You Only Landmark Once: U-Net Face Super Resolution with YOLO-World Landmark Heatmaps

Anna Briotto
anna.briotto@studenti.unipd.it

Riccardo Carraro
riccardo.carraro.10@studenti.unipd.it

Endi Hysa
endi.hysa@studenti.unipd.it

Abstract

We propose a lightweight U-Net architecture for face image super-resolution that reconstructs 128×128 *outputs from severely degraded* 16×16 *inputs, corresponding to an* $8\times$ *magnification. The model is trained with a composite objective that combines pixel-wise and perceptual losses with an additional heatmap-guided term. The heatmaps are derived from YOLO-World, a recent open-vocabulary object detector, which localizes key facial components such as the eyes, nose, and mouth. In contrast to prior approaches that require separately trained alignment or landmark networks, our method directly reuses detector outputs to build spatial weights for the loss, requiring no auxiliary training. The heatmaps are used exclusively as supervision, weighting reconstruction errors around semantically important regions. We evaluate the approach on the aligned CelebA dataset, comparing models trained with and without the heatmap loss. Incorporating the heatmap term yields higher PSNR, SSIM, and MS-SSIM, accelerates convergence, and produces reconstructions that are visually sharper and more realistic, particularly in fine facial details. By contrast, extensions with multiscale loss and deep supervision did not lead to measurable gains, suggesting limited benefit for compact architectures. Overall, our findings show that lightweight super-resolution networks can effectively exploit detection-driven priors from YOLO-World, achieving perceptually convincing results under extreme upscaling without relying on adversarial training.*

1. Introduction

Single Image Super-Resolution (SISR) is a fundamental task in computer vision, aiming to reconstruct a high-resolution image from its low-resolution counterpart. High-quality super-resolution has broad applications, ranging from the restoration of degraded media and enhancement of visual content, to aiding face recognition in surveillance and digital forensics [28, 19]. In all these scenarios, the ability to recover fine details in critical regions strongly influences the perceptual quality of the reconstructed image.

In the context of face image super-resolution, accurately reconstructing local facial structures, such as eyes, mouth, and nose, is crucial to maintaining both identity and visual realism. Several methods have explored the integration of landmark or attention information to guide the restoration of these key regions. However, deriving such guidance often requires dedicated face alignment networks or heatmap generators[12, 1], which can introduce complexity and present challenges for generalization and adaptation across different application domains. Some of these approaches, in particular [1], have troubles dealing with occluded landmarks, causing generation of artifacts,

In this work, we propose an approach that mitigates these critical points by using heatmaps extracted from YOLO-World detections [4]. This choice eliminates the need for training an auxiliary network and ensures that attention is focused only on visible features detected in the image. Our approach is evaluated on the aligned CelebA dataset [18] under the same distortion settings used by Kim *et al.*[12], allowing for direct comparison. Furthermore, instead of a GAN-based generator [15], which is widely adopted in the super resolution domain, we employ a U-Net architecture [22], which provides greater stability, faster convergence, and competitive or superior quantitative performance. The proposed pipeline could be applied to other domains where class-specific detections are available, enabling targeted super-resolution for diverse visual tasks such as identity-preserving face enhancement, restoration of partially occluded images, or content-aware image upscaling.

2. Related Work

2.1. Super-Resolution Approaches

Early deep learning-based SISR methods, such as SRCNN by Dong *et al.* [5], demonstrated the potential of convolutional neural networks for end-to-end image upscaling. Subsequent works improved reconstruction quality and efficiency through deeper architectures, residual learning, and enhanced upsampling strategies. For instance, EDSR [17] and its variants pushed state-of-the-art performance by re-

moving unnecessary normalization layers and increasing network depth. GAN-based models, notably SRGAN [15] and ESRGAN [24], introduced adversarial training and perceptual losses to enhance texture realism, often at the expense of training stability. To avoid these issues, we adopt a U-Net architecture [22], chosen for its stable optimization behavior and effectiveness in low-level vision tasks.

2.2. Landmark-guided SR

Landmark-guided approaches to face SR have shown strong performance in face super-resolution by leveraging facial priors; for example, Kim *et al.* [12] introduced a progressive upsampling framework that integrates attention maps obtained from a distilled Face Alignment Network (FAN), while other methods [3, 1] similarly rely on alignment or parsing networks, often requiring dedicated training. In particular, Kim *et al.* [12] use the heatmaps produced over the output generated by the model and calculate the loss with respect to the heatmap of the target image obtained with the distilled FAN.

