Exploring Open Domain Image Super-Resolution through Text

Kanchana Vaishnavi Gandikota [*1] **Paramanand Chandramouli** [*1]

Abstract

In this work, we propose for the first time a zero-shot approach for flexible open domain extreme super-resolution of images which allows users to interactively explore plausible solutions by using language prompts. Our approach exploits a recent diffusion based text-to-image (T2I) generative model. We modify the generative process of the T2I diffusion model to analytically enforce data consistency of the solution and explore diverse contents of null-space using text guidance. Our approach results in diverse solutions which are simultaneously consistent with input text and the low resolution images.

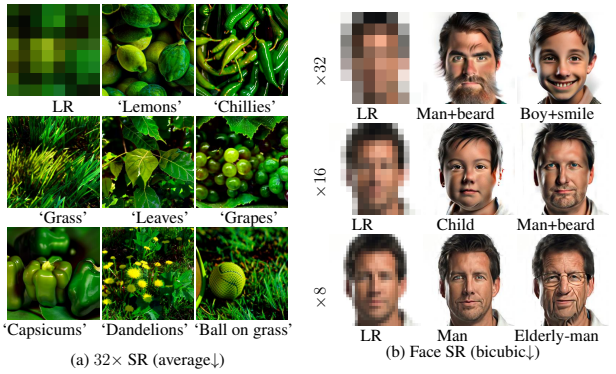

Figure 1. Text guided exploration of multiple perfectly consistent solutions to image super-resolution. Example solutions for input prompts of the form *'A photograph of* {*key word*}*'* are depicted.

1. Introduction

The goal of image super-resolution is to recover a high-quality image, given a low-resolution (LR) observation $\mathbf{y}$.

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{A}$, $\mathbf{x}$ and $\mathbf{n}$ represent the down-sampling operator, ground truth image and measurement noise respectively. Image super-resolution is highly ill-posed, especially at large super-resolution factors with many valid solutions satisfying the data consistency accurately. Yet, recent state-of-the-art supervised deep networks for super-resolution (Chan et al., 2021; Wang et al., 2022) recover only a single image from this solution space. On the other hand, deep learning based stochastic estimators also exist (Lugmayr et al., 2020; Li et al., 2022; Kawar et al., 2022; Wang et al., 2023), which use conditional or unconditional generative models to sample from the solution space. A few works also allow exploring the solution space, using graphical user inputs (Bahat & Michaeli, 2020) or semantic maps (Buhler et al., 2020), or text (Ma et al., 2022). Yet, even these methods are trained for specific classes and super-resolution factors. A flexible method which allows text-guided exploration of solutions space for various downsampling factors on open domain images does not exist.

In this paper, we propose for the first time a zero-shot approach to open domain image super-resolution using simple and intuitive text prompts. Our goal is to explore diverse, semantically accurate reconstructions which preserve data consistency with the low-resolution inputs for different large downsampling factors without explicitly training for these specific degradations. Towards this goal, we utilize a recent diffusion based text-to-image (T2I) generative model DALL-E2 (Ramesh et al., 2022) and adapt it for super-resolution, by modifying the reverse diffusion process. We incorporate the range space-null space decomposition (Schwab et al., 2018; Bahat & Michaeli, 2020; Chen & Davies, 2020; Wang et al., 2023) into the reverse diffusion to analytically enforce data consistency of the solutions, while exploring diverse contents of null-space as proposed in (Kawar et al., 2022; Wang et al., 2023). As text guided diffusion takes places in a down-sampled pixel space in DALL-E2, we adopt a two stage approach in enforcing the data consistency by modifying the measurement model accordingly in the down-sampled space. We propose an embeddings averaging trick to align text guidance with the input observation and improve reconstruction quality. Fig. 1 demonstrates that this approach can successfully recover solutions with high data consistency and semantic consistency with the text input.

[*]Equal contribution [1]Department of Computer Science, University of Siegen, Germany. Correspondence to: Kanchana Vaishnavi Gandikota <vaishnavigandikota[at]gmail.com>, Paramanand Chandramouli <paramanand[at]gmail.com>.

AI & HCI Workshop at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA, 2023.

