Unified Latent Space for Understanding and Generation via Semantic Auto-encoder

Supplementary Material

1. More Diffusion Transformer Generation Results

Comparison with DC-AE under different spatial compression ratios. Since S-AE employs a substantially higher channel dimension than conventional VAEs, we hypothesize that it may contain greater spatial redundancy. To examine this, we conduct an ablation study using the recent state-of-the-art model DC-AE as a baseline. We train S-AE under $32\times$ and $64\times$ spatial compression ratios and use both S-AE and DC-AE latents to train DiT models in a single-image overfitting setup. The training loss curves are shown in Figure 3, and qualitative comparisons are presented in Figure 5. As illustrated, the DiT model trained with S-AE latents converges faster and more smoothly than that with DC-AE, and produces higher-quality reconstructions, capturing fine details such as text and small facial features. We attribute these advantages to the semantic representation and fine-grained feature preservation of DINOv3: its higher-dimensional channels encode semantically meaningful local structures, enabling reduced spatial resolution while retaining detailed visual information during decoding.

More visualizations on fine-detail generation. To further demonstrate that the improvements of S-AE over RAE and VAE are consistent and generalizable, we present additional DiT training results based on the latent space of S-AE, as shown in Figure 8, Figure 7, and Figure 6. Across these examples—such as letter reconstruction, small facial features, and fine textures—DiT models trained on S-AE latents produce noticeably sharper and more faithful details compared to those trained on VAE or RAE latents. A comparison with the recent state-of-the-art VAE-based model DC-AE is also provided in Figure 5, highlighting S-AE's superior capability in preserving fine-grained visual structures.

2. More Evaluation on Generation Quality

In this section, we provide more qualititive and quantitative evaluation results on the generation quality of our S-AE compared to baseline methods like R-AE [**?**] and VAE [**?**]. As shown in Figure 1, we conducted more evaluation on high-resolution medical [**?**], remote sensing [**?**], and cartoon[1] domain-specific datasets. which are relatively small with each having 1600, 24, and 3140 samples. The results are shown in Fig 1 and Fig 2. Additional experiments on DiT training and reconstruction at different resolutions are in Tab.1. S-AE consistently produces more accurate geometry and details.

[1] https://huggingface.co/datasets/Norod78/cartoon-blip-captions

Table 1. Reconstruction metrics on low-data regimes and high-resolution datasets.

Model	PSNR↑/SSIM↑/rFID↓	
	Retina	Cartoon
VAE	40.14 / 0.95 / 8.30	33.28/0.93/0.17
R-AE	26.19 / 0.85 / 24.53	20.45/0.71/8.06
S-AE	**47.45 / 0.99 / 1.07**	**43.34/0.99/0.17**

3. More Ablations on the Model Backbone

In our experiments in our main paper, we used Dinov3 as the encoder backbone by default. In this section, we provide more ablation results where we replace the backbone with different models, as well as a comparison of efficiency. We applied identical training settings with SigLIP2-B and DINOv3-ViT-L. The results are reported in Tab.2. These results demonstrate that S-AE is not tied to a specific backbone, and its effectiveness generalizes to different pre-trained semantic encoders with distinct representation characteristics. However, the final performance is indeed related to the base encoder. The inference latency results measured on H20 GPU ($256\times$) are reported in Tab.2 which shows S-AE w/ SigLIP2-B e2e latency is 5% faster than VAE.

Table 2. Inference Efficiency by Latency (ms).

