# MACRO: Training-free Multi-plane Attention for Closeup Render Optimization

**Nitzan Hodos**[1,2], **Roy Amoyal**[1,3], **Lior Fritz**[1], **Ianir Ideses**[1], **Sagie Benaim**[1,4], **Netalee Efrat**[1]
[1]Amazon Prime Video [2]Technion, Israel Institute of Technology
[3]Ben Gurion University [4]Hebrew University of Jerusalem
https://nitzanhod.github.io/MACRO/

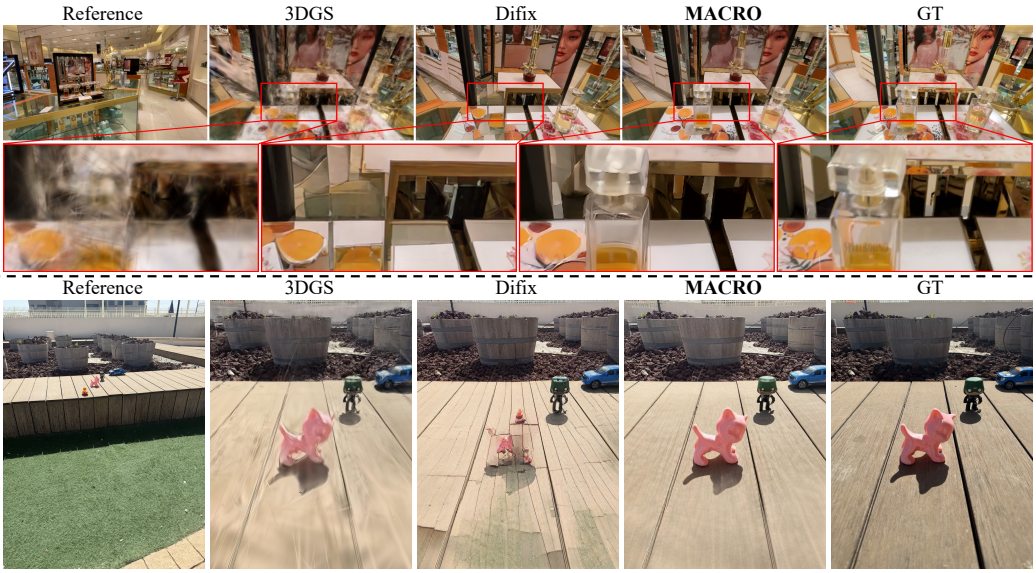


Figure 1: **Close-up novel view synthesis with MACRO. Top**: a scene from DL3DV-Closeup. Given a wide reference view, 3DGS produces a blurry close-up rendering. Difix generates a sharp image but transfers incorrect textures from the reference due to scale mismatch (see insets). MACRO faithfully restores details that match the ground truth. **Bottom**: a scene from MobileClose-10. Difix transfers the wood plank texture at the scale of the wide shot, producing visibly wrong patterns. MACRO recovers the correct close-up texture without retraining.

## Abstract

Close-up rendering, zooming into a scene well beyond any training camera, is important for virtual production and interactive 3D content, yet remains an open challenge. 3D Gaussian splatting (3DGS) enables high-fidelity, real-time novel view synthesis, but its rendering quality degrades at close range. Recent diffusion-based methods that enhance the rendering by conditioning on reference images from the training set produce significant artifacts in this setting. We analyze this failure and identify its root cause: the scale gap between the close-up and reference views. We show that the features in reference-conditioned enhancement models are not scale-invariant, causing cross-view attention to retrieve incorrect correspondences when the same content appears at different scales, and that this mismatch cannot be corrected in latent space because the VAE encoder is not scale-equivariant. Building on this analysis we introduce MACRO, Multi-plane Attention for Closeup Render Optimization, a training-free method for high-quality close-up novel view synthesis from 3DGS. MACRO resolves the scale gap by

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

leveraging the scene's known 3D structure: it decomposes the close-up into depth planes, crops and resizes references in image space to match the scale of each plane before encoding, and applies a depth-aware attention mask so each token attends only to scale-matched references. The method requires no architectural changes or additional training. We further contribute two new close-up novel view synthesis benchmarks, the first standardized evaluation protocol for this setting, and demonstrate state-of-the-art results on both, outperforming existing 3DGS and diffusion-based methods on both reconstruction and perceptual metrics.

# 1 Introduction

3D Gaussian splatting (3DGS) [8] has demonstrated impressive fidelity and speed for novel view synthesis. However, rendering quality deteriorates significantly when the synthesized viewpoint deviates from the training views. A particularly important case is close-up view synthesis: rendering zoomed-in views that are substantially closer to the scene than any training camera, often from different viewpoints. Close-up rendering is essential for applications such as virtual production, interactive 3D content, and immersive media, where users expect to freely explore scenes at any distance. This setting is fundamentally different from super-resolution or digital zoom: the close-up view may observe content that is only partially visible across multiple training images, with no single reference providing complete coverage. Moreover, the close-up viewpoint is out-of-distribution, producing severe artifacts in regions with insufficient training signal, while fine details that were unresolved or occluded at the original capture distance cannot be recovered from the 3DGS representation alone.

Since these missing details cannot be recovered from the 3D representation, recent works turn to diffusion-based enhancement methods [19, 4, 2, 26] that condition on real reference images from the training set. By leveraging the strong image priors of diffusion models together with actual scene observations, these methods can transfer genuine textures and details from the references to the degraded rendering, producing outputs faithful to the true scene. However, in this close-up setting these methods produce poor results. They assume that the target and reference views observe the scene at a similar scale, whereas in close-ups objects appear at completely different scales that vary spatially across the image, as a function of depth. Existing approaches that do target close-up synthesis [21, 31] attempt to bridge the gap gradually, generating intermediate views at progressively closer distances. While this progressive strategy reduces the distribution gap at each step, it does not tackle the actual root cause: the objects in the close-up and reference are at different scales. Progressive schemes are also computationally expensive, requiring multiple rounds of rendering, enhancement, and retraining of the 3D representation, with each step introducing accumulated errors.

In this work, we identify two key insights that explain and resolve this failure. First, the VAE and U-Net features used in diffusion-based enhancement are sensitive to scale: when the same scene content is encoded at different scales, the resulting latent representations differ, causing attention between the target and reference tokens to fail. Second, this scale mismatch cannot be corrected in latent space because the VAE encoder is not scale-equivariant [9]: $\mathcal{E}(\texttt{Resize}(\mathbf{x})) \neq \texttt{Resize}(\mathcal{E}(\mathbf{x}))$. The scale alignment must therefore be performed in image space, before encoding. We discuss these properties thoroughly in Section 4.1.

Building on these insights, we introduce **MACRO** (**M**ulti-plane **A**ttention for **C**loseup **R**ender **O**ptimization), a training-free method that resolves the scale mismatch. Our key observation is that diffusion-based enhancement models are 2D networks trained with 2D image priors, yet reference matching in close-up rendering is fundamentally a problem of geometric scale. While these 2D networks learn implicit 3D representations, their features are not scale-invariant, meaning they cannot reliably establish cross-view correspondences when the same content appears at drastically different zoom levels. Rather than forcing the network to overcome this mismatch purely in latent space, we propose MACRO, a training free light weight alternative. MACRO incorporates explicit 3D knowledge by leveraging the known 3D structure of the scene to decompose the rendered close-up into depth planes. For each depth plane, we compute scale-matched crops of the reference images in image space before VAE encoding, and apply a depth-aware attention mask so that each close-up token attends only to reference tokens at the matching scale. This naturally supports multiple reference views, with each depth plane attending to the most informative reference. MACRO requires

no architectural changes or training and operates as a preprocessing and attention masking strategy on top of existing enhancement models.

We demonstrate MACRO on top of the Difix [19] architecture and evaluate it on two close-up benchmarks we construct for this task. As shown in Figure 1, MACRO produces faithful close-up renderings where baselines fail. Quantitatively, MACRO improves perceptual metrics by up to $38\%$ over the best enhancement baseline and by up to $67\%$ over unenhanced 3DGS, while maintaining competitive reconstruction scores. To summarize, we make the following contributions:

- We analyze the sensitivity of diffusion-based enhancement to scale, revealing two key properties: (a) VAE and U-Net features are not scale-invariant, causing cross-view attention to fail under scale mismatch, and (b) the VAE encoder is not scale-equivariant, preventing correction in latent space and necessitating image-space solutions.
- We propose MACRO, a training-free, depth-based multi-plane attention mechanism that resolves scale mismatch by cropping and resizing reference images in image space before encoding, allowing correct cross-view attention.
- We introduce two close-up novel view synthesis benchmarks, **DL3DV-Closeup**, curated from DL3DV-10K [12], and **MobileClose-10**, captured specifically for this task, providing the first standardized evaluation protocol for close-up rendering under significant scale changes.
- We achieve state-of-the-art close-up rendering quality across both benchmarks, outperforming both 3DGS baselines and other diffusion-based approaches on reconstruction and perceptual metrics, while requiring no additional training.