2. Background

2.1. Denoising Diffusion Probabilistic Models (DDPM)

DDPM generative models (Ho et al., 2020) employ two diffusion processes: *i) A forward process* slowly noising a data sample $\boldsymbol{x}_0$ into Gaussian distribution $\mathcal{N}$ in T steps, with the evolution of a sample $\mathbf{x}_t$ at time-step t given by:

$$\begin{aligned} q(\mathbf{x}_t|\mathbf{x}_{t-1}) &:= \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \\ i.e., \quad \mathbf{x}_t &= \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \end{aligned} \tag{2}$$

where $\{\beta_t\}_{t=0}^T$ is the noise variance schedule. *ii) A learned reverse process* using iterative denoising to generate samples from the training data distribution $q(\mathbf{x})$ in T steps given by:

$$\begin{aligned} &p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \sigma_t^2\mathbf{I}), \quad \text{where} \\ &\boldsymbol{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\right) \quad \text{and} \\ &\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \ \text{ with } \ \alpha_t = 1-\beta_t, \ \text{and } \bar{\alpha}_t = \prod_{i=0}^{t}\alpha_i. \end{aligned} \tag{3}$$

Clean images are generated by iterative sampling (3) in the reverse diffusion process exploiting the learned neural network noise approximator $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$.

2.2. Range Space-Null Space Decomposition

When there is no measurement noise in (1), i.e. $\mathbf{y} = \mathbf{A}\mathbf{x}$, pseudoinverse operation $\mathbf{A}^\dagger\mathbf{y}$ produces the minimum norm solution with perfect data consistency. Any other sample of form $(\mathbf{A}^\dagger\mathbf{y} + \mathbf{x}_\delta)$ is also data consistent, as long as $\mathbf{x}_\delta$ lies in the null space of $\mathbf{A}$. Note that $\mathbf{x}$ can be decomposed as:

$$\mathbf{x} \equiv \mathbf{A}^\dagger\mathbf{A}\mathbf{x} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}. \tag{4}$$

Here $\mathbf{A}^\dagger\mathbf{A}\mathbf{x}$ is in the range space of $\mathbf{A}$ (with $\mathbf{A}\mathbf{A}^\dagger\mathbf{A}\mathbf{x} \equiv \mathbf{y}$) and the component $(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}$ is in the null space of $\mathbf{A}$, (with $\mathbf{A}(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x} \equiv \mathbf{0}$). Given an approximate solution $\bar{\mathbf{x}}$, Eq. 4 can be used to construct a data consistent solution (Bahat & Michaeli, 2020; Wang et al., 2023) given by $\hat{\mathbf{x}}$

$$\hat{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\bar{\mathbf{x}}. \tag{5}$$

2.3. Null Space Consistency in Diffusion Models

(Wang et al., 2023; Kawar et al., 2022) utilize the range space-null space decomposition in the reverse diffusion process to obtain data consistent solutions. At time step t,

$$\mathbf{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\sqrt{1-\bar{\alpha}_t}\right) \tag{6}$$

gives the estimate of the clean image. In noise-less case, a rectified data consistent estimate $\hat{\mathbf{x}}_{0|t}$ is obtained as:

$$\hat{\mathbf{x}}_{0|t} = \mathbf{A}^\dagger\mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}_{0|t}. \tag{7}$$

This rectified data consistent estimate is used in subsequent sampling from $p(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathbf{x}}_{0|t})$ in (Wang et al., 2023).

2.4. DALL-E2 unCLIP

The T2I model DALL-E2 (Ramesh et al., 2022) uses two main modules for text guided image generation: *i) a diffusion based prior* to produce CLIP image embeddings (Radford et al., 2021) z_i from encodings of the input prompt c, with z_i encapsulating what the prior *sees* from the text prompt. *ii) a conditional diffusion based decoder* to generate images conditioned on CLIP image embeddings and text embeddings $\mathbf{z} = \{z_i, z_t\}$. The framework is referred to as unCLIP, as it generates images by inverting CLIP embeddings. Text conditioned diffusion is performed in a down-sampled pixel space for improved computational efficiency, yielding a lower resolution image. This is subsequently super-resolved in a diffusion based module to obtain a higher resolution output.

3. Method

Given a low resolution image $\mathbf{y}$ with known downsampling operator $\mathbf{A}$, our goal is to generate data consistent solutions $\hat{\mathbf{x}}$ whose attributes can be varied using input text prompts c.