Model	**VAE**	**SAE**		
		Dinov3-H+	Dinov3-L	SigLip2-B
Encoder	7.37	23.65	12.09	4.94
Decoder	13.80	18.56	16.07	15.21

We note that while we used Dinov3 by default, our framework remains effective for different choices of backbones. The whole framework is essential for the final SOTA S-AE model. The key insight of our framework is identifying fundamental tradeoffs in semantic-based AEs between semantic abstraction and geometric structure preservation. Accordingly, we designed a semantic loss $\mathcal{L}_{\text{reg}}$ with its weight λ_{reg} to unify vision understanding and high-quality reconstruction in a unified latent space. Our proposed S-

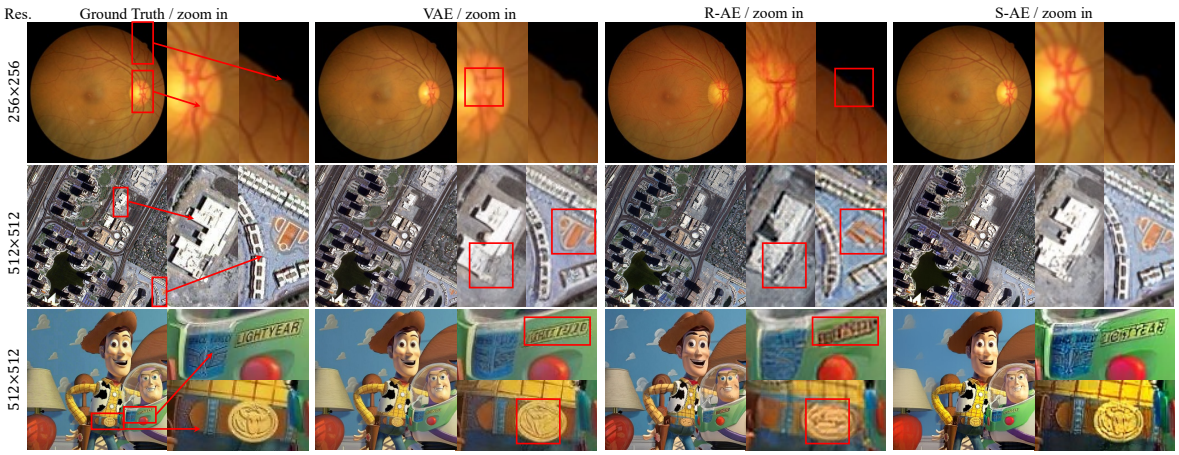

Figure 1. DiT training on domain-specific and low-data regimes datasets

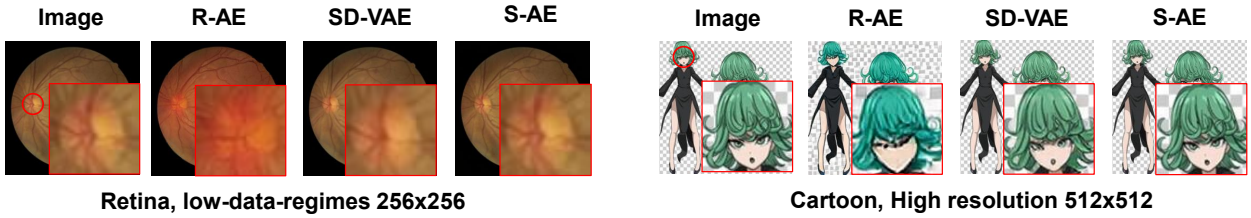

Figure 2. Qualitative comparison on low-data regimes, domain specfic, and high rsolution reconstruction

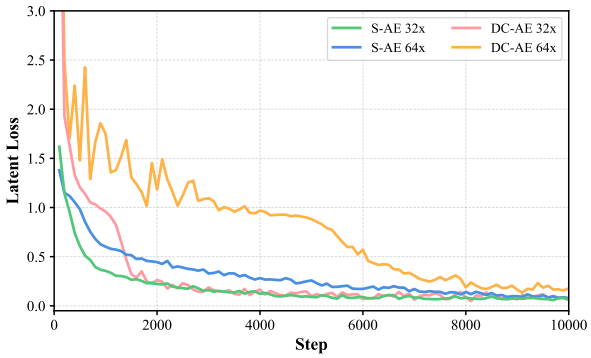

Figure 3. Training convergence speed comparison between S-AE and DC-AE with different spatial compression rates. It is clearly observed that S-AE converges more stably with faster convergence speed.