## 2 Related work

**High-fidelity novel view synthesis.** 3D Gaussian Splatting (3DGS) [8] enables real-time novel view synthesis, but rendering quality degrades for viewpoints far from the training cameras. Several works improve rendering fidelity without generative priors: Mip-Splatting [28] and multi-scale Gaussians [24] address aliasing across zoom levels, FreGS [29] improves high-frequency detail through frequency regularization, and Scaffold-GS [14] and Analytic Splatting [11] improve rendering precision. However, these methods cannot recover detail never observed in the training views, a fundamental limitation for close-up novel views captured far from any training camera.

**Close-up novel view synthesis.** Recent work directly targets the close-up rendering problem. [22] generates pseudo-label supervision for close-up viewpoints and retrains the 3DGS representation. Close-up-GS [21] progressively self-trains to refine Gaussians for close-up views. CloseUpShot [31] trains a point-conditioned diffusion model for close-up synthesis from sparse inputs. These methods all require retraining the reconstruction or a new generative model, which is both resource demanding and slow, while side-stepping the fundamental issues MACRO targets.

**Reference-guided rendering enhancement.** Recent diffusion-based methods condition on reference images from the training set through attention layers to improve novel view quality. One line of work enhances 3DGS renderings by transferring textures from clean references [20, 19, 4, 2, 26]. A second group takes a purely generative approach, synthesizing novel views without relying on a 3DGS reconstruction [32, 27, 6]. For both groups, close-up views present an out-of-distribution case: the attention between the target view and references operates under a scale mismatch, and purely generative approaches must additionally generalize to camera poses unseen during training. MACRO addresses this scale mismatch directly.

**Scale equivariance in neural networks.** Most vision model architectures are not scale-equivariant [17, 9]. Existing remedies include scale-equivariant architectures [18, 17, 16] and equivariance regularization [9], but these have not been demonstrated at the scale of the pretrained VAE/U-Net stacks used by reference-conditioned enhancement, and applying them would require retraining the enhancement backbone. MACRO instead resolves the scale mismatch in image space before encoding, requiring no retraining.

**Super-resolution for 3D Gaussian Splatting.** A related line of work renders 3DGS at resolutions higher than the training images [3, 23, 1], using sub-pixel constraints, latent feature fields, or 2D SR priors for multi-view consistent upsampling. These methods target uniform resolution upscaling

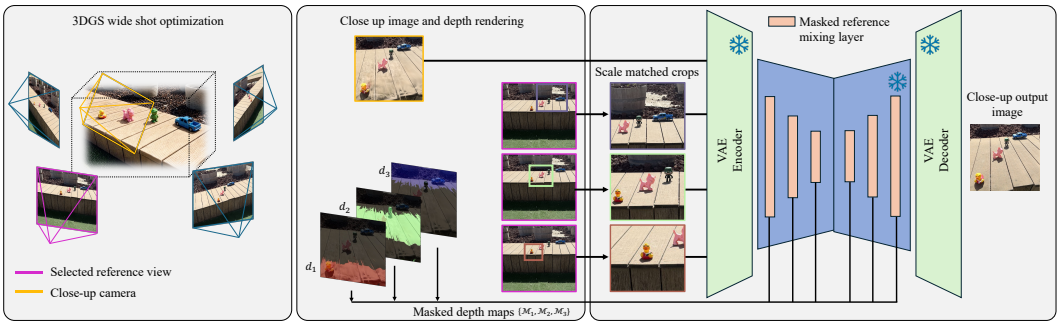


Figure 2: **MACRO pipeline.** Given a close-up rendering (**orange**) and its depth map from the 3DGS reconstruction, we decompose the scene into $P$ depth planes with centroid depths $d_1, d_2, \ldots, d_P$, yielding corresponding depth masks $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_P$. For each depth plane, we compute a scale-matched crop of the reference image (**pink**) via depth-guided cropping and resizing, producing $P$ reference crops per reference view (color coded according to the depth planes). The close-up input and scale-matched reference crops are encoded by the VAE and processed through a masked reference mixing layer, where each close-up token attends only to reference tokens from the corresponding depth plane. The VAE decoder produces the enhanced close-up output. The figure illustrates the single-reference case; when using $K$ references, this yields $K \times P$ crops (Section 4.3).

from the same viewpoints. In contrast, close-up novel views create spatially-varying scale changes (foreground magnifies more than background) and reveal previously occluded content, motivating our multi-plane decomposition.

## 3 Preliminaries: Reference-guided Enhancement

Several recent methods [19, 4, 2, 26] enhance novel view renderings from 3DGS by conditioning a diffusion model on real reference images from the training set. In all of these methods, the rendered target view and reference image are encoded into a shared latent space, and an attention mechanism fuses information between the two. Since we build our method on top of Difix [19], we describe its specific formulation below and present empirical results on its SD-Turbo U-Net backbone.

Given a target view $\tilde{\mathbf{I}}_{\text{target}}$ rendered from the 3DGS reconstruction and a clean reference image $\mathbf{I}_{\text{ref}}$ captured by a training camera, both are encoded into latent representations by a VAE encoder $\mathcal{E}$:

$$\mathbf{z}_{\text{target}} = \mathcal{E}(\tilde{\mathbf{I}}_{\text{target}}), \quad \mathbf{z}_{\text{ref}} = \mathcal{E}(\mathbf{I}_{\text{ref}}). \tag{1}$$

The latent tokens are processed by the U-Net with *reference mixing layers*. In each such layer, the target and reference tokens are concatenated along the token dimension and joint self-attention is applied, allowing every target token to attend to every reference token:

$$[\mathbf{z}'_{\text{target}};\, \mathbf{z}'_{\text{ref}}] = \texttt{SelfAttn}([\mathbf{z}_{\text{target}};\, \mathbf{z}_{\text{ref}}]), \tag{2}$$

where $[\cdot\,;\,\cdot]$ denotes concatenation along the token dimension. This enables the transfer of textures and details from the clean reference to the degraded rendering. Finally, the final predicted noise is used to analytically compute the cleaned latent $\mathbf{z}'_{\text{cleaned}}$, which is passed through the VAE's decoder to produce the final enhanced output $\hat{\mathbf{I}}_{\text{target}} = \mathcal{D}(\mathbf{z}'_{\text{cleaned}})$.

## 4 Method

Given a 3DGS reconstruction of a scene, our goal is to enhance a close-up view $\tilde{\mathbf{I}}_{\text{cu}}$ rendered from a close-up camera $\pi_{\text{cu}}$, using $N$ reference images $\{\mathbf{I}^n_{\text{ref}}\}_{n=1}^N$ captured by training cameras $\{\pi^n_{\text{ref}}\}_{n=1}^N$ and the depth map $\mathbf{D}_{\text{cu}}$ rendered from the 3DGS reconstruction. Figure 2 provides an overview of the MACRO pipeline. In Section 4.1, we analyze the scale mismatch problem that motivates our approach. We then present the three components of MACRO: multi-plane depth decomposition (Section 4.2), scale-matched reference cropping (Section 4.3), and depth-masked attention (Section 4.4).

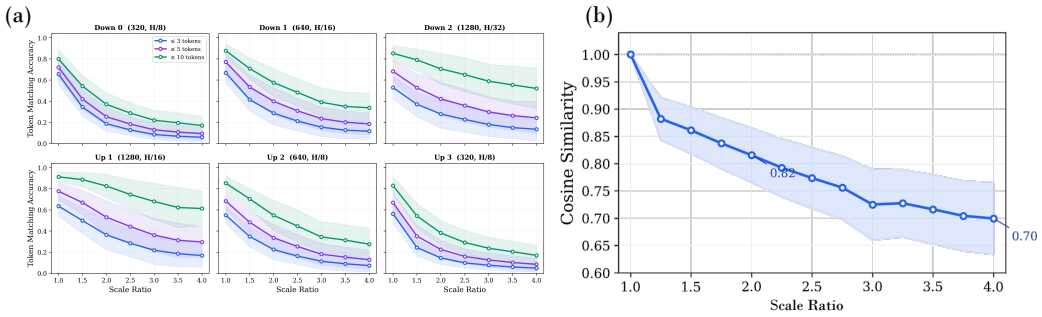


Figure 3: **Scale sensitivity analysis. (a)** Fraction of U-Net attention tokens whose nearest-neighbor key retrieval falls within a radius of 3, 5, or 10 tokens of the ground-truth correspondence, across six attention layers (Down 0–2, Mid, Up 1–2). Matching accuracy degrades at all layers as scale increases, confirming that cross-view attention fails under scale mismatch. Note that at coarser layers (Down2, Up1) the token matching accuracy is higher as there are less tokens overall. **(b)** Cosine similarity between the VAE latent of a resized image and the resized VAE latent of the original, i.e., $\mathcal{E}(\texttt{Resize}_s(\mathbf{x}))$ vs. $\texttt{Resize}_s(\mathcal{E}(\mathbf{x}))$. Similarity drops from 1.0 to 0.70 at 4× scale, showing that the VAE is not scale-equivariant and scale correction in latent space is less effective.

