{R}, z) &= ELBO(\mathbf{Y}|\mathbf{R}) - \frac{1}{\alpha} \exp(\alpha ELBO(D_{\theta}(z))) \\
 L_{D_{\theta}}(\mathbf{Y}|\mathbf{R}, z) &= ELBO(\mathbf{Y}|\mathbf{R}) + \gamma ELBO(D_{\theta}(z)) \\
 \text{where } ELBO(\mathbf{Y}|\mathbf{R}) &= \mathbb{E}_{z \sim Q(z|\mathbf{Y}, \mathbf{R})} [\log P_{\theta}(\mathbf{Y}, \mathbf{R}|z)] \\
 &- D_{KL}[Q_{\phi}(z|\mathbf{Y}, \mathbf{R})||P(z)] \leq \log P_{\theta}(\mathbf{Y}|\mathbf{R})
 \end{aligned} \tag{1}$$

In Eq (1), it can be seen that the process includes two steps: 1) fix the decoder and optimize the encoder to distinguish through the ELBO value, between HR images (high ELBO) and SR images (low ELBO) and 2) fix the encoder and optimize the decoder to “fool” the encoder with photo-realistic SR images. In such a way, we can push the distribution of the SR data close to the HR data. Note that the natural advantage of using Soft-IntroVAE is that we use a **soft exponential function over the ELBO** to improve the training stability. Meanwhile, using a reference image as the condition to learn the joint $p(\mathbf{Y}|\mathbf{R})$ probability so that the matched features from the reference images can be extracted for aiding reconstruction.

For clarification, the detailed structure of the encoder and decoder is shown in Figure 2. For the encoder, it takes the reference, HR/SR image and 2D grid as inputs. A *to pixel*

sampler converts them to pixel-coordinate 1D vectors. The self-attention and cross-attention is applied to learn the latent distribution. The decoder combines the HF signal and LF signal together to form the final SR image.

Positional Encoding for frequency band splitting. Another key component of the proposed SVAE-SR is positional encoding (PE). Inspired by [13], we expand the 2D grid map to much wider frequency bands. It can better fit the data with a high-frequency variation. It is specifically useful for image SR because the missing details around the edges and textures are high-frequency signals. Neural networks are biased to low-frequency reconstruction, which leads to over-smooth visual quality. Using PE can explicitly lift the feature for high-frequency mapping. Mathematically, given signal o ,

$$\lambda(o) = \left(\sin(2^0\pi o), \cos(2^0\pi o), \dots, \sin(2^{L-1}\pi o), \cos(2^{L-1}\pi o) \right) \quad (2)$$

Eq (2) maps the signal to an L -degree frequency band. We further split it into two parts, the first $L//2$ frequency bands for the low-frequency coding (LFC) and the rest for the high-frequency coding (HFC).

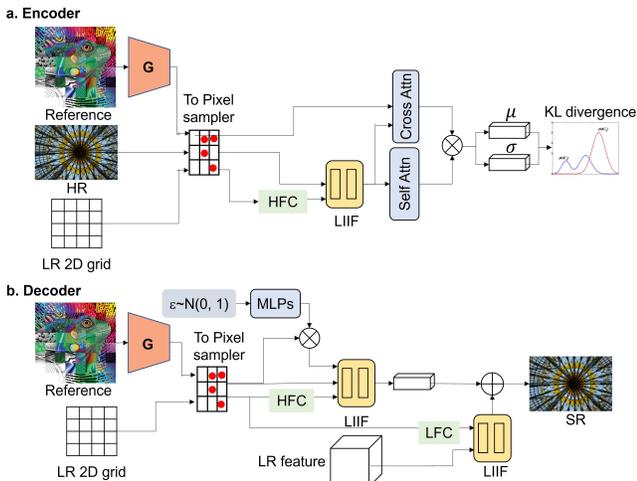


Fig. 2: The detailed structure of the encoder and decoder. The encoder takes the HR/SR image and reference image as inputs to compute their features correlations for joint distributions. The high-frequency positional Encoding (HFC) ensures the estimated features expand to the high-frequency bands. The decoder combines the LR features and sampled high-frequency features and pass them to the LIIF model for pixel estimation.

Training Loss. We train SVAE-SR using the l_1 loss between SR and HR images and KL divergence as follows:

$$L = L_1 + \lambda \|Y' - Y\|^1 + \beta KL[Q_\phi(z|\mathbf{Y}|\mathbf{R})||N(0, 1)], \quad (3)$$

where λ and β are the weighting parameters to balance pixel distortion and KL losses.