AE architecture and protocol achieve SOTA reconstruction while preserving classification accuracy.

4. More Visualizations on Semantic and Reconstruction Quality

We further provide visual evidence showing that S-AE delivers superior reconstruction fidelity and produces semantically meaningful latent representations. As illustrated in Figure 9 and Figure 10, S-AE distinguishes different image components more clearly than VAE, indicating stronger structural understanding. Compared to RAE, whose latents highlight relevant regions but contain substantial noise, S-AE generates cleaner and more disentangled latent activations, reflecting its ability to preserve fine-grained details while maintaining semantic organization.

5. More Detailed Related Works

Generative models have exhibited remarkable success across diverse domains, driven by continual progress in how models represent and generate data. Early advances were led by Generative Adversarial Networks (GANs) [**?** **?** **?**], which pioneered adversarial learning to synthesize photorealistic imagery from latent codes. More recently, diffusion models [**?**] have redefined generative modeling by formulating data synthesis as an iterative denoising process, achieving unprecedented performance in text-to-image and video generation [**?** **?** **?** **?** **?**]. To efficiently represent visual information, diffusion models are typically trained within the latent spaces of Variational Autoencoders (VAEs) [**?** **?**], which compress pixel data into a lower-dimensional manifold, enabling high-fidelity yet computationally efficient synthesis. Complementarily, autoregressive models [**?** **?** **?**] have adopted GPT-style [**?**] token prediction paradigms to handle discrete visual tokens, further demonstrating the importance of learned latent representations in generative efficiency and visual realism.

Figure 4. Comparison with DC-AE under different spatial compression rates. Note that R-AE uses a default spatial compression rate of 16, and VAE of 8.

Figure 5. Comparison with DC-AE under different spatial compression rates. Note that R-AE uses a default spatial compression rate of 16, and VAE of 8.

Recent research has increasingly focused on enhancing these latent representations through **semantic auto-encoders**, which integrate pretrained vision-language or self-supervised encoders into the VAE framework. Methods such as VA-VAE [**?**], MAETok [**?**], DC-AE 1.5 [**?**], and l-DEtok [**?**] leverage pretrained encoders like MAE or DAE [**?**] to infuse semantics into VAE latent spaces, substantially improving both reconstruction fidelity and generative consistency. However, their reliance on heavily compressed, low-dimensional latents limits the preservation of geometric and structural details essential for downstream tasks. In contrast, semantic VAEs that directly reconstruct from frozen representation features—e.g., DINO or SigLIP embeddings—demonstrate that with a simple ViT-based decoder, one can achieve reconstruction quality on par with or exceeding SD-VAE [**?**], while maintaining semantically rich and discriminative latent spaces.

From the generative modeling perspective, these semantic auto-encoders bridge the gap between representation learning and data synthesis. Works such as REPA [**?**], DDT [**?**], and REG [**?**] show that aligning diffusion transformer latents with semantic encoder spaces accelerates convergence and enhances image quality, while ReDi [**?**] further extends this idea by generating both VAE latents and principal components of DINOv2 features within a unified diffusion framework. Together, these advances suggest that semantically structured latent spaces serve not only as compact encodings for reconstruction but also as shared representational grounds for understanding and generation. By directly training diffusion models on semantic latents, we enable faster convergence, improved generalization, and more coherent visual synthesis, paving the way for next-generation **semantic VAEs** that unify perception and generation under a single latent representation paradigm.