In contrast, our method leverages recent advances in open-vocabulary object detection. Cheng *et al.* [4] introduced YOLO-World, an extension of YOLO[21] capable of detecting arbitrary user-defined concepts in real time. We repurpose YOLO-World to generate heatmaps of facial landmarks without any task-specific retraining. These heatmaps are used to spatially weight reconstruction errors around landmark regions, providing a lightweight and domain-adaptable alternative to FAN-based supervision. In fact, instead of recomputing the heatmap of the model's output and compare it to the heatmap of the target image, we use the heatmap of the target image to weight more errors committed around relevant areas of the face. This aligns with broader trends in super-resolution that emphasize region-specific priors [28, 25], while avoiding the need to train specialized auxiliary networks.

3. Method

3.1. Dataset

We conduct our experiment on the aligned version of the CelebA dataset [18], a large-scale collection of more than 200,000 celebrity face images with variability in pose, background and illumination. Following Kim *et al.* [12], each 128×128 HR image is bicubically downsampled to 16×16 to generate the LR input, ensuring comparability with prior work. For computational feasibility, we use a 42k subset comprising 30k training, 6k validation, and 6k testing images. Hyperparameters were tuned on a smaller 10k subset before scaling to the full 30k training set. This dataset configuration ensures a fair evaluation againts Kim *et al.* [12]

3.2. Architecture

We propose a lightweight U-Net variant for face super-resolution, mapping 16×16 inputs to 128×128 outputs. An overview of the full model architecture is shown in Figure 1. The design follows the classic encoder–bottleneck–decoder structure [22], with targeted modifications to improve detail preservation, convergence, and output fidelity.

The **encoder** consists of four convolutional blocks with progressively increasing depth. Each block applies two 3×3 convolutions with LeakyReLU activations (α=0.2), using a stride of 2 for spatial downsampling instead of max-pooling. This choice is motivated by prior findings [20] showing that strided convolutions improve learning stability and flexibility in generative models. The number of filters doubles at each stage, starting from 48.

At the deepest level, a **bottleneck** block further processes the compressed representation with two additional convolutional layers.

The **decoder** mirrors the encoder's structure. Each stage performs bilinear upsampling followed by feature concatenation with the corresponding encoder output (skip connection), and a double convolutional block. This strategy has been shown to enhance spatial detail reconstruction in image translation tasks [22, 9].

After restoring the original input resolution (16×16), the model includes a **learned upsampling module** that progressively increases the spatial resolution through three steps: $16\rightarrow32\rightarrow64\rightarrow128$. Each step uses nearest-neighbor interpolation, followed by a 3×3 convolution and LeakyReLU activation. This approach, inspired by Dong *et al.* [6], avoids checkerboard artifacts commonly observed with transposed convolutions and allows sharper reconstructions.

At full resolution (128×128), a **refinement head** composed of five residual blocks [7] is used to fine-tune the output. Residual learning accelerates convergence and improves detail restoration in super-resolution tasks [17]. A 1×1 convolution is applied before the residual stack to reduce feature dimensionality, and a final 1×1 projection layer outputs the RGB image.

Finally, a **global skip connection** adds a bilinearly upsampled version of the low-resolution input directly to the output. This mechanism, commonly adopted in SR models [15, 17], improves convergence and helps maintain color consistency.

The full model contains approximately 7.3 million parameters.

3.3. Heatmap-based loss

Our idea is to compute all the heatmaps of the dataset and use them only as weighted mask over the errors committed in the output, penalizing more errors committed in important face landmark making the model prioritize those

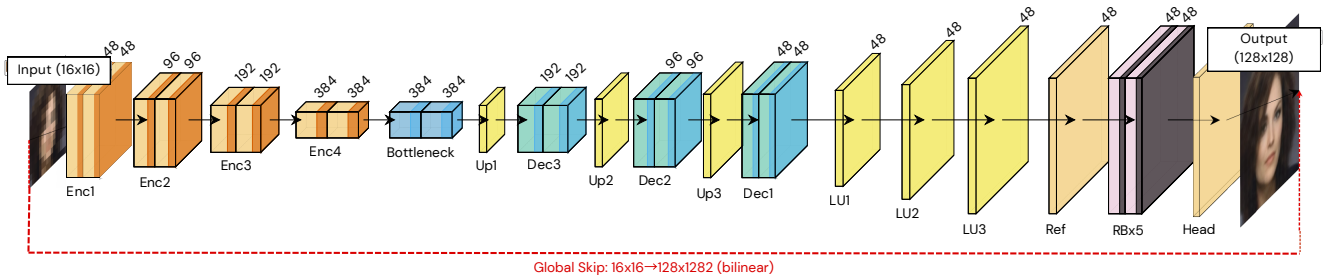

Figure 1. Efficient U-Net architecture for image super-resolution, transforming a low-resolution 16×16 input into a high-resolution 128×128 output performing a $8\times$ upsampling. The encoder blocks (Enc) halve the spatial size at each stage, while the decoder blocks (Dec) and upsampling blocks (Up, LU) progressively restore and increase resolution. The refinement stage (Ref, RBx5) further enhances spatial details. A global skip connection (bilinear interpolation) adds the upsampled input to the output for stability.

regions. We first use YOLO-World [4] to detect facial components in the target images, and subsequently generate pixel-aligned heatmaps. In contrast to Kim *et al.* [12], who rely on distilled FAN landmarks, our approach leverages YOLO-World's open-vocabulary framework to define prompts for identifying facial landmarks, without requiring a task-specific auxiliary network. The time needed to generate 50k heatmaps is about 30 minutes on a Tesla T4 GPU, using YOLO-World [4], specifically the lightweight YOLOv8s-World variant deployed via the official Ultralytics framework [10].