### 4.1 Motivation: Scale mismatch in close-up rendering

We first analyze why existing reference-guided enhancement fails in the close-up setting. We identify two properties: (1) the learned features are sensitive to scale, causing attention to retrieve incorrect correspondences, and (2) this mismatch cannot be corrected in latent space due to the non-equivariance of the VAE encoder.

The reference mixing mechanism of Eq. (2) implicitly assumes that corresponding content in the target and reference views appears at a similar spatial scale: tokens representing the same scene content should produce similar features so that attention retrieves the correct match. This assumption holds when both views have comparable fields of view. However, when the camera moves closer to the scene, objects appear at a much larger scale than in the reference images, and the same content is encoded into different feature representations, breaking the attention matching.

**Attention fails under scale mismatch.** To quantify this effect, we simulate the close-up/reference scenario by cropping a sub-region from a native 4K image and resizing it to the processing resolution (emulating a close-up), while separately resizing the full image to the same resolution (emulating a reference). We encode both views with the VAE, run each through the U-Net, and at each self-attention layer compute the query projection $\mathbf{Q} = W_q(\mathbf{h})$ from the close-up and the key projection $\mathbf{K} = W_k(\mathbf{h})$ from the reference. For each close-up query token, we retrieve its nearest neighbor in the reference key space and measure the spatial displacement from the ground-truth correspondence. Figure 3(a) reports the fraction of tokens whose match falls within 3, 5, or 10 tokens of the correct position across six U-Net attention layers. Matching accuracy degrades monotonically with increasing scale ratio, confirming that the pretrained attention cannot establish reliable correspondences under scale mismatch. We visualize this in Figure 4, showing how attention to the full reference retrieves incorrect correspondences, while our scale-matched crops restore correct matching.

**Scale correction must happen in image space.** One might attempt to fix this mismatch through geometric constraints on the attention, e.g., by warping or rescaling the latent representations after encoding. However, this approach is incorrect because the VAE encoder $\mathcal{E}$ is not scale-equivariant [9]: encoding a resized image does not produce the same result as resizing the encoded latent:

$$\mathcal{E}(\texttt{Resize}_s(\mathbf{x})) \neq \texttt{Resize}_s(\mathcal{E}(\mathbf{x})), \tag{3}$$

where $\texttt{Resize}_s$ denotes spatial rescaling by factor $s$. We verify this in Figure 3(b), which measures the cosine similarity between $\mathcal{E}(\texttt{Resize}_s(\mathbf{x}))$ and $\texttt{Resize}_s(\mathcal{E}(\mathbf{x}))$ as a function of scale factor $s$ on 200 images from DL3DV-10K [12]. This leads to the core principle underlying MACRO: scale alignment must be performed in image space, before VAE encoding.

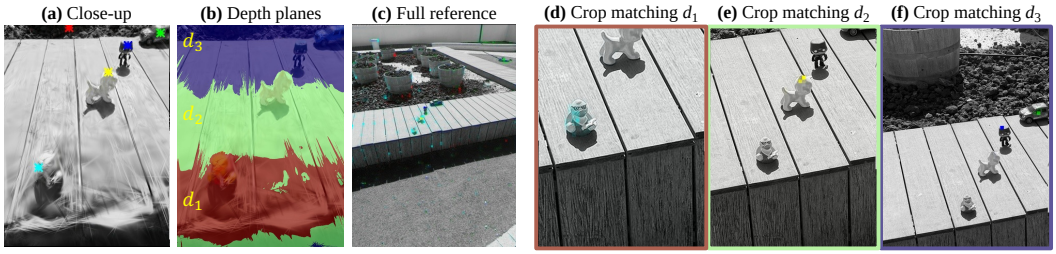


Figure 4: **Attention mask visualization.** **(a)** A rendered close-up with color-coded query tokens. **(b)** Depth plane assignments. **(c)** Standard (unmasked) attention maps from each query token to the full reference image; the duck query token incorrectly attends to the wood deck, explaining the texture artifacts visible in Difix results (Figure 1). **(d–f)** Our masked attention constrains each query token to attend only to tokens on the scale-matched crop of its depth plane. Within each crop, the attention correctly retrieves the corresponding scene content.

### 4.2 Multi-plane depth decomposition

Given the analysis above, we now present the design of MACRO. In a close-up view, objects at different depths appear at different scales relative to the reference images. A foreground object close to the camera undergoes a larger scale change than a background object. To handle this, we decompose the close-up rendering into $P$ depth planes, where each plane groups pixels at similar depths and thus similar scale factors. Given the rendered depth map $\mathbf{D}_{\text{cu}}$, we partition the pixels into $P$ groups using $k$-means clustering on the depth values:

$$\{d_1, d_2, \ldots, d_P\} = \texttt{k-means}(\mathbf{D}_{\text{cu}}, P), \tag{4}$$

where $d_p$ is the centroid depth of plane $p$. Each pixel $i$ in the close-up view is assigned to its nearest depth plane, yielding a partition $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_P\}$ of the close-up image, where $\mathcal{M}_p$ is the set of pixel indices belonging to plane $p$. Clustering is performed in the reciprocal $1/z$-component space.

Given that $N$ views were used to train the 3DGS to reconstruct the full scene, most of them are not useful when enhancing a very specific close-up. We therefore apply per close-up a greedy choice of $\{\mathbf{I}^k_{\text{ref}}\}_{k=1}^K$ references that cover the input angle best where $K \leq N$, and also eliminate references who do not see the close-up due to occlusion. Then only these $K$ references are used for the enhancement pipeline. We found that setting $K = 3$ references and $P = 3$ depth planes to be optimal and use these values throughout our experiments. More details about the hyperparameter search are in Section 5.3.

### 4.3 Scale-matched reference cropping

For each depth plane $p$ with centroid depth $d_p$, we compute a scale-matched crop of each reference image that shows the corresponding scene content at the same scale as it appears in the close-up view. Given $K$ reference views and $P$ depth planes, this produces a total of $K \times P$ reference crops.

**Scale factor.** For depth plane $p$, the scale factor between the close-up and reference view is determined by the ratio of the object's depth in the reference to its depth in the close-up:

$$s_p = \frac{d_p^{\text{ref}}}{d_p^{\text{cu}}}, \tag{5}$$

where $d_p^{\text{cu}}$ and $d_p^{\text{ref}}$ are the depths of the plane-$p$ content as seen from the close-up and reference cameras, respectively. Objects closer to the close-up camera (small $d_p^{\text{cu}}$) yield a larger scale factor, meaning they appear proportionally larger in the close-up than in the reference. The crop size in the reference image is set to cover a region of $1/s_p$ of the reference field of view, so that after resizing to the processing resolution, the content appears at the same scale as in the close-up.

**Crop location and upsampling.** We project the 3D points at depth $d_p$ visible in the close-up into the reference camera and center the crop on the projected region. The crop is upsampled to match the input resolution (1k in our experiments) using a state of the art super-resolution model PFT-SR

[13] while maintaining aspect-ratio. The resulting scale-matched reference patch $\mathbf{I}_{\text{ref}}^{k,p}$ is encoded as $\mathbf{z}_{\text{ref}}^{k,p} = \mathcal{E}(\mathbf{I}_{\text{ref}}^{k,p})$. Because cropping and resizing occur in image space before encoding, the attention mechanism can establish correct correspondences (Section 4.1).

### 4.4 Depth-masked Multi-plane Attention

The close-up render and all $K \times P$ scale-matched reference crops are encoded through the VAE and processed jointly by the U-Net mixing blocks (Section 3). To ensure correct scale texture copying, for each depth plane $p$ we construct an attention mask $\mathbf{A} \in \{0,1\}^{S_l \times KS_l}$ that controls which close-up tokens can attend to the reference crops of that plane's scale, where $S_l = H_l \times W_l$ is the number of spatial latent tokens of a single view at mixing block $l$. Each close-up token from depth plane $p$ attends only to reference tokens from the corresponding scale-matched patch for plane $p$. Let $\mathcal{M}_l^p$ denote the set of close-up token indices belonging to plane $p$, and let $\mathcal{R}_l^{k,p}$ denote the reference token indices corresponding to the scale-matched crop for plane $p$ from reference $k$, both downsampled to the resolution of $H_l, W_l$. Then:

$$\mathbf{A}[i, (k-1)S_l + j] = 1 \iff i \in \mathcal{M}_l^p \text{ and } j \in \mathcal{R}_l^{k,p}. \tag{6}$$

We then construct the full multi-view attention mask $\mathbf{A}' \in \{0,1\}^{(KP+1)S_l \times (KP+1)S_l}$, where per-view self attention is fully allowed for both input and reference views, cross-reference attention is blocked, and input-reference cross-view attention is determined by $\mathbf{A}$. The modified reference mixing layer (Eq. (2)) then becomes:

$$\mathbf{z}_l' = \texttt{MaskedSelfAttn}(\mathbf{z}_l, \mathbf{A}'), \tag{7}$$

where $\mathbf{z}_l = [\mathbf{z}_{\text{cu}}; \mathbf{z}_{\text{ref}}^{1,p}; \ldots; \mathbf{z}_{\text{ref}}^{K,P}]$ is the concatenation of the close-up and all reference latents at U-Net mixing block $l$, and the mask $\mathbf{A}$ is applied to the attention scores before the softmax. Notably, this requires only a single forward pass through the diffusion model. The mask controls the information flow without requiring multiple runs.