$$\begin{aligned} \textit{Data Consistency}: \quad & \mathbf{A}\hat{\mathbf{x}} \equiv \mathbf{y}, \\ \textit{Semantic Consistency}: \quad & \hat{\mathbf{x}} \sim q(\mathbf{x}|\mathbf{c}), \end{aligned} \tag{8}$$

where $q(\mathbf{x}|\mathbf{c})$ denotes the distribution of images with semantic meaning provided by the text prompt $\mathbf{c}$. Towards this goal we employ null space consistency enforcement described in Sec. 2.3 in the conditional reverse diffusion process of the unCLIP model. To take into account two-stage diffusion in theunCLIP model, we adopt a two-stage consistency enforcement. We first recover a lower resolution $\mathbf{x}_{LR}$ using a modified measurement $\mathbf{A}_{LR}$ in the down-sampled space of text-conditioned decoder $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$, conditioned on embeddings $\mathbf{z}$ produced from the text input $\mathbf{c}$. The current estimate of the low resolution clean image at each step is given by:

$$\mathbf{x}_{LR_{0|t}} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_{LR_t} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_{LR_t}, t|\mathbf{z})\sqrt{1-\bar{\alpha}_t}\right). \tag{9}$$

and the consistency rectified estimate is given as

$$\hat{\mathbf{x}}_{LR_{0|t}} = \mathbf{A}^\dagger{}_{LR}\mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger{}_{LR}\mathbf{A}_{LR})\mathbf{x}_{LR_{0|t}}. \tag{10}$$

For the subsequent diffusion for super-resolution using the model $\boldsymbol{\zeta}_{\boldsymbol{\theta}}$, we consider the actual measurement operator $\mathbf{A}$, with corresponding null space consistency rectification. Our final two stage approach is summarized in Algorithm 1. In practice, we accelerate the reconstruction by starting at an earlier time step $t_0 < T$ for both the reverse diffusion processes, and use fewer number of steps between between $[1, t_0]$ in the reverse diffusion. Due to this acceleration, and text conditioned diffusion in the smaller dimensional pixel space, our method can produce fast reconstructions.

Embeddings Averaging Trick When the image embedding z_i as imagined by the prior does not structurally align

Algorithm 1 Our Approach

$\mathbf{x}_{LR_T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, ..., 1$ **do**
  $\mathbf{x}_{LR_{0|t}} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{x}_{LR_t} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_{LR_t}, t|\mathbf{z}) \sqrt{1 - \bar{\alpha}_t} \right)$
  $\hat{\mathbf{x}}_{LR_{0|t}} = \mathbf{A}^{\dagger}{}_{LR}\mathbf{y} + (\mathbf{I} - \mathbf{A}^{\dagger}{}_{LR}\mathbf{A}_{LR})\mathbf{x}_{LR_{0|t}}$
  $\mathbf{x}_{LR_{t-1}} \sim p_1(\mathbf{x}_{LR_{t-1}}|\mathbf{x}_{LR_t}, \hat{\mathbf{x}}_{LR_{0|t}})$
end for
$\mathbf{x}_{LR} \leftarrow \mathbf{x}_{LR_0}$
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, ..., 1$ **do**
  $\mathbf{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{x}_t - \boldsymbol{\zeta}_{\boldsymbol{\theta}}(\mathbf{x}_t, t|\mathbf{x}_{LR}) \sqrt{1 - \bar{\alpha}_t} \right)$
  $\hat{\mathbf{x}}_{0|t} = \mathbf{A}^{\dagger}\mathbf{y} + (\mathbf{I} - \mathbf{A}^{\dagger}\mathbf{A})\mathbf{x}_{0|t}$
  $\mathbf{x}_{t-1} \sim p(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathbf{x}}_{0|t})$
end for
return $\mathbf{x}_0$

with the observation, it leads to unrealistic images. To alleviate this, we propose to modify z_i for better structural consistency with the degraded input:

$$z_i = (1 - \lambda) z_{i_{prior}} + \lambda \mathcal{E}(\mathbf{A}^{\dagger}\mathbf{y}) \tag{11}$$

where $\mathcal{E}$ is the CLIP encoder used in training the DALL-E2 unCLIP model. This embeddings averaging trick improves structural consistency of the image embedding with the input observation. We found reasonable outputs with $\lambda \in [0, 0.6]$ with lower values of λ for higher SR factors, see Fig. 2 for result with and without this trick.

Implementing $\mathbf{A}$ and $\mathbf{A}^{\dagger}$ In case of down-sampling by averaging with scale n, $\mathbf{A}$ becomes the average pooling operator, and corresponding $\mathbf{A}^{\dagger}$ would replicate the pixels n^2 times. To super-resolve images produced by bicubic down-sampling, we use SVD to construct the pseudo-inverse (Kawar et al., 2022) $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, $\mathbf{A}^{\dagger} = \mathbf{V}\mathbf{\Sigma}^{\dagger}\mathbf{U}^T$.