Each detected region is cropped and processed to enhance structural details: we apply Scharr [23] and Canny [2] edge detectors, normalize the responses, and smooth them with a Gaussian blur (kernel size 15×15, $\sigma = 3$). To avoid assigning uniform importance across the entire bounding box, we modulate the edge maps with spatial fading functions with respect to the pixel coordinates (x, y), normalized to $[-1, 1]$ relative to the box center:

- **Gaussian fade:** $f(x, y) = \exp(-2(x^2 + y^2))$, emphasizing the center of small features (e.g., eyes, nose, mouth).
- **Inverse Gaussian fade:** $f(x, y) = 1 - \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)$ with $\sigma = 0.6$, emphasizing the contours of larger regions (e.g., face, head).

Class-specific weights were manually assigned based on qualitative inspection of the output heatmaps, with the goal of emphasizing visually prominent facial regions. We set the weight to 4.5 for eyes; 4.0 for eyebrows, mouth, nose, chin, ears, and face; 3.0 for nose tip; and 2.0 for head. The contributions from all detections are accumulated into a combined and normalized heatmap, that matches the resolution of the target image, as shown in figure 2. This enables a direct use as weighting masks in the reconstruction loss. Specifically, we define a weighted pixel loss that reweights reconstruction errors according to the pre-

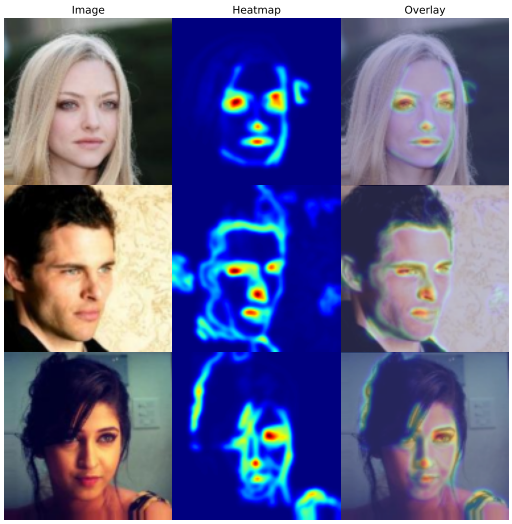

Figure 2. Heatmaps generatad with YOLO-World

computed YOLO-World heatmaps. Let $H \in [0, 1]^{H \times W}$ be the normalized heatmap. The weighted loss is:

$$\mathcal{L}_{\text{att}} = \frac{1}{N} \sum_{i=1}^{N} \frac{w_i \cdot |\hat{I}_i - I_i|}{\sum_{j=1}^{N} w_j}, \tag{1}$$

where $w_i = \mathit{floor} + (1 - \mathit{floor}) \cdot H_i^{\gamma}$, $\hat{I}_i$ and I_i are the predicted and ground-truth pixel intensities, N is the total number of pixels, $\gamma > 1$ controls the selectivity of hot regions, and $\mathit{floor} \in (0, 1]$ ensures non-zero weight outside the salient areas.

In this way, errors in high-importance regions (e.g., eyes, mouth) contribute more strongly to the total loss, while errors in less relevant areas are still considered but down-weighted.

3.4. Other Loss Functions

To train our super-resolution network, we employ a composite loss function that balances pixel accuracy, percep-

tual similarity, semantic awareness, and visual realism. The overall loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pix}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{att}}\mathcal{L}_{\text{heat}} + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}}, \quad (2)$$

where each component is scaled by a tunable weight λ.

Pixel Loss We use the standard mean squared error (MSE) to enforce low-level fidelity between the predicted and ground-truth images:

$$\mathcal{L}_{\text{pix}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{I}_i - I_i\right\|^2, \quad (3)$$

where $\hat{I}_i$ and I_i denote the predicted and ground-truth pixel intensities, and N is the number of pixels.

Perceptual Loss To capture high-level semantic structure, we adopt a perceptual loss computed on feature activations of a pretrained VGG16 network, following [11]. Specifically, we compare features from conv1_2, conv2_2, and conv3_3 using MSE:

$$\mathcal{L}_{\text{perc}} = \sum_{l=1}^{3}\left\|\phi_l(\hat{I}) - \phi_l(I)\right\|^2, \quad (4)$$

where $\phi_l(\cdot)$ denotes the feature extractor at layer l. Inputs are normalized to ImageNet statistics after mapping from $[-1, 1]$ to $[0, 1]$.