## 5 Experiments

We evaluate MACRO on two close-up novel view synthesis benchmarks that we construct as part of this work, addressing the lack of standardized evaluation protocols for this setting. In Section 5.1, we describe the evaluation benchmarks and metrics. In Appendix B, we provide implementation details. In Section 5.2, we compare MACRO against existing methods. In Section 5.3, we ablate key design choices. In Section 5.4, we discuss limitations.

### 5.1 Benchmarks and Metrics

**Benchmarks.** Existing novel view synthesis benchmarks do not systematically evaluate close-up rendering, where the target camera is significantly closer to the scene than any training view. To address this gap, we introduce two new benchmarks that test different aspects of close-up novel view synthesis. **DL3DV-Closeup** is our primary benchmark, comprising 44 diverse in-the-wild scenes curated from DL3DV-10K [12] by selecting scenes with large depth disparity between camera positions, yielding naturally occurring close-up/far view pairs with a total of 314 evaluation pairs. Our newly curated **MobileClose-10** contains 10 scenes captured specifically for the close-up task using a smartphone camera, and serves to evaluate generalization beyond DL3DV with 39 far/close evaluation pairs. Full details on the construction of both benchmarks are provided in Appendix D.

**Metrics.** We report PSNR and SSIM as standard pixel-aligned reconstruction metrics. However, these measures do not fully capture whether the enhanced close-up faithfully reproduces the semantic content and fine-grained structure of the scene: two images can have similar PSNR yet differ substantially in perceptual fidelity, particularly when diffusion-based enhancement introduces plausible but spatially shifted details. We therefore additionally report three perceptual metrics: LPIPS [30], DreamSim [5], and DINOv2 [15] feature distance. These operate in learned feature spaces and better reflect whether the output preserves the identity and structure of the original scene content.

**Baselines.** We compare against the reconstruction methods 3DGS [8] and Mip-Splatting [28], the purely generative NVS method SEVA [32], and the enhancement methods GSFixer [26] and Difix [19].

Table 1: **Quantitative comparison on both benchmarks.** DSim = DreamSim, DINO = DINOv2 feature distance. MACRO achieves the best perceptual metrics on both benchmarks while maintaining competitive reconstruction scores.

| | *DL3DV-Closeup (44 scenes)* | | | | | *MobileClose-10 (10 scenes)* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | PSNR↑ | SSIM↑ | LPIPS↓ | DSim↓ | DINO↓ | PSNR↑ | SSIM↑ | LPIPS↓ | DSim↓ | DINO↓ |
| 3DGS [8] | 16.00 | 0.499 | 0.457 | 0.261 | 0.462 | <u>14.58</u> | **0.441** | 0.578 | 0.293 | 0.512 |
| Mip-Splatting [28] | **16.26** | **0.514** | 0.478 | 0.257 | 0.440 | 14.41 | 0.428 | 0.667 | 0.335 | 0.521 |
| SEVA [32] | 12.35 | 0.335 | 0.544 | 0.156 | 0.263 | 13.91 | 0.401 | 0.647 | 0.270 | 0.417 |
| GSFixer [26] | 15.91 | 0.494 | 0.378 | <u>0.121</u> | 0.250 | 14.51 | 0.401 | 0.614 | <u>0.250</u> | 0.408 |
| Difix [19] | 15.81 | 0.477 | <u>0.343</u> | 0.128 | <u>0.205</u> | 14.55 | 0.407 | <u>0.557</u> | 0.253 | <u>0.352</u> |
| Difix3D [19] | 14.69 | 0.427 | 0.394 | 0.149 | 0.248 | 13.76 | 0.418 | 0.582 | 0.258 | 0.406 |
| Difix3D+ [19] | 14.33 | 0.404 | 0.393 | 0.155 | 0.243 | 13.36 | 0.385 | 0.579 | 0.270 | 0.394 |
| PR-Difix3D | 15.15 | 0.433 | 0.399 | 0.153 | 0.262 | 13.68 | 0.393 | 0.563 | 0.292 | 0.428 |
| PR-Difix3D+ | 14.42 | 0.379 | 0.430 | 0.178 | 0.271 | 13.00 | 0.329 | 0.596 | 0.320 | 0.437 |
| **MACRO (ours)** | <u>16.17</u> | <u>0.505</u> | **0.330** | **0.105** | **0.152** | **14.60** | <u>0.433</u> | **0.551** | **0.168** | **0.217** |




Figure 5: **Qualitative results on DL3DV-Closeup.** Only in MACRO is the text on the "Essentials Kit" readable and the animal figures faithfully preserved (see crops).

We also evaluate progressive variants that attempt to bridge the distribution gap of the close-up setting gradually, as discussed in Section 1: Difix3D progressively renders and enhances intermediate views at incrementally closer distances, distilling results back into the 3DGS representation at each step. PR-Difix3D (Progressive Reference) also updates the reference images at each step to decrease the scale gap between the rendered and reference, implementing the strategy of Close-up-GS [21] which did not release code. The "+" variants (Difix3D+, PR-Difix3D+) apply an additional enhancement step at the final iteration. Our GSFixer setup similarly reflects the interpolated trajectory strategy of CloseUpShot [31] (see Appendix B.2), as its implementation was also unavailable at the time of submission. FlowR [4] and ArtiFixer [2] were not available either.

### 5.2 Results

Table 1 reports quantitative results on both benchmarks and Figures 5 and 6 show qualitative comparisons on DL3DV-Closeup and MobileClose-10 respectively (additional qualitative results on DL3DV-Closeup are provided in Appendix A). Although 3DGS and Mip-Splatting avoid the texture artifacts of enhancement methods, their renderings remain blurry and lack fine detail at close range. SEVA produces sharp outputs but diverges from the ground truth, likely because close-up views are out of distribution relative to its training set and it relies solely on camera parameters without leveraging the scene's 3DGS reconstruction. Difix improves perceptual metrics over 3DGS, confirming that the diffusion model adds detail, but it copies wrong textures from the reference, and due to the scale difference, even when the attention mechanism looks at the correct location, the textures are transferred at the wrong scale (see Figure 6 last two rows - cactus and fence scales). Difix3D does not improve, as the progressive enhancement eventually reaches the wrong-scale regime and the inconsistent intermediate views degrade the 3DGS model. PR-Difix3D updates the reference progressively but does not consistently improve over Difix3D, and remains far below MACRO on all perceptual metrics.

MACRO achieves the best perceptual metrics on both benchmarks while maintaining competitive PSNR, confirming that scale-matched references enable the model to transfer correct scene content rather than hallucinating plausible but misaligned details. Results are consistent across DL3DV-Closeup and MobileClose-10, demonstrating generalization.

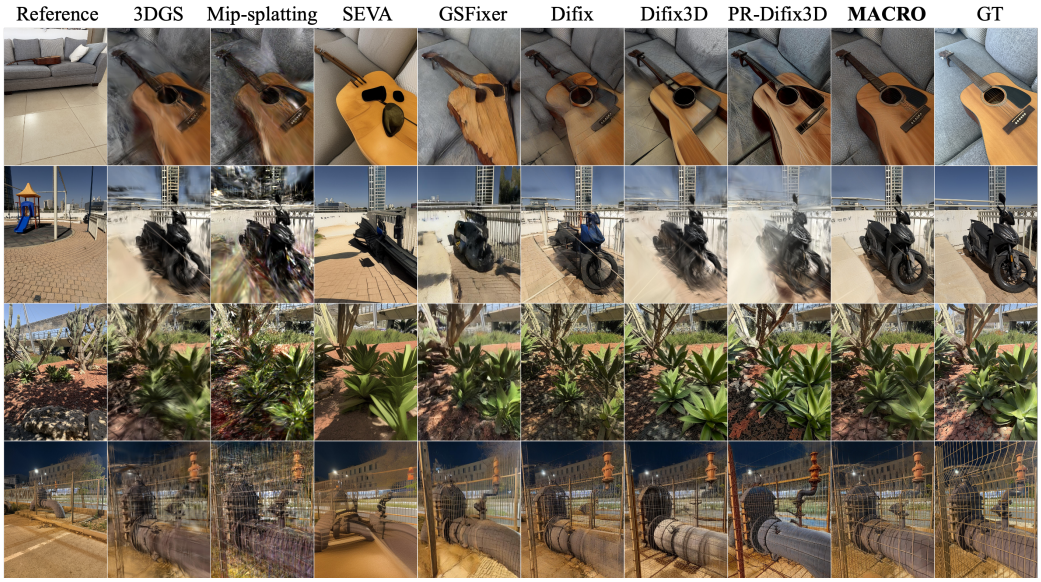


Figure 6: **Qualitative results on MobileClose-10.** MACRO produces close-up renderings that faithfully match the ground truth (up to exposure differences, as described in Appendix D), while baseline methods either remain blurry or transfer incorrect textures from the reference images. Best viewed zoomed-in on a digital screen.