Table 1: PSNR comparison between ours and other state-of-the-art methods in various upsampling scales. RDN trains different models for different scales. MetaSR, LIIF, and ours use one model for all scales and are trained with continuous random scales uniformly sampled in $\times 1 \sim \times 4$. We also test on out-of-distribution scenarios in $\times 6$ and $\times 8$.

Dataset	Method	In-distribution			Out-of-distribution	
		$\times 2$	$\times 3$	$\times 4$	$\times 6$	$\times 8$
Set5	RDN	38.24	34.71	32.47	-	-
	RDN-MetaSR	38.22	34.63	32.38	29.04	26.96
	RDN-LIIF	38.17	34.68	32.50	29.15	27.14
	SVAE-SR (ours)	38.23	34.72	32.60	29.23	27.24
Set14	RDN	34.01	30.57	28.81	-	-
	RDN-MetaSR	33.98	30.54	28.78	26.51	24.97
	RDN-LIIF	33.97	30.53	28.80	26.64	25.15
	SVAE-SR (ours)	34.01	30.58	28.86	26.72	25.23
B100	RDN	32.34	29.26	27.72	-	-
	RDN-MetaSR	32.33	29.26	27.71	25.90	24.83
	RDN-LIIF	32.32	29.26	27.74	25.98	24.91
	SVAE-SR (ours)	32.36	29.32	27.80	26.06	5.00
Urban100	RDN	32.89	28.80	26.61	-	-
	RDN-MetaSR	32.92	28.82	26.55	23.99	22.59
	RDN-LIIF	32.87	28.82	26.68	24.20	22.79
	SVAE-SR (ours)	32.92	28.88	26.73	24.29	22.87

4. EXPERIMENTS

Implementation details. We train our model on DIV2K [24] and Flickr2K [2] datasets. They both contain images with resolutions larger than 1000×1000 . Same as [7], we first extract LR patches with the size of 48×48 . Then we randomly sample upsampling scales α in uniform distribution $U(1, 4)$. The corresponding HR patches $48\alpha \times 48\alpha$ are then used to sample 48^2 pixels to form coordinate-RGB pairs. For the reference image, we randomly choose one from Wikiart [25] that is widely used in style transfer. The testing datasets include Set5 [26], Set14 [27], Urban100 [28] and DIV2K validation [24]. We train with Adam optimizer with a learning rate of 10^{-4} and a batch size of 32 for 100k iterations (8hrs on two NVIDIA V100 GPUs).

General image super-resolution. SVAE-SR performs efficient super-resolution with high reconstruction quality. To show its effectiveness, we compare it to four state-of-the-art methods: Bicubic, EDSR [2], RDN [4], LIIF [7] and MetaSR [8] in Table 1. For demonstration, we use continuous upsampling scales α in uniform distribution $U(1, 4)$, and we also test on out-of-training-distribution scenarios, where larger unseen upsampling scales, namely $6 \times \sim 8 \times$, are evaluated on unknown images. We can find that using our proposed method can achieve arbitrary image enlargement with superior performance across different datasets. Especially on out-of-distribution scenarios, ours performs even better with $+0.08 \sim +0.09$ dB in PSNR.

We show visual comparisons in Figure 3. It can be seen that our method can restore the fine textures with better visual quality, like the windows in *78004*, floor strides in *148024*, lighting rays in *0828* and the handrails in *0851*.