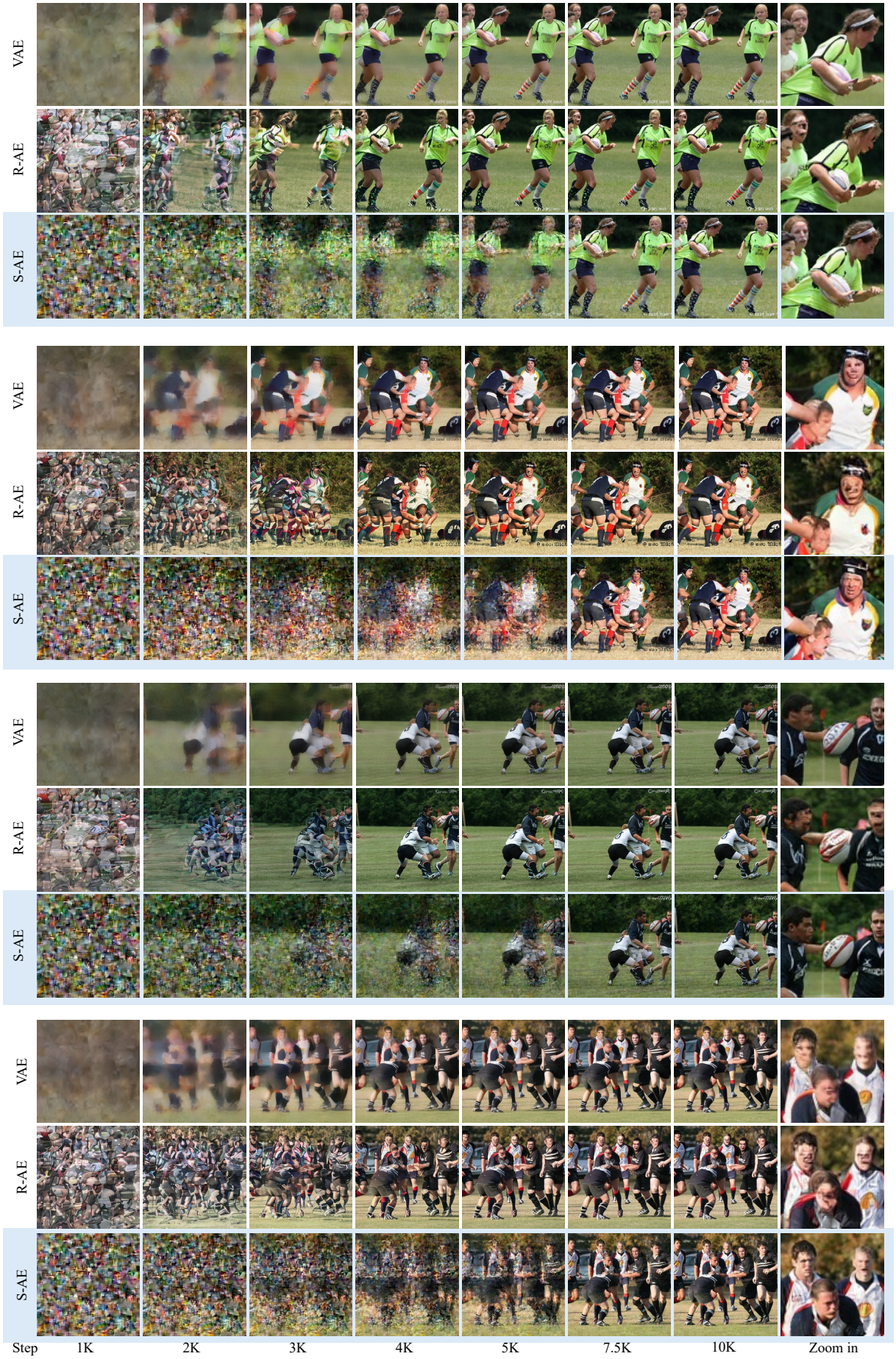

Figure 6. More visualizations for the single image diffusion training with different auto-encoders.