### 5.3 Ablation studies

We ablate key design choices of MACRO on DL3DV-Closeup (Table 2). Reducing from $K{=}3$ to $K{=}1$ references drops perceptual metrics, confirming that multiple references provide complementary coverage. To isolate the contribution of scale matching from the multi-plane mechanism, we evaluate a *single-plane* variant ($P{=}1$, no mask) that applies one global scale-matched crop per reference. This variant alone already outperforms all enhancement baselines on every perceptual metric (Table 1), confirming that image-space scale correction is the primary driver. The full multi-plane formulation improves further, and its gain grows with the depth range within the close-up: on close-ups with a near-to-far depth ratio of at least 10, MACRO improves DreamSim by $10.5\%$ and DINOv2 by $10.4\%$ over the single plane (Appendix C.1), since a single crop can match only one scale. Notably, adding planes *without* the mask performs worse than a single crop, showing that the mask is what lets each token use the reference at its own scale. Replacing PFT-SR with LANCZOS upsampling yields comparable PSNR but worse perceptual scores, and backward-warping the reference instead of cropping causes large degradation, as depth noise in the warp destroys high-frequency texture. Finally, MACRO is robust to depth noise: perturbing the rendered depth by up to $20\%$ changes LPIPS by at most $0.007$ and DreamSim and DINOv2 by at most $0.001$ (Appendix C.2).

To choose the number of references $K$ and number of depth planes $P$ we performed grid search over a representative subset of 8 scenes from DL3DV-Closeup, as shown in Figure 7. We evaluated $K \in \{1, 2, 3, 4, 6, 8, 12\}$ and $P \in \{1, 2, 3, 5, 8, 10, 16\}$. We found that setting $K = P = 3$ resulted in maximal PSNR and minimal LPIPS at a moderate runtime; increasing $P$ has diminishing returns on quality. For $KP \geq 60$ we encountered OOM issues which limited the search, and higher $KP$ values also increased runtime - mainly attributed to the enhancer itself, as attention is quadratic in nature. The rest of the pipeline (reference selection, warping, cropping and super resolution) are all relatively negligible in terms of processing time. We therefore use MACRO with $K = P = 3$ throughout the paper.

### 5.4 Limitations

MACRO crops and resizes references to match the close-up scale, but the upsampled crops lose high-frequency details, introducing a blur-realism tradeoff. A dedicated super-resolution stage on the final output could alleviate this. Additionally, the number of depth planes $P$ is currently fixed

Table 2: **Ablation study on DL3DV-Closeup.** Each row varies a single component from the full MACRO pipeline. $K, P$: number of reference views and depth planes; $P{=}1$ uses a single global scale-matched crop per reference. Mask: depth-aware attention masking. Upsample: method used to resize reference crops, SR = PFT-SR [13].

| | Components | | | Metrics | | | | |
|---|---|---|---|---|---|---|---|---|
| Config | $K, P$ | Mask | Upsample | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DreamSim ↓ | DINOv2 ↓ |
| **MACRO (full)** | 3, 3 | ✓ | SR | 16.17 | 0.505 | **0.330** | **0.105** | **0.152** |
| $K{=}1$ | 1, 3 | ✓ | SR | <u>16.18</u> | <u>0.506</u> | 0.349 | 0.120 | 0.185 |
| $P{=}1$ | 3, 1 | — | SR | 16.13 | 0.501 | <u>0.332</u> | <u>0.108</u> | <u>0.157</u> |
| No mask | 3, 3 | — | SR | 16.12 | 0.501 | 0.333 | 0.110 | 0.168 |
| LANCZOS | 3, 3 | ✓ | LANCZOS | **16.20** | **0.509** | 0.360 | <u>0.108</u> | 0.164 |
| Warp | 3, 3 | ✓ | — | 15.97 | 0.474 | 0.399 | 0.211 | 0.412 |

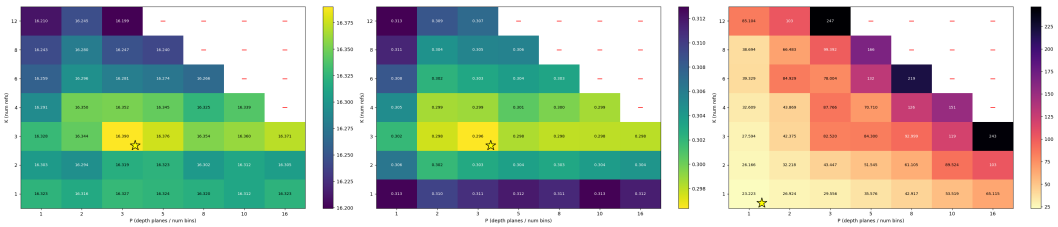

Figure 7: **Hyperparameter search of** $K, P$. **left**: PSNR, **middle**: LPIPS, **right**: end to end runtime in seconds. We observe optimal LPIPS,PSNR for $K = P = 3$, while configurations with $KP \geq 60$ cause OOM errors.

rather than adapted to the depth complexity of each view, leaving room for a per-image adaptive scheme. Moreover, MACRO relies on the 3DGS-rendered depth map for plane decomposition and crop computation, which may be inaccurate in regions with insufficient training views. Off-the-shelf depth estimators did not improve over the 3DGS depth, as they yield sparse depth at close range. Rarely, faint seams appear at plane boundaries; letting boundary tokens attend to adjacent planes is a simple remedy. Finally, distilling MACRO outputs back into the 3DGS with standard schemes degrades quality (Appendix C.3); close-up-aware 3DGS densification is a possible solution.

## 6 Conclusion

We addressed close-up novel view synthesis from 3DGS, where reference-conditioned diffusion enhancement fails because the underlying features are not scale-invariant and the VAE is not scale-equivariant, ruling out latent-space correction. MACRO resolves this by decomposing the close-up into depth planes and providing scale-matched reference crops through a depth-aware attention mask, injecting 3D knowledge into a 2D pipeline without architectural changes or finetuning. On two close-up benchmarks, MACRO improves perceptual metrics over the strongest enhancement baseline by up to 26% and 38% respectively, and by up to 67% and 58% over unenhanced 3DGS. As a training-free method, MACRO is readily applicable to settings where close-up training data is unavailable, such as fixed camera arrays used for dynamic scenes. Given the plug and play nature of MACRO, we expect that leveraging it on top of stronger generative backbones would result in further improvements, closing the gap in close-up novel view synthesis.

## Acknowledgments and Disclosure of Funding

This work was funded by Amazon. The authors did not receive any third-party funding or support for this work. All authors were employed by Amazon (Prime Video) while conducting this work. The authors declare no competing interests beyond the affiliations listed.

## References


[1] Pranav Asthana, Alex Hanson, Allen Tu, Tom Goldstein, Matthias Zwicker, and Amitabh Varshney. Splatsure: Selective super-resolution for multi-view consistent 3d gaussian splatting, 2026. URL https://arxiv.org/abs/2512.02172.

[2] Riccardo de Lutio, Tobias Fischer, Yen-Yu Chang, Yuxuan Zhang, Jay Zhangjie Wu, Xuanchi Ren, Tianchang Shen, Katarina Tothova, Zan Gojcic, and Haithem Turki. Artifixer: Enhancing and extending 3d reconstruction with auto-regressive diffusion models. In *SIGGRAPH*, 2026.

[3] Xiang Feng, Tieshi Zhong, Shuo Chang, Weiliu Wang, Chengkai Wang, Yifei Chen, Tongyu Hu, Yuhe Wang, Zhenzhong Kuang, Xuefei Yin, and Yanming Zhu. IE-SRGS: an internal-external knowledge fusion framework for high-fidelity 3d gaussian splatting super-resolution. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor, editors, *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 3957–3965. AAAI Press, 2026. doi: 10.1609/AAAI.V40I5.37398. URL https://doi.org/10.1609/aaai.v40i5.37398.

[4] Tobias Fischer, Samuel Rota Bulò, Yung-Hsu Yang, Nikhil Keetha, Lorenzo Porzi, Norman Müller, Katja Schwarz, Jonathon Luiten, Marc Pollefeys, and Peter Kontschieder. FlowR: Flowing from sparse to dense 3d reconstructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[5] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, pages 50742–50768, 2023.

