

Rethinking RGB Color Representation for Image Restoration Models

Jaerin Lee¹ JoonKyu Park¹ Sungyong Baik² Kyoung Mu Lee^{1,3}

Abstract

Image restoration models are typically trained with a pixel-wise distance loss defined over the RGB color representation space, which is well known to be a source of blurry and unrealistic textures in the restored images. The reason, we believe, is that the three-channel RGB space is insufficient for supervising the restoration models. To this end, we augment the representation to hold structural information of local neighborhoods at each pixel while keeping the color information and pixel-grainedness unharmed. The result is a new representation space, dubbed augmented RGB (*a*RGB) space. Substituting the underlying representation space for the per-pixel losses facilitates the training of image restoration models, thereby improving the performance without affecting the evaluation phase. Notably, when combined with auxiliary objectives such as adversarial or perceptual losses, our *a*RGB space consistently improves overall metrics by reconstructing both color and local structures, overcoming the conventional perception-distortion trade-off.

1. Introduction

Since VDSR (Kim et al., 2016b) and EDSR (Lim et al., 2017), training an image-to-image deep neural network has been a promising practice for dealing with image restoration tasks. This has led to much interest in the learning objectives for better supervision of image restoration networks. Most of the works have focused on exploiting semantic prior knowledge (Wang et al., 2018b; Zhang et al., 2019; Park et al., 2023), typically in the form of adversarial (Goodfellow et al., 2014) and VGG perceptual losses (Johnson et al., 2016). These loss functions are regarded as add-ons to a per-pixel RGB distance, which has been a unanimous choice regardless of the underlying restoration problem.

¹ASRI, Department of ECE, Seoul National University, Seoul, Korea ²Department of Data Science, Hanyang University, Seoul, Korea ³Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Korea. Correspondence to: Kyoung Mu Lee <kyoungmu@snu.ac.kr>.

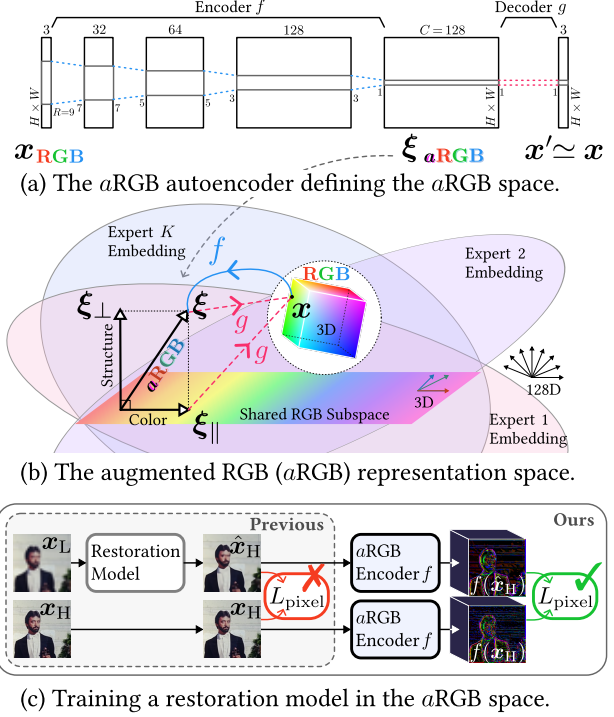


Figure 1. The *a*RGB representation space. Our augmented RGB (*a*RGB) space is designed to replace the RGB space for calculating per-pixel losses to train image restoration models. Unlike conventional per-pixel distances over the RGB space, the same distances defined over our *a*RGB space convey pixel-grained structural information, which is crucial for high-fidelity image reconstruction. Our *a*RGB space also enjoys interpretability, for any embedding ξ is orthogonally decomposable into the color encoding ξ_{\parallel} and the structure encoding parts ξ_{\perp} .