LPIPS Loss. Following Kim *et al.* [12] We also incorporate the Learned Perceptual Image Patch Similarity (LPIPS) [29], which compares deep features extracted from VGG networks, fine-tuned to match human perceptual judgments:

$$\mathcal{L}_{\text{lpips}} = \text{LPIPS}(\hat{I}, I). \quad (5)$$

By encouraging similarity in deep feature space, LPIPS enhances the perceptual realism of the reconstructed images, promoting sharper textures and more human-aligned reconstructions. This aligns with the growing trend in super-resolution literature that shifts focus from signal fidelity to perceptual quality [24].

3.5. Metrics

Following Kim *et al.* [12], we evaluate the quality of the reconstructed images using Peak signal-to-noise ratio (PSNR), Structural Similarity Index (SSIM) [26], and Multi-scale Structural Similarity (MS-SSIM). **PSNR** is a pixel-wise fidelity measure, while **SSIM** accounts for local structural consistency by considering luminance, contrast and texture. **MS-SSIM** [26] extends SSIM to multiple resolutions, providing a more robust assessment of visual quality in line with human perception. This combination of metrics ensures that both low-level fidelity and perceptual realism are properly captured, allowing direct comparison with prior work. Beyond numerical evaluation, we also performed qualitative inspection of validation reconstructions throughout training, and later of test outputs, to verify whether the generated images preserved perceptual realism and semantic consistency.

4. Experiments

4.1. Training Details

Our model is trained on the aligned CelebA dataset using low-resolution inputs of 16×16 and high-resolution ground truths of 128×128. We use a batch size of 64 and train for up to 200 epochs using a single Tesla T4 GPU with mixed-precision training enabled. Reproducibility is ensured by fixing seeds and enforcing deterministic behavior [8]. One single epoch (training and validation together) required about 7 minutes using 30k images. The overall training time for each experiment can be inferred from Table 1 based on the epoch reached. In addition to metric logging, validation samples were inspected every 5 epochs to ensure that training improvements translated into perceptually better reconstructions. The complete training pipeline has been uploaded to GitHub.[1]

Optimizer and Learning Rate We use Adam optimizer [13] with a learning rate of 10^{-4}, no weight decay, and default momentum parameters (β_1=0.9, β_2=0.999). To handle stagnation, we employ `ReduceLROnPlateau` as our learning rate scheduler, which has shown superior performance over fixed schedules in super-resolution and generative tasks [15, 24]. Early stopping is based on a custom validation criterion combining perceptual quality (LPIPS [30]), structural integrity (SSIM), and signal fidelity (PSNR). SSIM and PSNR are truncated once they exceed predefined thresholds, after which the criterion becomes increasingly sensitive to LPIPS. This reflects prior findings that fidelity metrics saturate despite gains in visual realism [17, 24], making LPIPS a more reliable signal for continued perceptual improvement.

Dynamic Loss Weighting Inspired by recent work on curriculum learning and progressive loss scheduling [24, 31], we implement a time-dependent weighting strategy for our multi-term loss. At each epoch e, the total loss is given by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pix}} + \lambda^{(e)}_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda^{(e)}_{\text{att}}\mathcal{L}_{\text{att}} + \lambda^{(e)}_{\text{lpips}}\mathcal{L}_{\text{lpips}}, \quad (6)$$

where $\lambda^{(e)}$ values evolve over time based on a predefined schedule. The training schedule is divided into three phases.

[1] https://github.com/Carraro-Riccardo/Light-Weight-Super-Resolution-UNet-with-YOLO-landmark-heatmaps

Method	PSNR ↑			SSIM ↑			MS-SSIM ↑			Best Epoch ↓
	Train	Val	Test	Train	Val	Test	Train	Val	Test	
Kim *et al.** [12]	–	–	22.66	–	–	0.6850	–	–	0.9020	—
Model with Heatmap Loss	25.55	24.78	**24.84**	0.7170	0.6881	**0.6910**	0.9264	0.9132	0.9142	**58**
Model without Heatmap Loss	25.36	24.21	24.31	0.7135	0.6676	0.6745	0.9234	0.9032	0.9048	114
Model with Multiscale Loss	25.59	24.78	24.83	0.7180	0.6883	**0.6910**	0.9271	0.9138	**0.9147**	**58**
Model with Deep Supervision	25.59	24.77	24.72	0.7171	0.6871	0.6856	0.9266	0.9132	0.9127	68

Table 1. Results on CelebA. *trained on the whole dataset. Bold indicates best results per block.