[6] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T. Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models, 2024. URL https://arxiv.org/abs/2405.10314.

[7] Sergio Izquierdo, Mohamed Sayed, Michael Firman, Guillermo Garcia-Hernando, Daniyar Turmukhambetov, Javier Civera, Oisin Mac Aodha, Gabriel Brostow, and Jamie Watson. Mvsanywhere: Zero-shot multi-view stereo, 2025. URL https://arxiv.org/abs/2503.22430.

[8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433. URL https://doi.org/10.1145/3592433.

[9] Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. EQ-VAE: equivariance regularized latent space for improved generative image modeling. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/kouzelis25a.html.

[10] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r, 2024. URL https://arxiv.org/abs/2406.09756.

[11] Zhihao Liang, Qi Zhang, Wenbo Hu, Lei Zhu, Ying Feng, and Kui Jia. Analytic-splatting: Anti-aliased 3d gaussian splatting via analytic integration. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVII*, Lecture Notes in Computer Science, pages 281–297. Springer, 2024. doi: 10.1007/978-3-031-72643-9_17. URL https://doi.org/10.1007/978-3-031-72643-9_17.

[12] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3d vision. In *IEEE/CVF Conference*

*on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22160–22169. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02092. URL https://doi.org/10.1109/CVPR52733.2024.02092.

[13] Wei Long, Xingyu Zhou, Leheng Zhang, and Shuhang Gu. Progressive focused transformer for single image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2279–2288, 2025.

[14] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20654–20664, 2024.

[15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

[16] Md Ashiqur Rahman and Raymond A. Yeh. Truly scale-equivariant deep nets with fourier layers, 2023. URL https://arxiv.org/abs/2311.02922.

[17] Ivan Sosnovik, Artem Moskalev, and Arnold Smeulders. Disco: accurate discrete scale convolutions. *arXiv preprint arXiv:2106.02733*, 2021.

[18] Daniel E. Worrall and Max Welling. Deep scale-spaces: Equivariance over scale, 2019. URL https://arxiv.org/abs/1905.11697.

[19] Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26024–26035, 2025.

[20] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. Reconfusion: 3d reconstruction with diffusion priors, 2023. URL https://arxiv.org/abs/2312.02981.

[21] Jiatong Xia and Lingqiao Liu. Close-up-gs: Enhancing close-up view synthesis in 3d gaussian splatting with progressive self-training, 2025. URL https://arxiv.org/abs/2503.09396.

[22] Jiatong Xia, Libo Sun, and Lingqiao Liu. Enhancing close-up novel view synthesis via pseudo-labeling, 2025. URL https://arxiv.org/abs/2503.15908.

[23] Shiyun Xie, Zhiru Wang, Xu Wang, Yinghao Zhu, Chengwei Pan, and Xiwang Dong. Supergs: Super-resolution 3d gaussian splatting enhanced by variational residual features and uncertainty-augmented learning, 2024. URL https://arxiv.org/abs/2410.02571.

[24] Zhiwen Yan, Weng Fei Low, Yu Chen, and Gim Hee Lee. Multi-scale 3d gaussian splatting for anti-aliased rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 20923–20931. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01977. URL https://doi.org/10.1109/CVPR52733.2024.01977.

[25] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

[26] Xingyilang Yin, Qi Zhang, Jiahao Chang, Ying Feng, Qingnan Fan, Xi Yang, Chi-Man Pun, Huaqi Zhang, and Xiaodong Cun. Gsfixer: Improving 3d gaussian splatting with reference-guided video diffusion priors. *CoRR*, abs/2508.09667, 2025. doi: 10.48550/ARXIV.2508.09667. URL https://doi.org/10.48550/arXiv.2508.09667.

[27] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis, 2024. URL https://arxiv.org/abs/2409.02048.

[28] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 19447–19456. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01839. URL https://doi.org/10.1109/CVPR52733.2024.01839.

[29] Jiahui Zhang, Fangneng Zhan, Muyu Xu, Shijian Lu, and Eric P. Xing. Fregs: 3d gaussian splatting with progressive frequency regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 21424–21433. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02024. URL https://doi.org/10.1109/CVPR52733.2024.02024.

[30] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL https://arxiv.org/abs/1801.03924.

[31] Yuqi Zhang, Guanying Chen, Jiaxing Chen, Chuanyu Fu, Chuan Huang, and Shuguang Cui. Closeupshot: Close-up novel view synthesis from sparse-views via point-conditioned diffusion model, 2025. URL https://arxiv.org/abs/2511.13121.

[32] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models, 2025. URL https://arxiv.org/abs/2503.14489.

# MACRO: Training-free Multi-plane Attention for Closeup Render Optimization

# Supplementary Material

The supplementary material provides additional qualitative results, extended ablation studies, implementation details, and further details on the construction of the close-up benchmarks.

## Appendix Contents:

A. Additional Qualitative Results (Appendix A)

B. Implementation Details (Appendix B)

C. Additional Ablations (Appendix C)

D. Benchmark Construction Details (Appendix D)

# A Additional qualitative results

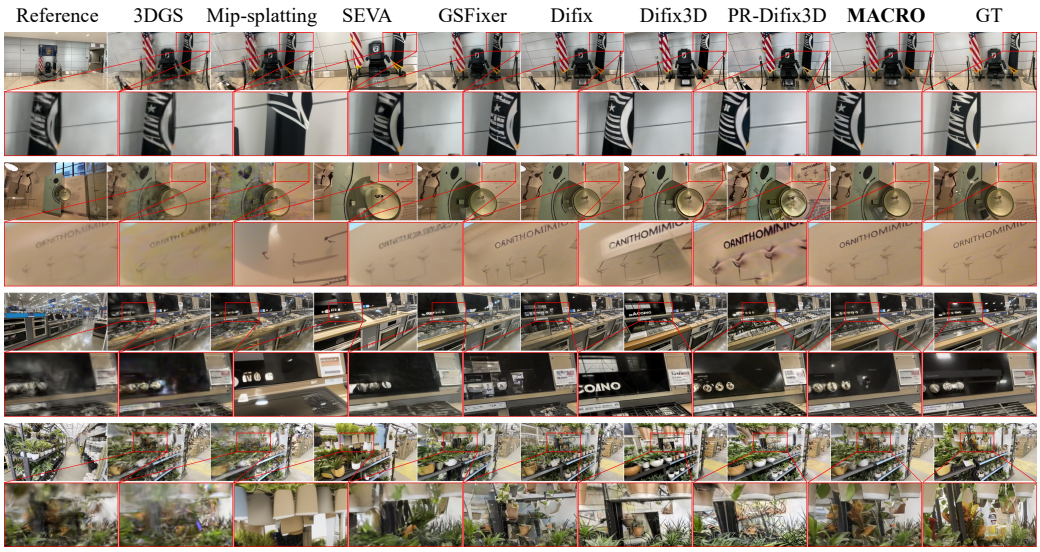


Figure 8: **Additional qualitative results on DL3DV-Closeup.** Four examples shown, each with a full-frame view and a crop below for detail. MACRO produces more faithful renderings that better match the ground truth. Best viewed zoomed-in on a digital screen.

# B Implementation details

## B.1 MACRO implementation details

**Reference selection.** For each close-up render, MACRO selects $K$ reference views from the training set by greedy set cover over pixel observability: each candidate's 3dgs rasterized depth map is unprojected to world coordinates and reprojected into the close-up camera, and references are added one at a time to maximize newly covered pixels. References who cannot see the close-up completely due to full occlusion determined by reference view depth are discarded. We use the 3DGS rasterized depth throughout; we tested alternative SOTA depth estimation methods for reference elimination module, including MVSAnywhere[7] and Mast3r[10], which produced negligible differences. These methods were not applicable to estimate input depth however, as the input close-up render is too corrupted.

**Compute resources.** Because MACRO is training-free, the total compute is dominated by inference: per close-up, the depth-plane construction, reference selection, cropping, and resizing add negligible overhead on top of the underlying enhancement backbone, and the super-resolution module adds a small fixed cost per crop. MACRO's per-view runtime is therefore close to that of its backbone [19], which is itself a single-step distilled diffusion model. All experiments were run on a single NVIDIA P5 instance over approximately one week. MACRO inference takes $\sim$34 seconds per image ($540 \times 960$ resolution) on DL3DV-Closeup and $\sim$104 seconds on MobileClose-10 ($1071 \times 1428$ resolution). These are much faster compared to the other solutions that either require multi-step video diffusion or iterative finetuning of the 3dgs representation, which can take over 30 minutes per scene.

## B.2 Baseline adaptations for close-up novel view synthesis

**Difix3D, Difix3D+, and Progressive-Reference variants.** Unlike Difix [19], which applies a single post-processing correction to the rendered close-up view at inference time, Difix3D [19] progressively refines the 3DGS representation by adding diffusion-corrected rendered views along the path that connects the training views to the close-up camera. In addition to the original implementation, we evaluate a variant that progressively updates the reference set, which we call **Progressive-Reference Difix3D (PR-Difix3D)**. After each refinement round, the newly corrected intermediate views are

added as candidate references for subsequent iterations. **PR-Difix3D+**, is the analogous extension of Difix3D+.