4. Experiments and Results

We perform our experiments using the publicly available implementation of the unCLIP model (Lee et al., 2022) trained on 115M image-text pairs, with resolutions of 64×64 for text conditioned decoder and 256×256 for the subsequent super-resolution. We use this model directly without further fine-tuning for super-resolution with large SR factors. This problem is severely ill-posed and allows exploration of a larger solution space, and is therefore an ideal setting to test our method on exploring diverse solutions. Fig. 3 shows qualitative results of our approach for natural images and faces. We compare sample reconstructions of our approach with (Wang et al., 2023) trained on CelebaHQ dataset (Karras et al., 2018) on sample face images from Set5 (Bevilacqua et al., 2012), and on an image of an African man. Since both the approaches impose data consistency in reverse diffusion, their solutions achieve LR PSNR values > 50 dB.

Figure 2. $\times 16$ SR results with (bottom) and without (top) averaging trick with λ=0.4, and text prompt 'a high-res photo of a cat'.

However, the vanilla DDNM offers no scope of controlling the output using text. Further, even for a single text input, our results show a greater diversity in pose, backgrounds and lighting and content. On the other hand, solutions of (Wang et al., 2023) show limited diversity indicating limited variations in lighting and pose in the training data. On face images such as the 'baby' or unaligned face of a 'woman' from Set5, the solutions recovered by (Wang et al., 2023) are unsatisfactory, indicating that diffusion model trained on a specific domain, does not generalize well to images that are slightly out of distribution. Even when restoring the down-sampled version of an image of a face of an African man, the solutions of (Wang et al., 2023) either contain some artifacts (for $\times 16$ SR) or exhibit a very limited diversity in terms of pose, expression and perceived race of the reconstructed face image. Recent work (Salminen et al., 2020) on biases in StyleGAN (Karras et al., 2019) demonstrates severe racial bias in generation (generating pictures of white people 72.6%), which is inherited by algorithms using StyleGAN for reconstruction (Menon et al., 2020). It would be interesting to investigate the presence of such biases in diffusion models. In contrast to the diffusion model trained on CelebA-HQ faces, our results demonstrate greater diversity in pose, expression, age, lighting and background, and use of text effortlessly enables reconstruction of faces with varying personal attributes including age, gender, race.

5. Discussion and Conclusion

We proposed for the first time a flexible method for image super-resolution through user provided text prompts through the use of a pretrained text-to-image diffusion model DALL-E2 unCLIP. Our approach generates semantically accurate solutions satisfying perfect data consistency. The performance of the proposed method depends on and is limited by the generative capabilities of DALL-E2 unCLIP, and it inherits the biases of it's training data. Our work opens up a promising direction of developing efficient tools for text

guided exploration of image recovery.

References

Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., and Anderson, P. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8948–8957, 2019.

Bahat, Y. and Michaeli, T. Explorable super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2716–2725, 2020.

Bevilacqua, M., Roumy, A., Guillemot, C., and Alberi-Morel, M. L. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *British Machine Vision Conference*. BMVA press, 2012.

Buhler, M. C., Romero, A., and Timofte, R. Deepsee: Deep disentangled semantic explorative extreme super-resolution. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Chan, K. C., Wang, X., Xu, X., Gu, J., and Loy, C. C. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14245–14254, 2021.

Chen, D. and Davies, M. E. Deep decomposition learning for inverse imaging problems. In *Proc. European Conference on Computer Vision*, pp. 510–526. Springer, 2020.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019. doi: 10.1109/CVPR.2019.00453.

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022.

Lee, D., Kim, J., Choi, J., Kim, J., Byeon, M., Baek, W., and Kim, S. Karlo-v1.0.alpha on coyo-100m and cc15m. `https://github.com/kakaobrain/karlo`, 2022.

Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., and Chen, Y. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

Lugmayr, A., Danelljan, M., Van Gool, L., and Timofte, R. Srflow: Learning the super-resolution space with normalizing flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 715–732. Springer, 2020.

Ma, C., Yan, B., Lin, Q., Tan, W., and Chen, S. Rethinking super-resolution as text-guided details generation. *arXiv preprint arXiv:2207.06604*, 2022.

Menon, S., Damian, A., Hu, S., Ravi, N., and Rudin, C. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 2437–2445, 2020.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Salminen, J., Jung, S.-g., Chowdhury, S., and Jansen, B. J. Analyzing demographic bias in artificially generated facial pictures. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–8, 2020.

Schwab, J., Antholzer, S., and Haltmeier, M. Deep null space learning for inverse problems: convergence analysis and rates. *Inverse Problems*, 35, 2018.