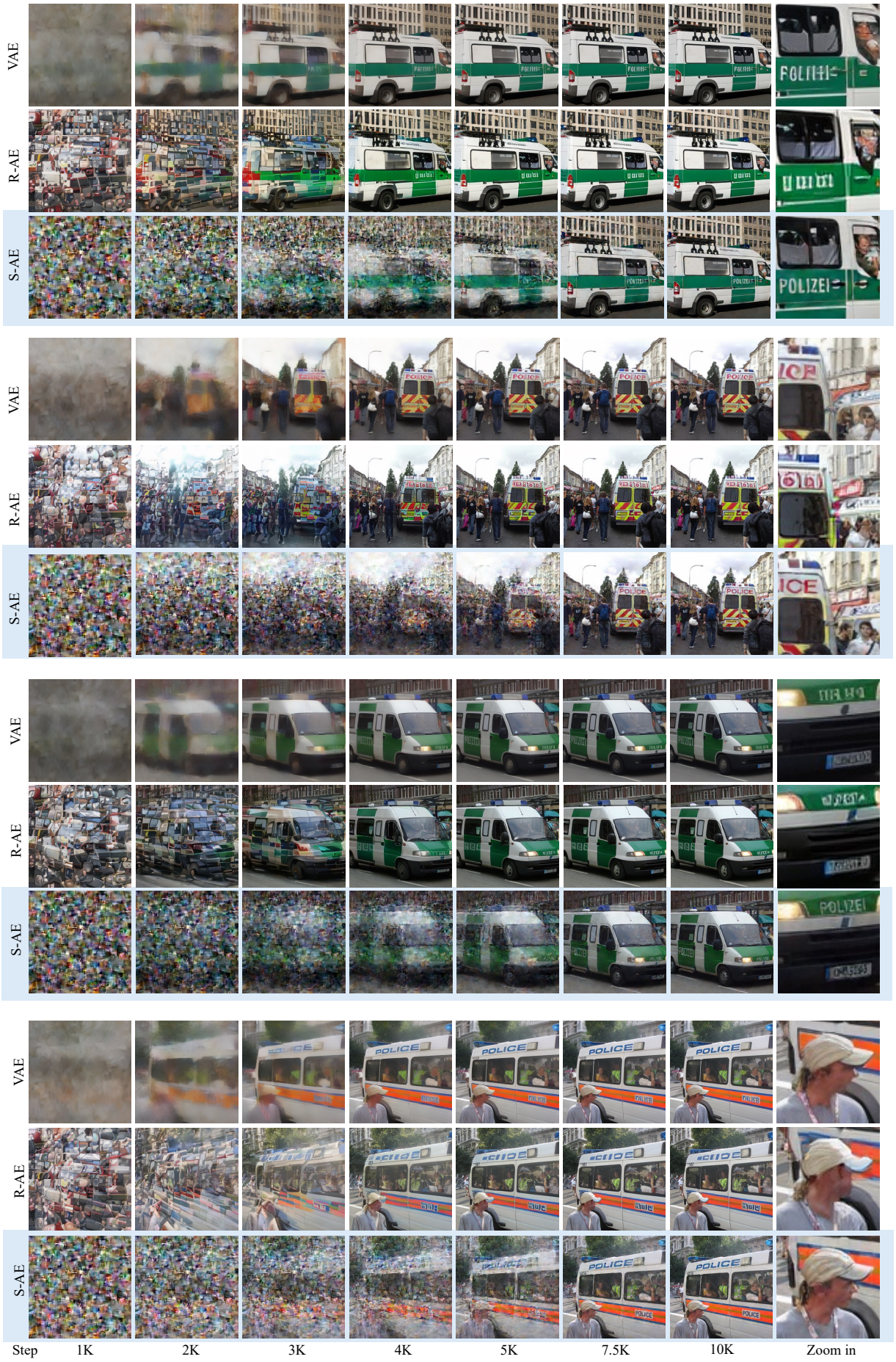

Figure 7. More visualizations for the single image diffusion training with different auto-encoders.