**GSFixer.** GSFixer [26] is a video diffusion model based on a fine-tuned CogVideoX-5B [25], designed to correct 3DGS rendering artifacts. To adapt GSFixer to our close-up benchmarks, we construct a 49-frame sequence for each target close-up view. Specifically, we select the two best training references using the same heuristic as Difix3D. We then interpolate frames from the first reference to the close-up view and from the close-up view to the second reference, placing the target close-up at the center of the sequence.

# C Additional ablations

## C.1 Single-plane ablation

To quantify how much of MACRO's gain comes from scale correction alone, we evaluated a single-global-crop variant ($P{=}1$, $K{=}3$, no depth mask) against full MACRO on all 44 DL3DV-Closeup scenes (Table 2). Averaged over the benchmark, full MACRO improves over the $P{=}1$ variant by $+3.1\%$ on DreamSim ($0.108 \rightarrow 0.105$), $+3.0\%$ on DINOv2 ($0.157 \rightarrow 0.152$) and $+0.5\%$ on LPIPS ($0.332 \rightarrow 0.330$). This gap, however, is far from uniform: it concentrates exactly where multiple planes are needed. A single global crop can match only one scale, so it suffices for planar close-ups but fails where a view spans a wide range of depths. Stratifying by within-view depth disparity, the gain of full MACRO over the $P{=}1$ variant increases monotonically with the depth ratio. Averaged over the 33 close-ups in the highest stratum (depth ratio $\geq 10$), full MACRO improves DreamSim from $0.142$ to $0.127$ ($+10.5\%$), DINOv2 from $0.219$ to $0.196$ ($+10.4\%$), and LPIPS from $0.374$ to $0.365$ ($+2.3\%$). On individual views the effect is even more pronounced; for example, on a scene with depth ratio $\approx$13, DreamSim improves from $0.254$ to $0.180$ ($+29\%$), DINOv2 from $0.592$ to $0.448$ ($+24\%$), and LPIPS from $0.414$ to $0.390$ ($+6\%$).

We see this as reflecting two distinct contributions. First, the diagnostic analysis of Section 4.1: applying it as a single global crop, with no additional machinery, training, or 3DGS re-optimization, already outperforms the enhancement baselines (Table 1) on all three perceptual metrics by $+3.4\%$ to $+23.5\%$. Second, the full multi-plane formulation, which is the general form of scale matching rather than an add-on to it: because a close-up observes content across a range of depths, the correct reference scale is not global, and the single global crop is its degenerate planar special case ($P{=}1$). The median within-view depth ratio in DL3DV-Closeup is only $3.5\times$, so planar cases dominate the aggregate, but the full formulation is what keeps MACRO correct on the harder, non-planar cases where a single crop fails.

## C.2 Robustness to depth errors

MACRO relies on the 3DGS-rendered close-up depth, which can be inaccurate in under-observed regions. To quantify sensitivity, we perturb the rendered depth with multiplicative Gaussian noise, $\tilde{\mathbf{D}} = \mathbf{D}_{\text{cu}}(1 + \sigma\mathcal{N}(0,1))$, and rerun the full pipeline (Table 3). Even at $\sigma{=}20\%$, metrics remain essentially unchanged, since depth enters only through aggregate statistics: per-plane median depths for the scale factors (Eq. (5)) and $k$-means centroids, pooled to latent-token resolution, for the mask (Eq. (6)). MACRO also performs a single feed-forward pass, so depth errors cannot accumulate across steps.

**Alternative depth sources.** We also tested off-the-shelf estimators, MVSAnywhere [7] and MASt3R [10], in place of the 3DGS depth. They gave equal or worse results: when fit to wide-shot images and queried at a close-up, their depth is sparse, and naive completion does not densify it correctly. Because the depth is used to partition the close-up render itself, the 3DGS depth is more consistent with that image. Close-up-specific depth completion could close this gap.

## C.3 Distillation into 3DGS

Following Difix3D [19], one may distill enhanced close-ups back into the 3DGS. We evaluate three variants: (1) the progressive distillation of Difix3D, (2) that of Difix3D+, and (3) a single distillation step directly from MACRO-enhanced close-ups (Table 4). Variants 1 and 2 degrade on all metrics;

Table 3: **Depth perturbation on DL3DV-Closeup.** The rendered close-up depth is perturbed with multiplicative Gaussian noise of standard deviation $\sigma$.

| Config | Mask | $K, P$ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DreamSim ↓ | DINOv2 ↓ |
|---|---|---|---|---|---|---|---|
| MACRO | ✓ | 3, 3 | 16.17 | 0.505 | 0.330 | 0.105 | 0.152 |
| $\sigma$=10% | ✓ | 3, 3 | 16.17 | 0.505 | 0.336 | 0.104 | 0.152 |
| $\sigma$=20% | ✓ | 3, 3 | 16.18 | 0.505 | 0.337 | 0.105 | 0.152 |

variant 3 degrades on PSNR, SSIM and LPIPS while slightly improving DINOv2. This mirrors the Difix vs. Difix3D trend in Table 1, although the degradation is milder because scale errors are already mitigated. We attribute this to standard 3DGS fine-tuning being ill-suited to the high-disparity close-up setting; dedicated schemes, such as densifying the close-up region or reweighting close and far camera losses, are a promising direction.

Table 4: **Distilling MACRO into 3DGS on DL3DV-Closeup.** Variants (1) and (2) follow the distillation of Difix3D and Difix3D+, respectively; variant (3) distills directly from MACRO-enhanced close-ups.

| Configuration | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DreamSim ↓ | DINOv2 ↓ |
|---|---|---|---|---|---|
| MACRO (no distillation) | **16.17** | **0.505** | **0.330** | **0.105** | 0.152 |
| Distillation (variant 1) | 15.56 | 0.469 | 0.366 | 0.116 | 0.177 |
| Distillation (variant 2) | 15.32 | 0.454 | 0.356 | 0.117 | 0.168 |
| Distillation (variant 3) | 15.81 | 0.484 | 0.339 | **0.105** | **0.146** |

# D Benchmark Construction Details

We provide additional details for the close-up novel view synthesis benchmarks used in our experiments. We first describe the construction of the DL3DV-Closeup benchmark in Appendix D.1, followed by MobileClose-10 (Appendix D.2), which was captured specifically for this task.

**Zoom statistics.** Estimating the effective zoom factor for closeups directly from COLMAP or 3DGS-rendered depth is challenging due to sparsity and noise. We therefore use MASt3R [10] depth estimation for the scene (including the closeup GT images), which provides cleaner depth. We then estimate the effective zoom factor as Zoom $= \|\mathbf{c}_W - \mathbf{P}\|/\|\mathbf{c}_C - \mathbf{P}\|$, where $\mathbf{c}_C$ and $\mathbf{c}_W$ are the closeup and wide camera centers and $\mathbf{P}$ is the median of the 3D points unprojected from the closeup that fall inside the wide view. Across DL3DV-Closeup we measure Zoom $= \times 2.59 \pm 0.73$ (range $\times 1.46$-$\times 5.81$), and across MobileClose-10 Zoom $= \times 3.24 \pm 1.03$ (range $\times 1.47$-$\times 5.16$), confirming that both benchmarks contain genuine closeups with substantial scale gap from the training views.

**Exposure differences.** In some scenes from both benchmarks, the close-up and wide views were captured under slightly different exposure settings. Reference-conditioned enhancers, including MACRO, inherit the wide-shot's exposure when transferring appearance from the reference. This caps the achievable PSNR and SSIM on these scenes for all reference-conditioned methods we evaluate, and is one reason the absolute reconstruction scores in Table 1 are modest. Since all enhancement baselines are affected equally, the comparison between methods remains fair.

## D.1 DL3DV-Closeup Benchmark

We construct DL3DV-Closeup from the DL3DV-10K benchmark [12], which contains 140 scenes. We filter scenes according to depth disparity and valid close-up/far-view relationships, resulting in 44 scenes and a total of 314 close-up evaluation images. We show examples of far and close-up views in Appendix D.1.1, and describe the construction of the benchmarks in Appendices D.1.2- D.1.5.

### D.1.1 Dataset Examples

Figure 9 shows example scenes from our DL3DV-Closeup benchmark. Each example includes far training views and corresponding close-up evaluation views, illustrating the large scale changes and partial-view setting targeted by the benchmark.