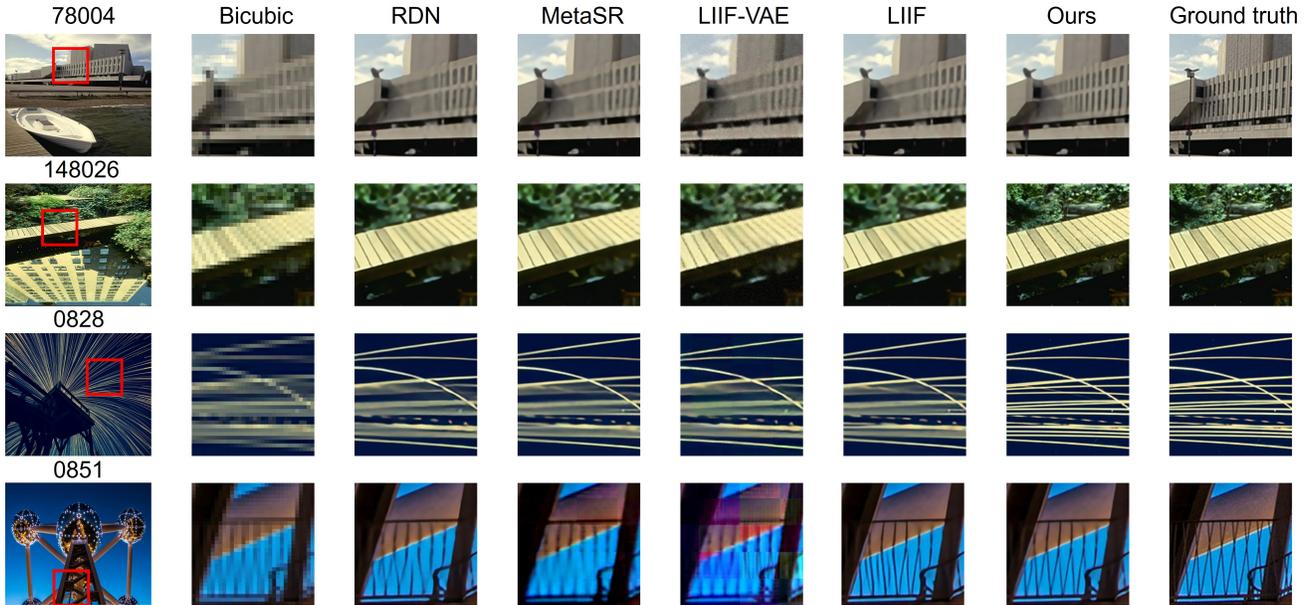


Fig. 3: Visual comparison to the state of the arts. We show $4\times$ super-resolution on BSD100 (78004, 148026) and DIV2K-validation (0828,0851) datasets. We enlarge the regions in red boxes for better visualization.

Table 2: Ablation studies on our methods with or without Soft-IntroVAE and PE structures for image SR. The tests are done on DIV2K validation in PSNR(dB).

Method	In-distribution			Out-of-distribution		
	$\times 2$	$\times 3$	$\times 4$	$\times 6$	$\times 12$	$\times 18$
Bicubic	31.01	28.22	26.66	24.82	22.27	21.00
LIIF	34.99	31.26	29.27	26.99	23.89	22.34
LIIF-VAE	34.89	31.20	29.16	26.90	23.81	22.29
LIIF-GAN	34.78	31.12	29.03	26.58	23.53	22.07
LIIF-PE	35.01	31.29	29.30	27.02	23.93	22.37
SVAE-SR w/o PE	35.02	31.30	29.31	27.04	24.01	22.42
SVAE-SR w/ PE	35.05	31.33	29.35	27.08	24.06	22.47

Ablation studies. Effect of Soft-IntroVAE for SR. To show the effectiveness of the proposed SVAE-SR, we compare it with LIIF (baseline with pure MLP structure), LIIF-VAE (LIIF with ordinary VAE structure), LIIF-GAN (LIIF with ordinary GAN structure) and proposed SVAE-SR. From rows 2 to 4 and SVAE-SR w/o PE in Table 2, we can see that with the same LIIF network, using Soft-IntroVAE (SVAE-SR w/o PE) achieve $+0.03 \sim +0.12dB$ in PSNR. On the other hand, using LIIF-GAN actually obtains lower PSNR, about $-2 \sim -4dB$.

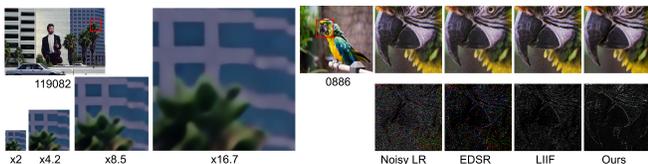


Fig. 4: Visualization on 1) $2\times \sim 16\times$ continuous upsampling, and 2) image super-resolution on the noisy image with noise level $\tau = 1.5$. We show the residual image between SR and HR images to enhance the differences.

Effect of positional encoding for SR. To illustrate the effect of the positional encoding for SR. From Table 2, we conduct the comparisons among LIIF (baseline with pure MLP structure) (row 2), LIIF-PE (baseline with positional encoding) (row 5) and ours with or without PE (row 6 and 7). We can see that using PE can improve the reconstruction quality by about $+0.03 \sim 0.05dB$ in PSNR.

In Figure 4, we show two more our results on continuous image super-resolution between $2\times \sim 16\times$. We can see that the window patterns are well restored across different scenarios. We also mentioned that the proposed SVAE-SR has the ability to overcome the noise for real image SR. We show one example using LR image with Gaussian random noise of intensity 1.5. We can see that our method achieves better results compared to other methods.