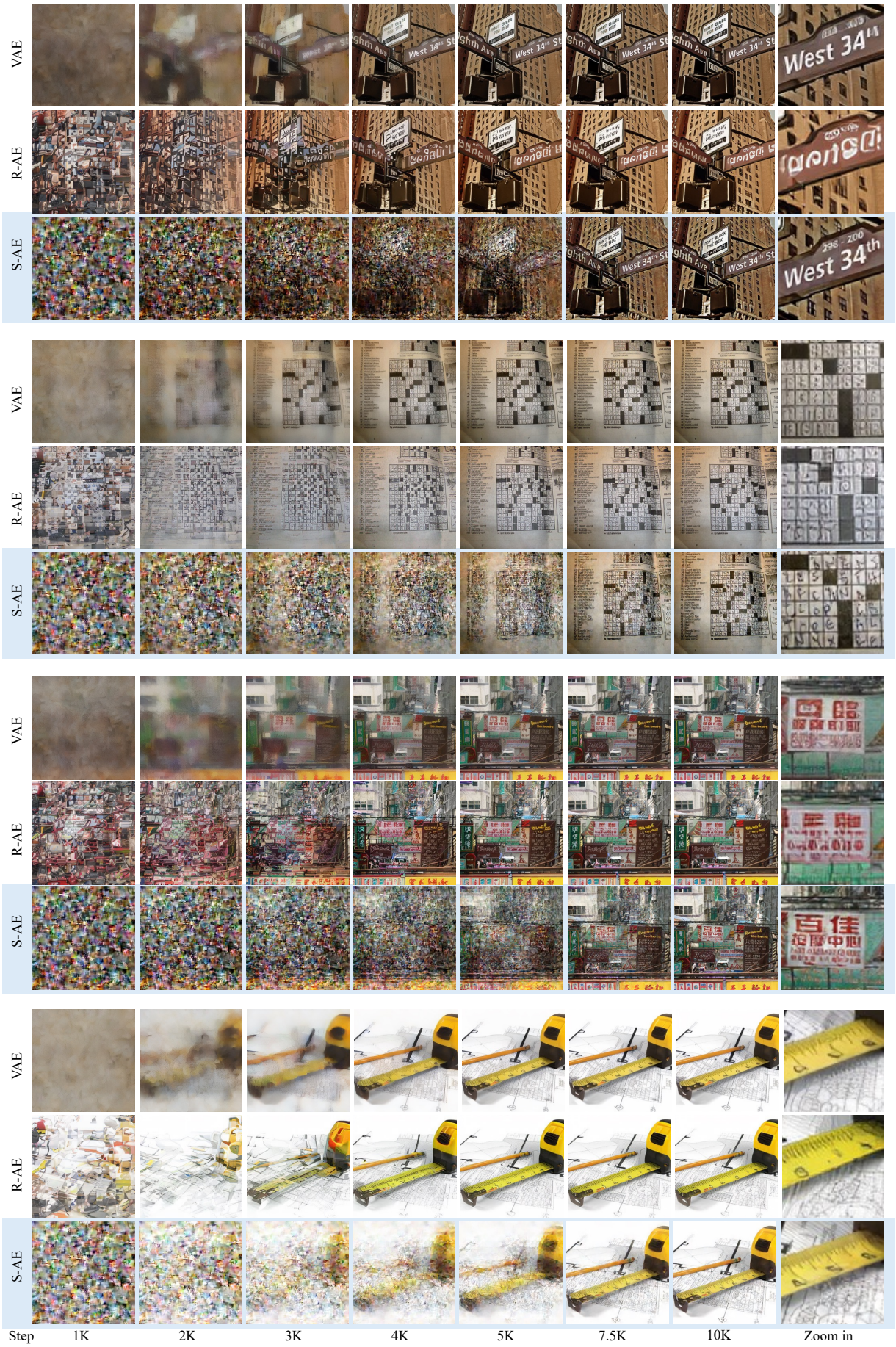

Figure 8. More visualizations for the single image diffusion training with different auto-encoders.