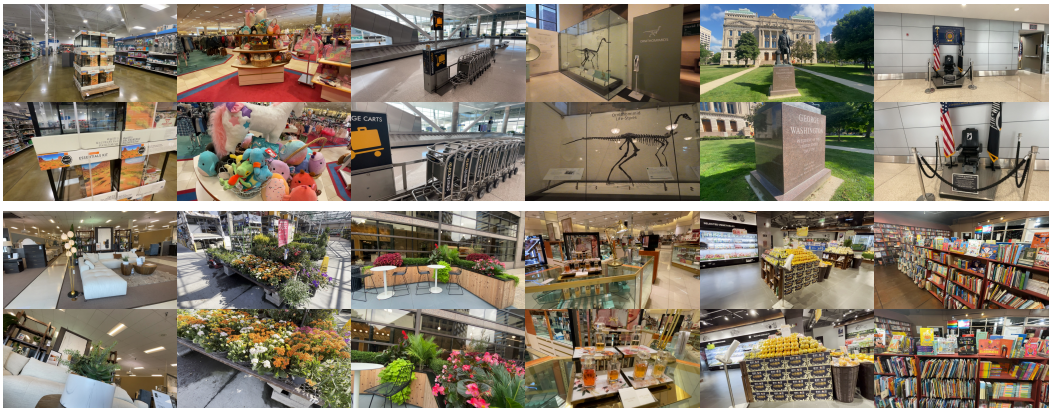

Figure 9: **Example scenes from DL3DV-Closeup.** Odd rows show far training views, even rows show corresponding close-up evaluation views. Each column is one scene. Best viewed zoomed-in on a digital screen.

### D.1.2 Per-Camera Depth Estimation

For each scene, we use the COLMAP sparse reconstruction $\mathcal{X} = \{x_m\}_{m=1}^{M}$ and camera poses $\{(R_i, t_i)\}_{i=1}^{N}$. Each camera $\pi_i$ is represented by its pose $(R_i, t_i)$ and viewing direction $\mathbf{v}_i$, where $\mathbf{v}_i$ denotes the unit optical axis.

We define the set of visible points for camera $i$ as:

$$\mathcal{V}_i = \left\{ x_m \in \mathcal{X} \;\middle|\; \frac{\langle x_m - t_i,\ \mathbf{v}_i \rangle}{\|x_m - t_i\|} > \cos\left(\tfrac{\theta_{\text{fov}}}{2}\right),\ \langle x_m - t_i,\ \mathbf{v}_i \rangle > 0 \right\}, \tag{8}$$

where $\theta_{\text{fov}}$ is the diagonal field of view derived from the camera intrinsics. This corresponds to 3D points within the camera's field of view and in front of the camera, providing a coarse approximation of visible scene content.

We define the *median scene depth* for camera $i$ as:

$$d_i = \text{median}_{x_m \in \mathcal{V}_i} \langle x_m - t_i,\ \mathbf{v}_i \rangle. \tag{9}$$

This corresponds to the median depth of visible content along the camera viewing direction.

We further compute the 10th and 90th percentile depths, $d_i^{10}$ and $d_i^{90}$, and define the within-view depth span as $s_i = d_i^{90}/d_i^{10}$, capturing the range of depths observed within a single view.

### D.1.3 Candidate Pair Construction

For each scene, cameras are partitioned into a *close* set (bottom 40% by $d_i$) and a *far* set (top 40% by $d_i$).

We define candidate close-up/far pairs $(i, j)$, where $i$ is a close camera and $j$ is a far camera. A pair must satisfy:

- Depth ratio $\frac{d_j}{d_i} \geq 2$,
- Intersection-over-Union (IoU) $\geq 40\%$, computed between the sets of visible 3D points $\mathcal{V}_i$ and $\mathcal{V}_j$,
- Viewing angle difference between $10^\circ$ and $60^\circ$.

Let $N_{\text{cand}}$ denote the number of such valid candidate pairs in the scene.

#### D.1.4 Scene Selection

We rank scenes using a disparity score that captures both cross-view and within-view depth variation:

$$S = \frac{\sigma(\{d_i\})}{\mu(\{d_i\})} \cdot (1 + \log(1 + \bar{s})) \cdot \left(1 + \log\left(1 + \frac{\max_i d_i^{10}}{\min_i d_i^{10}}\right)\right) \cdot \left(1 + \log\left(1 + \frac{1}{\min_i d_i}\right)\right) \cdot \log(1 + N_{\text{cand}}) \tag{10}$$

where $\bar{s} = \frac{1}{N}\sum_i s_i$, and $\sigma(\cdot)$ and $\mu(\cdot)$ denote standard deviation and mean, respectively.

The score is composed of the following factors:

- **Cross-view depth variation:** $\frac{\sigma(\{d_i\})}{\mu(\{d_i\})}$ measures the coefficient of variation of per-camera depths, capturing how much scene depth varies across viewpoints.
- **Within-view depth span:** $\bar{s}$ summarizes the depth range within individual views, where $s_i = d_i^{90}/d_i^{10}$.
- **Near-depth variation:** $\frac{\max_i d_i^{10}}{\min_i d_i^{10}}$ captures variation in the closest visible content across views.
- **Closeness:** $\min_i d_i$ encourages the presence of cameras observing very close scene content.
- **Pair count:** $N_{\text{cand}}$ denotes the number of valid candidate close-up/far pairs, ensuring that meaningful pairs exist in the scene.

Logarithmic scaling is applied to stabilize the contribution of each term and prevent extreme values from dominating the score.

Scenes with $N_{\text{cand}} = 0$ are assigned a score of zero and discarded.

#### D.1.5 Train/Test Split and Evaluation Pairs

For each selected scene, frames are sorted by their median depth $d_i$ and divided as follows:

- Top 30% (farthest views): training pool,
- Bottom 30% (closest views): test candidates.

The middle 40% is discarded to enforce a clear separation between training and test views.

From the training pool, we select $N = 16$ views using farthest-point sampling over camera positions to maximize spatial coverage of the scene geometry.

### D.2 MobileClose-10 Benchmark

In addition to DL3DV-Closeup, we introduce MobileClose-10, a close-up benchmark captured specifically for this task. The dataset was collected using an iPhone 17 Pro Max across 10 diverse real-world scenes.

Each scene contains 15-18 far-view images used for training, together with 2-6 close-up evaluation views, resulting in a total of 39 close-up evaluation images. The close-up views are captured with significant changes in camera distance relative to the training views, with a median zoom factor of $3.42\times$ between the close-up evaluation views and the corresponding far views. Example far and close-up views from this benchmark are shown in Figure 10.

This dataset is designed to emphasize challenging close-up novel view synthesis scenarios, where the target views exhibit substantial scale changes with the training views.



Figure 10: **Example scenes from MobileClose-10.** The first row shows far training views, while the second row shows the close-up views. Each column represents examples from one scene.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions: the scale-sensitivity analysis of diffusion enhancers (non-invariance of features and non-equivariance of the VAE), the proposed training-free MACRO method, and the two new close-up benchmarks on which we demonstrate state-of-the-art results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5.4 discusses the limitations of MACRO.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be

used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not contain formal theoretical results. The scale-sensitivity properties discussed in Section 4.1 are established empirically (Figure 3) and supported by citations to prior work on scale equivariance in neural networks.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sections 4.2-4.4 describe the MACRO pipeline in full. Appendix B and C provide implementation details including hyperparameters, reference selection, runtime, and baseline adaptations. Appendix D details benchmark construction. Code and data are publicly available at the project page linked under the title.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case

of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code, the MobileClose-10 dataset and the DL3DV-Closeup splits are publicly available at the project page linked under the title, with instructions to reproduce the main results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 and Appendices B and D cover the evaluation and implementation details (reference selection, inference resolution, runtime) and Section 5.3 documents the hyperparameter search that selected $K = P = 3$. MACRO is training-free, so training hyperparameters (optimizer, learning rate) apply only to the 3DGS reconstruction and baseline methods, where we follow the standard original implementations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report shaded error regions for the scale-sensitivity analysis in Figure 3. For the main method comparison in Table 1, we report means over the 44 DL3DV-Closeup and 10 MobileClose-10 scenes (the mean of per-scene means over their 314 and 39 evaluation pairs) without error bars or significance tests. Multiple full evaluation runs would be too computationally expensive given the inference cost of the diffusion-based baselines.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide a description of the resources required in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Reviewed the code of ethics, it does.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: MACRO's improvements for close-up rendering are an important step for applications in virtual production, interactive 3D content, and immersive media, who wish to enable users to freely explore captured scenes at any distance; like other reference-conditioned diffusion enhancers, it inherits the general risks of generative image models, including the potential to produce convincing but inaccurate visual content.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: MACRO releases no pretrained generative model, and the MobileClose-10 dataset consists of 10 generic scenes captured by us with no people or private content, posing no meaningful misuse risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes. Existing assets (DL3DV-10K, the baseline methods and pretrained models) are cited, and their licenses and terms of use are respected; our release includes the corresponding license and credits.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code and datasets (MobileClose-10, DL3DV-Closeup splits) are documented in the README of the code repository, linked from the project page, including setup, data layout and evaluation instructions. MobileClose-10 was captured by the authors and contains no people or private content.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used solely for formatting, editing or standard assistance, not in the core methodology or research ideation.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.