Wang, Y., Hu, Y., and Zhang, J. Panini-net: Gan prior based degradation-aware feature interpolation for face restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2576–2584, 2022.

Wang, Y., Yu, J., and Zhang, J. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations*, 2023.

'A crown with blue jewels and diamonds..'

'A fat blue jay sitting on a broken limb of a leafless tree.'

a) Exploring multiple consistent solutions for the same text prompt $\times 16$ SR on samples from nocaps dataset (Agrawal et al., 2019).

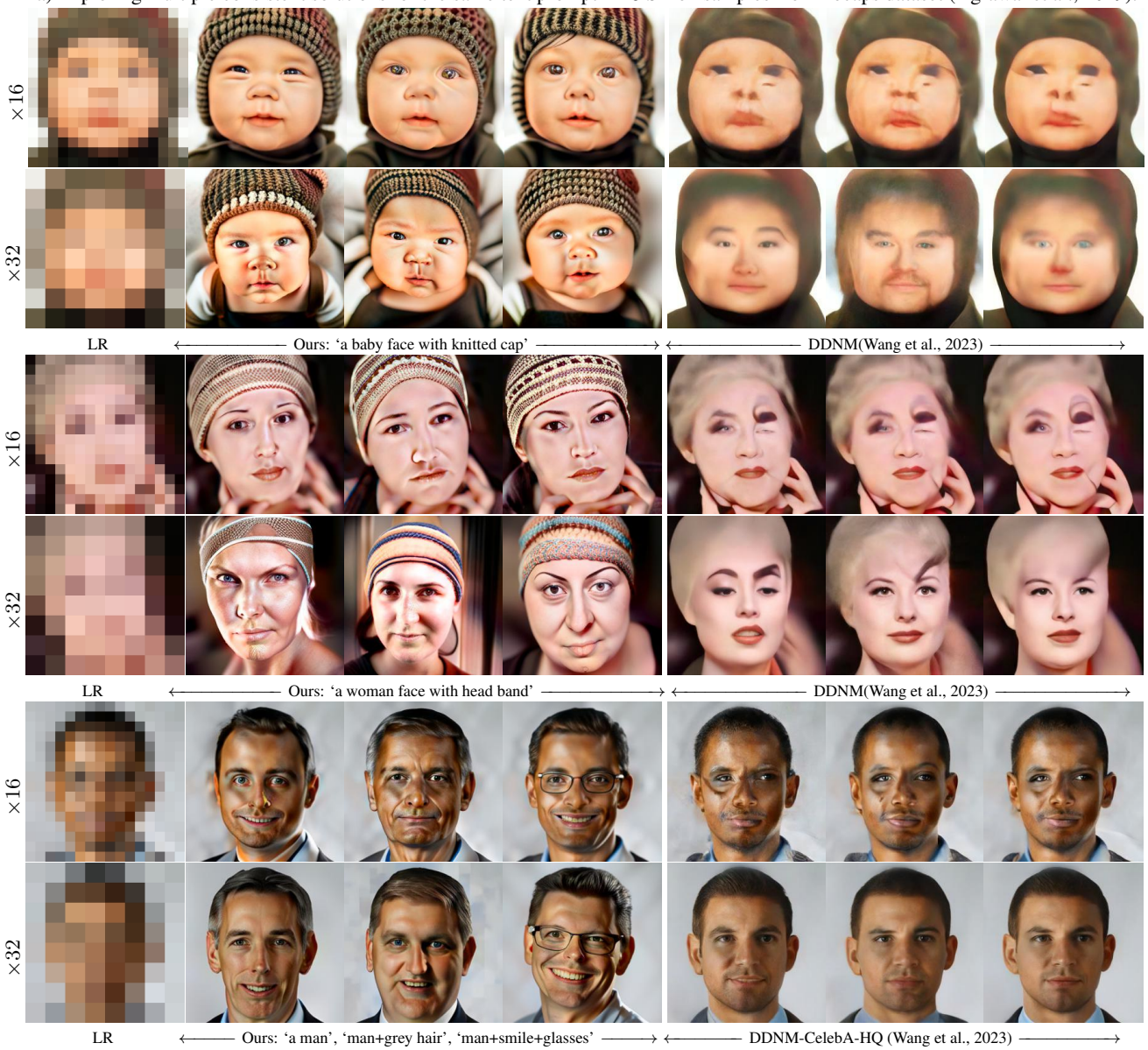

Figure 3. Qualitative evaluation, (top) $\times 16$ SR on natural images, (bottom) comparison with (Wang et al., 2023) on faces.