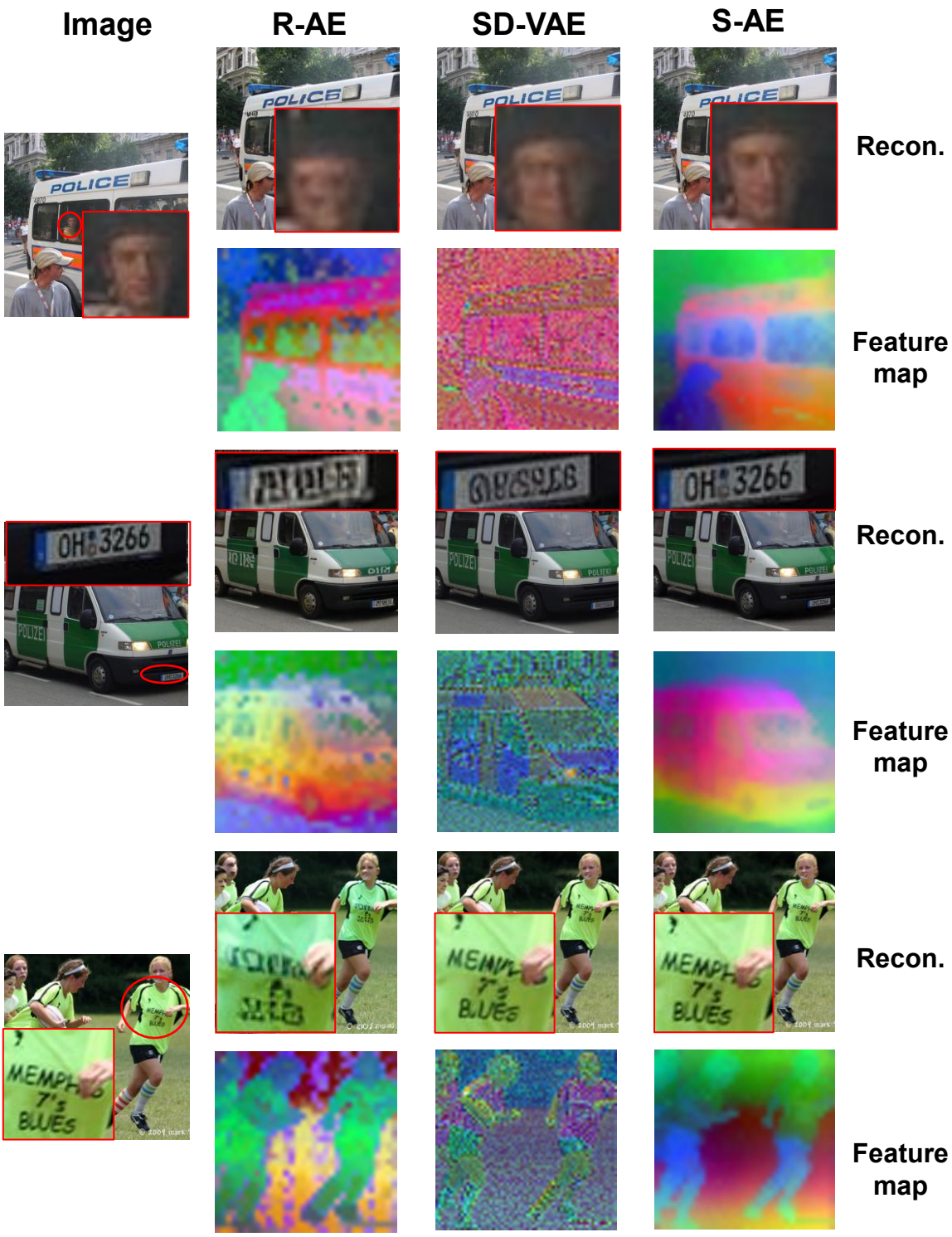

Figure 9. Visualization of reconstruction results of different auto-encoders. We select some extremely hard cases, consisting of texts and small humans. Semantic Auto-encoder achieves strong reconstruction while other methods contain artifacts.