Computational cost. Our method achieves relative real-time inference in a Nvidia V100 GPU, approximately 0.3s in $4\times$ upsampling and 1.1s in $16\times$ upsampling.

5. CONCLUSION

In this paper, we propose a Soft-IntroVAE for continuous image super-resolution (SVAE-SR). It combines the advantages of VAE and LIIF to explore the continuous latent space interpolation, which results in arbitrary upsampling with photo-realistic visual quality. In the meantime, the proposed positional encoding scheme expands the signal to much wider frequency bands, which avoids the network bias in the low-frequency domain. Experimental results on qualitative and quantitative comparisons show that our proposed SVAE-SR achieves outstanding performance and points in a promising direction in robust real-image super-resolution.

6. REFERENCES

- [1] Zhi-Song Liu, Li-Wen Wang, Chu-Tak Li, Wan-Chi Siu, and Yui-Lam Chan, "Image super-resolution via attention based back projection networks," in *Proc. ICCV-Workshops*, 2019, pp. 3517–3525.
- [2] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. CVPR-Workshops*, 2017, pp. 1132–1140.
- [3] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," *Proc. ECCV*, 2018.
- [4] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image restoration," *Proc. CVPR*, vol. 43, no. 7, pp. 2480–2495, 2021.
- [5] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proc. ECCV-Workshops*, September 2018.
- [6] Wan-Chi Siu Zhi-Song Liu, Li-Wen Wang, Chu-Tak Li, Marie-Paule Cani, and Yui-Lam Chan, "Unsupervised real image super-resolution via generative variational autoencoder," in *Proc. CVPR-Workshops*, June 2020.
- [7] Yinbo Chen, Sifei Liu, and Xiaolong Wang, "Learning continuous image representation with local implicit image function," in *Proc. CVPR*, 2021, pp. 8624–8634.
- [8] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun, "Meta-sr: A magnification-arbitrary network for super-resolution," in *Proc. CVPR*, 2019, pp. 1575–1584.
- [9] Zhiqin Chen and Hao Zhang, "Learning implicit fields for generative shape modeling," in *Proc. CVPR*, 2019, pp. 5932–5941.
- [10] Mateusz Michalkiewicz, Jhony Kaesemodel Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson, "Implicit surface representations as layers in neural networks," in *Proc. ICCV*, 2019, pp. 4742–4751.
- [11] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas Funkhouser, "Local implicit grid representations for 3d scenes," in *Proc. CVPR*, 2020, pp. 6000–6009.
- [12] Peng Songyou, Niemeyer Michael, Mescheder Lars, Pollefeys Marc, and Geiger Andreas, "Convolutional occupancy networks," in *Proc. ECCV*, 2020.
- [13] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Proc. ECCV*, 2020.
- [14] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger, "Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision," in *Proc. CVPR*, 2020, pp. 3501–3512.
- [15] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger, "Texture fields: Learning texture representations in function space," in *Proc. ICCV*, 2019, pp. 4530–4539.
- [16] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proc. CVPR*, 2022, pp. 5460–5469.
- [17] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen, "Single image super-resolution via a holistic attention network," in *Proc. ECCV*, 2020, p. 191–207.
- [18] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang, "Second-order attention network for single image super-resolution," in *Proc. CVPR*, 2019, pp. 11065–11074.
- [19] Wan-Chi Siu Zhi-Song Liu and Li-Wen Wang, "Reference based image super-resolution via variational autoencoder," in *Proc. CVPR-Workshops*, June 2021.
- [20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS*. 2017, vol. 30, Curran Associates, Inc.
- [21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NeurIPS*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 2672–2680. Curran Associates, Inc., 2014.
- [22] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," 2014.
- [23] Tal Daniel and Aviv Tamar, "Soft-introvae: Analyzing and improving the introspective variational autoencoder," in *Proc. CVPR*, June 2021, pp. 4391–4400.
- [24] Radu Timofte and et al., "Ntire 2017 challenge on single image super-resolution: Methods and results," in *Proc. CVPR*. 8 2017, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1110–1121, IEEE Computer Society.
- [25] "K. nichol. painter by numbers, wikiart," <https://www.kaggle.com/c/painter-by-numbers>, 2016.
- [26] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie line Alberi Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proceedings of the British Machine Vision Conference*. 2012, pp. 135.1–135.10, BMVA Press.
- [27] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," 06 2010, vol. 6920, pp. 711–730.
- [28] J. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. CVPR*, June 2015, pp. 5197–5206.