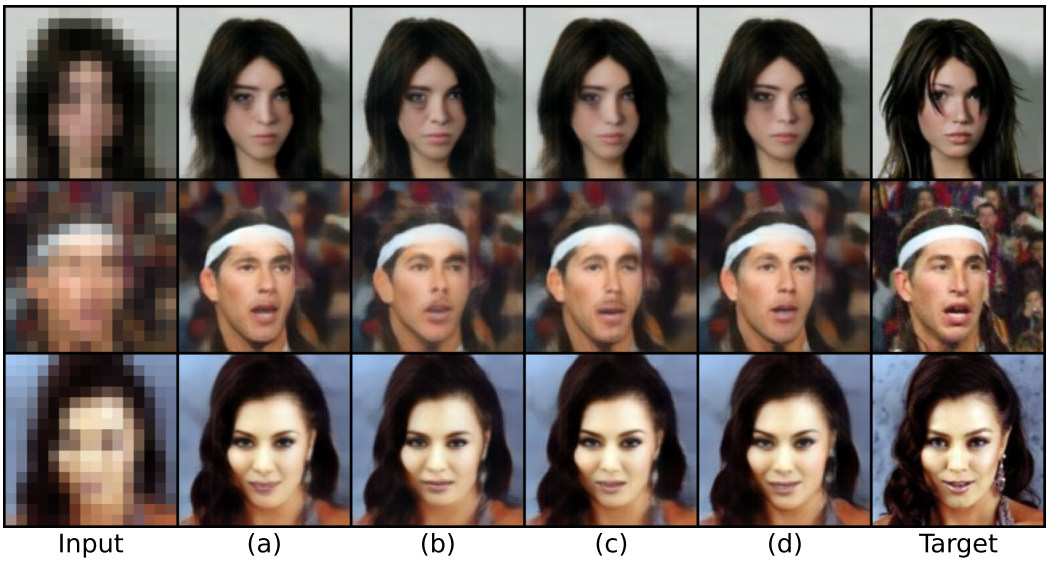

Figure 3. Visual results on representative CelebA test examples. From left to right: Input (upsampled for visualizzation purposes), (a) Model with Heatmap Loss, (b) Model without Heatmap Loss, (c) Model with Multiscale Loss, (d) Model with Deep Supervision and Target image

In the warm-up phase (0–30%), perceptual and attention weights increase linearly while LPIPS remains inactive. During the mid phase (30–85%), all weights are gradually ramped toward their target values. Finally, in the late phase (85–100%), LPIPS is slightly boosted, attention is reduced, and the perceptual loss is held constant to emphasize fine-grained perceptual details [15, 24]. This adaptive scheduling aims to improve training stability and allows the model to gradually shift focus from low-level accuracy to high-level fidelity.

4.2. Effect of Heatmap Loss

We begin by evaluating our core contribution, the integration of YOLO-World-derived landmark heatmaps into a U-Net framework for face super-resolution. Quantitative results are reported in Table 1. The baseline model, trained without heatmap weighting, achieves 24.31 dB PSNR, 0.6745 SSIM and 0.9048 MS-SSIM. Incorporating landmark-guided loss yields consistent gains, improving performance to 24.84 dB PSNR, 0.6910 SSIM and 0.9142 MS-SSIM. Beyond these margins, the model using the Heatmap Loss converges in nearly half the epochs required by the baseline, suggesting a more stable optimization trajectory. We attribute these improvements to the ability of landmark-driven weighting to emphasize semantically critical regions such as eyes and mouth, which leads the model to recover sharper facial details. This effect is visible in Fig. 3 and further supported by qualitative comparisons in the Appendix A, where landmark emphasis enhances visual fidelity in regions most salient to human observers. While distilled FAN-based priors have been shown to mitigate similar issues [12], our results indicate that YOLO-World detections provide an effective alternative, focusing supervision only on visible structures and thereby reducing the risk of noisy or inconsistent guidance.

4.3. Extensions

To assess whether additional supervisory signals can improve our heatmap-guided model, we investigate two common strategies from the image restoration literature:

multiscale loss and deep supervision.

Multiscale Loss In the multiscale setting, reconstruction objectives are enforced not only at the final 128×128 resolution but also at downsampled targets of 64×64 and 32×32. Such supervision has been shown to stabilize training and capture frequency-dependent details [12, 14]. Ground-truth images and YOLO-World heatmaps are downsampled via area interpolation with subsequent renormalization. For each scale s, we compute a composite loss $\mathcal{L}_s$ (pixel, heatmap-weighted, perceptual, LPIPS when applicable) and aggregate across scales as

$$\mathcal{L}_{\text{MS}} = \sum_{s\in\{1,\frac{1}{2},\frac{1}{4}\}} \alpha_s \cdot \mathcal{L}_s, \tag{7}$$

with weights $\alpha_{1.0} = 1.0$, $\alpha_{1/2} = 0.2$, $\alpha_{1/4} = 0.05$, favoring full-resolution fidelity while providing auxiliary guidance. Perceptual terms (VGG, LPIPS) are applied only for scales $\geq 96\times96$, where structural detail is preserved [15, 30]. As reported in Table 1, this variant performs within 0.05 dB PSNR and 0.002 SSIM of the heatmap-guided baseline, indicating that while multiscale losses may regularize optimization, they bring negligible gains in final reconstruction fidelity.