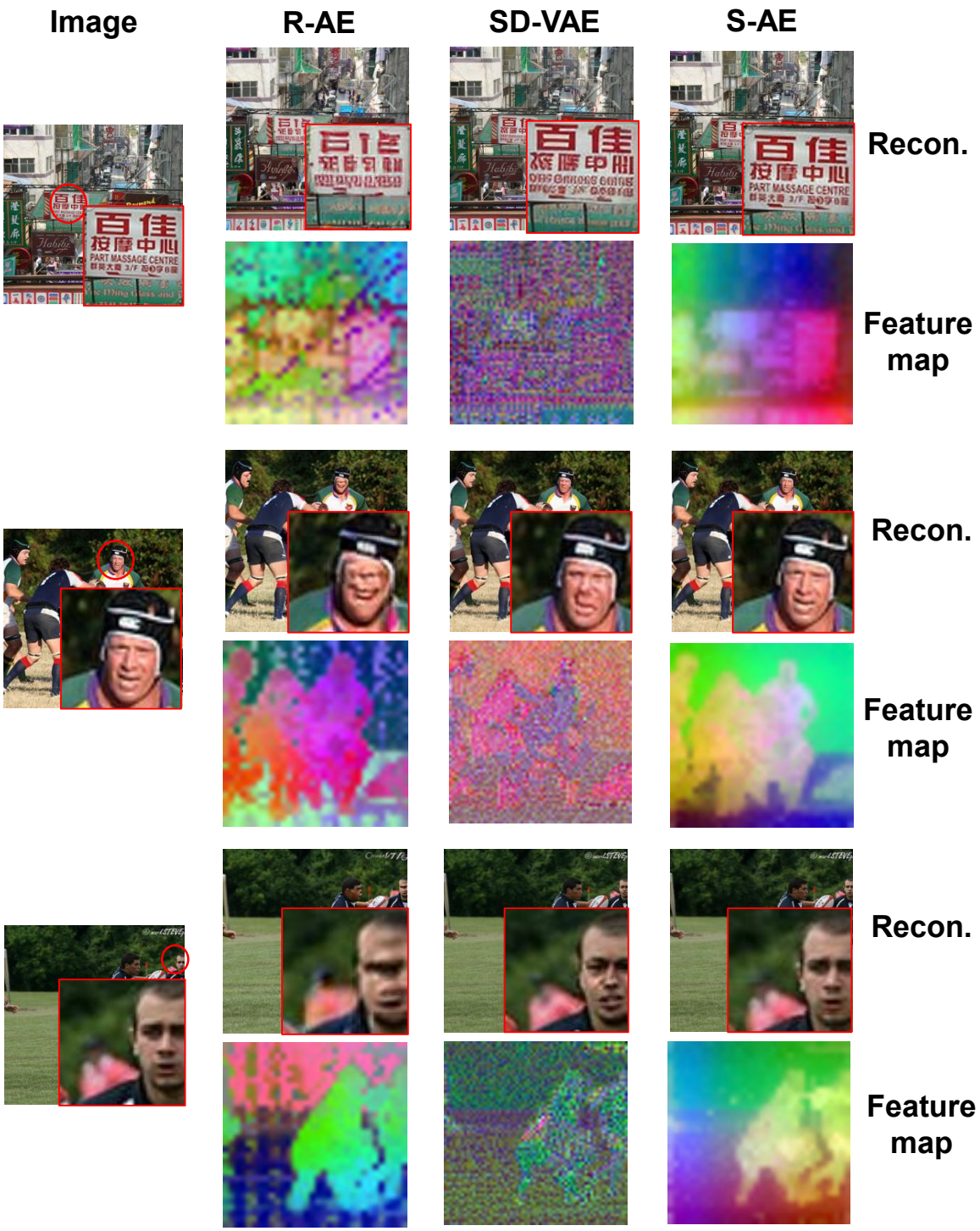

Figure 10. Visualization of reconstruction results of different auto-encoders. We select some extremely hard cases, consisting of texts and small humans. Semantic Auto-encoder achieves strong reconstruction while other methods contain artifacts.