Deep Supervision Deep supervision introduces auxiliary prediction heads at intermediate decoder stages, encouraging intermediate feature maps to align with downsampled ground truth [16, 27, 14]. Our network outputs $\hat{I}^{(1)} \in \mathbb{R}^{128\times128}$ along with auxiliary predictions $\hat{I}^{(1/2)} \in \mathbb{R}^{64\times64}$ and $\hat{I}^{(1/4)} \in \mathbb{R}^{32\times32}$. The same composite formulation in Eq. 7 supervises each resolution. This provides stronger gradient flow without affecting inference-time complexity. Quantitatively, performance remains essentially unchanged relative to the base Heatmap-Guided model. Qualitatively, reconstructed images do not reveal perceptible improvements, and in some cases appear less faithful to the reference (Fig. 3). Nevertheless, convergence dynamics differ: as shown in Appendix B, the model exhibits smoother, albeit slower, training curves compared to the baseline, suggesting improved stability of gradient propagation.

Results Table 1 summarizes the outcomes of the two extensions. Both multiscale supervision and deep supervision applied to the Heatmap-Guided model yield accuracy levels nearly indistinguishable from that of the Heatmap-Guided base model, with differences within typical experimental variability in PSNR and SSIM. These findings indicate that, in our lightweight U-Net architecture, auxiliary losses primarily serve to stabilize training but do not translate into measurable improvements in reconstruction fidelity. While it is conceivable that more substantial gains could be obtained in higher-capacity networks, such exploration lies beyond the scope of the present work. Our main conclusion is that YOLO-World-guided heatmaps remain the dominant factor for performance, while additional supervisory mechanisms play at most a secondary role in shaping training dynamics and final accuracy.

5. Conclusion

This work addresses extreme face super-resolution ($8\times, 16\times16 \rightarrow 128\times128$) using a lightweight U-Net guided by facial landmark heatmaps generated via YOLO-World. These heatmaps serve as spatial weights in the reconstruction loss, emphasizing errors in semantically important regions such as eyes, nose and mouth.

A significant contribution is the adoption of YOLO-World, an open-vocabulary object detector, to extract landmarks directly from ground-truth images without requiring the training or fine-tuning of auxiliary alignment networks like FAN. This approach reduces computational overhead by avoiding inference-time heatmap generation on model predictions, a requirement in previous works such as Kim *et al.* [12]. The method thus offers a practical, flexible, and resource-efficient alternative for inducing spatially-aware supervision.

Quantitative results show consistent improvements in PSNR, SSIM, and MS-SSIM metrics, alongside faster convergence. Qualitative evaluations support these gains, revealing sharper and more realistic reconstructions in critical facial regions. Further studies indicate that, within the capacity constraints of the lightweight U-Net, additional supervisory schemes like multiscale losses and deep supervision mainly stabilize training without substantially enhancing final accuracy.

Nonetheless, several avenues remain for future exploration. Validation on less controlled datasets such as unaligned CelebA or "faces in the wild" is needed to assess robustness to pose, illumination, and occlusion variations. Optimizing YOLO-World's landmark detection via prompt engineering or fine-tuning could further improve supervision quality. Finally, the generality of leveraging open-class, detector-driven priors invites application to other domains requiring region-focused super-resolution or detail enhancement.

In summary, by leveraging YOLO-World-derived heatmaps for spatially-weighted supervision, the proposed approach offers a lean, adaptable framework that improves super-resolution quality under extreme degradation, without incurring the costs or complexities of auxiliary network training or adversarial methods.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2018.

[2] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, 8(6):679–698, 1986.

[3] Yu Chen, Yu-Kun Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2492–2501, 2018.

[4] Yu Cheng, Fangyun Wei, Xiangyu Zhang, Jingdong Wang, Wei Yang, Yu Qiao, and Dahua Lin. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. In *IEEE transactions on pattern analysis and machine intelligence*, volume 38, pages 295–307. IEEE, 2015.

[6] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. *European Conference on Computer Vision (ECCV)*, pages 391–407, 2016.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[8] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.

[10] Glenn Jocher, Ayush Chaurasia, Laughing, Jiacong Fang, Jirka Borovec, Alex Wong, et al. Ultralytics YOLOv8: Cutting-edge object detection models. https://github.com/ultralytics/ultralytics, 2023. Accessed: 2025-08-20.

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision (ECCV)*, pages 694–711. Springer, 2016.

[12] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7177–7186, 2019.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[14] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017.

[15] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[16] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *AISTATS*, 2015.

[17] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[19] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. *IEEE signal processing magazine*, 20(3):21–36, 2003.

[20] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016.

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

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

[23] Hanno Scharr. Optimal operators in digital image processing. *Dissertation, University of Heidelberg*, 2000.

[24] 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 *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[25] Ying Wang, Xintao Wang, Chao Dong, and Xiaoou Tang. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 23:3104–3121, 2021.

[26] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. *Signals, Systems and Computers*, 2:1398–1402, 2003.

[27] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015.

[28] Wenming Yang, Xuechen Zhang, Yanan Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single

image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12):3104–3121, 2019.

[29] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.

[30] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[31] Ying Zhang, Xuanqin Zhang, Chenglong Li, Yongtao Wang, Bineng Zhong, and Guohui Sun. Designing a practical perceptual loss: Learning discriminative feature representations for image quality assessment. *IEEE Transactions on Image Processing*, 30:7049–7061, 2021.

A. Further test visualizations

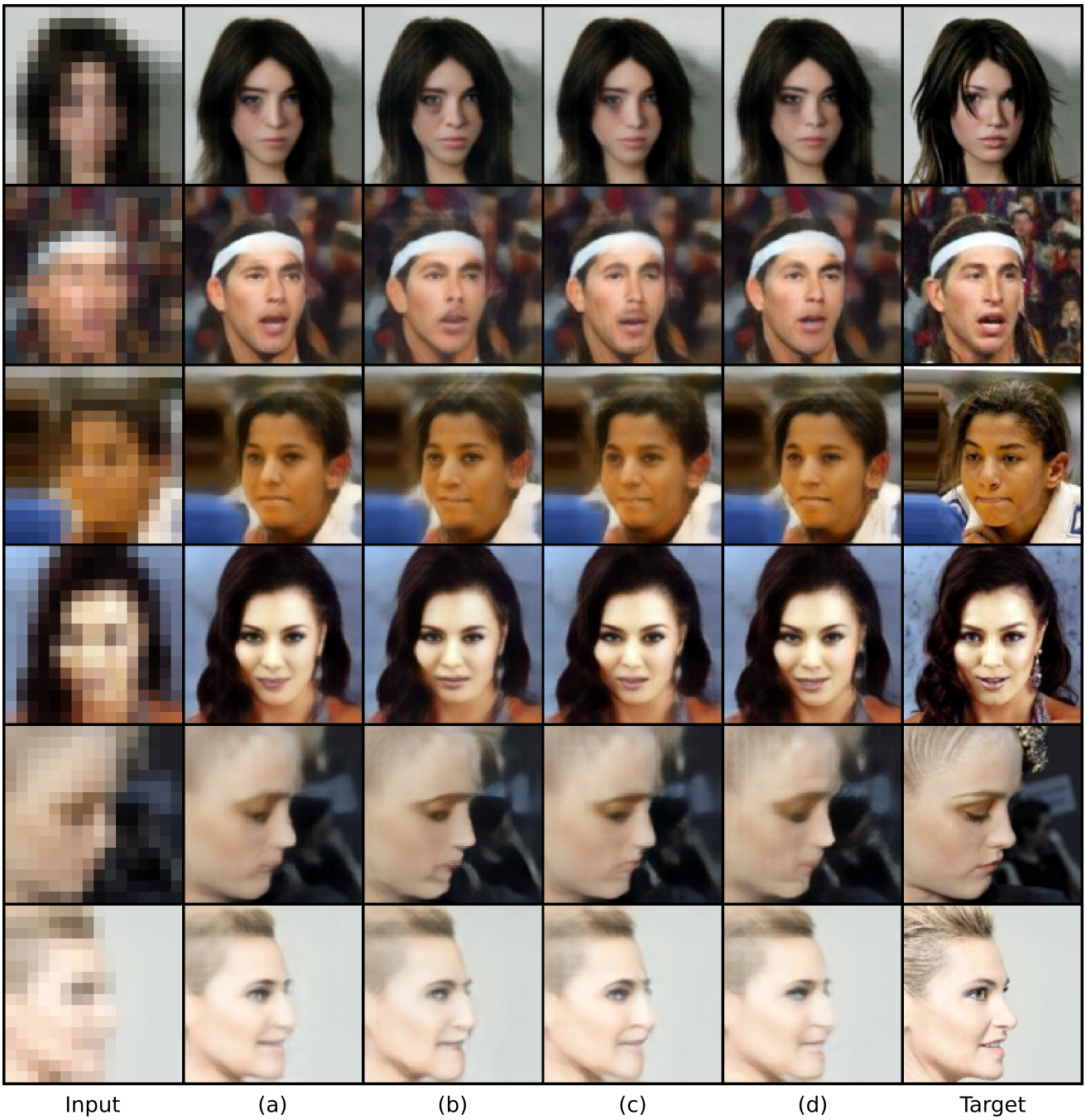

Figure 4. Visual results on representative CelebA test examples. From left to right: Input (upsampled for visualizzation purposes), (a) Model with Heatmap Loss, (b) Model without Heatmap Loss, (c) Model with Multiscale Loss, (d) Model with Deep Supervision and Target image

B. Validation metrics

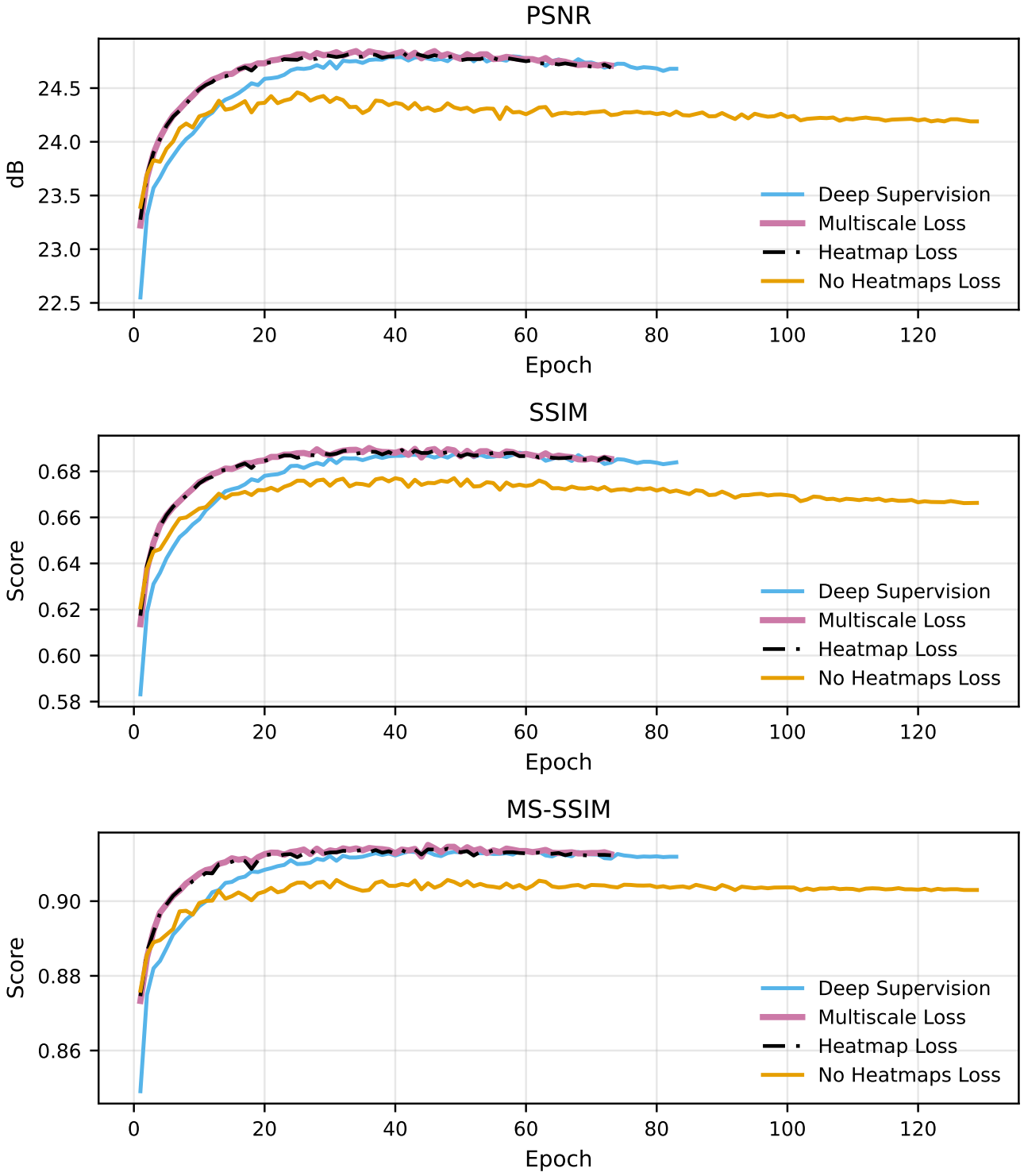

Figure 5. Validation curves of PSNR, SSIM and MS-SSIM on the CelebA test subset. We compare the baseline without heatmap weighting, the heatmap-guided model, and the variants with multiscale loss and deep supervision. Landmark-guided weighting yields higher scores and faster convergence relative to the baseline, while multiscale and deep supervision provide additional stability across epochs.