Unfortunately, as highlighted in (Ledig et al., 2017), this per-pixel loss defined over the RGB color representation is the primary cause of blurriness commonly found in the restoration results. We attribute these well-known shortcomings to the lack of local structural information within the three-dimensional feature at each pixel. Since per-pixel distance metrics are applied to each pixel independently, the network is guided towards a mean RGB value estimator for each pixel. Hence, the result appears to be blurry in a global sense. Nevertheless, the per-pixel RGB color difference has been considered a necessary evil due to its *pixel-grained*

supervision. The image restoration model should be trained to preserve the extremely dense, pixel-grained correspondence between the low-quality inputs, the reconstructions, and the ground truth images. This is the reason previous solutions utilizing auxiliary loss functions such as perceptual loss (Johnson et al., 2016) and adversarial loss (Ledig et al., 2017; Zhang et al., 2021) still rely on the per-pixel loss in the RGB space. Perceptual losses using classifier backbones reduce the spatial dimension of the image, and thus fail to provide pixel-perfect structural information even if they are used along with a per-pixel RGB difference loss. Adversarial losses, on the other hand, do not rely on the pairwise distances but only on the distributional shift of each unit patch received by the discriminator network, inevitably leading to inaccurate restoration. Consequently, the common practice of mixing the per-pixel RGB distance with those additional losses cannot provide accurate, pixel-grained supervision of color and local structure.

Instead, we seek our solution by directly changing the per-pixel loss. Focusing on the deficiency of information in the underlying RGB representation space, we propose to *augment* this representation with local structural information. This leads to our *augmented RGB (aRGB) representation space*, serving as a substitute for the traditional RGB space over which per-pixel losses are defined. Our solution relies on a translating autoencoder consisting of a nonlinear, mixture-of-experts (Jacobs et al., 1991) encoder and a linear decoder that translates images between the RGB and the *aRGB* spaces. This architecture ensures almost perfect preservation of the color information (> 60 dB PSNR) while embedding diverse, multimodal distributions of local image structure.

Overall, our contributions can be summarized as follows:

- We present a novel approach to solve the perception-distortion trade-off by training a network over alternative representation space—our *aRGB* space. This involves creating a novel autoencoder to obtain the *aRGB* space.
- With only a few lines of additional code, our method is seamlessly applied to any existing restoration models.
- The pixel-wise loss over our *aRGB* space not only enhances distortion metrics but also consistently improves perceptual metrics when combined with traditional perceptual objectives, all without affecting the testing phase.
- We provide comprehensive analysis on our *aRGB* space for interpretable low-level image representation.

2. Related Work

Pairwise loss in image restoration. Training a deep neural network that translates low-quality images into high-quality estimates has undoubtedly become the standard way

of solving image restoration. While most of the advancements have been made in the network architecture (Kim et al., 2016b; Lim et al., 2017; Nah et al., 2017; Tong et al., 2017; Wang et al., 2018b; Zhang et al., 2018b; Waqas Zamir et al., 2021; Liang et al., 2021; Waqas Zamir et al., 2022; Chen et al., 2022), the importance of loss functions is also widely acknowledged. Since SRCNN (Dong et al., 2016), the first pioneer, employed the MSE loss, the first image restoration models had been trained with the MSE loss (Kim et al., 2016a;b; Nah et al., 2017; Zhang et al., 2017). However, after EDSR (Lim et al., 2017) reported that better convergence can be achieved with L_1 loss, various pairwise loss functions are explored. LapSRN (Lai et al., 2017) rediscovers Charbonnier loss (Bruhn et al., 2005), a type of smooth L_1 loss, for image super-resolution, which is also employed in image deraining (Jiang et al., 2020) with a new edge loss, defined as a Charbonnier loss between Laplacians, which is then employed in general restoration by MPRNet (Waqas Zamir et al., 2021). NAFNet (Chen et al., 2022), on the other hand, uses the PSNR score directly as a loss function. In accordance with these approaches, we attempt a more general approach to design a representation space over which those loss functions can be redefined.

Structural prior of natural images. It is generally recognized that a convolutional neural network, either trained (Simonyan & Zisserman, 2015) or even untrained (Ulyanov et al., 2018), contains structural prior that resonates with the internal structure of natural images. Attempts to exploit this information include the perceptual loss (Johnson et al., 2016; Zhang et al., 2018a; Ding et al., 2020). Adversarial losses (Goodfellow et al., 2014) can also be seen as utilization of structural priors (Wang et al., 2018b; Zhang et al., 2019; Park et al., 2023), as they rely on the gradients calculated from the structural differences between real and restored images. On the other hand, dual domain-based losses (Cui et al., 2023; Cho et al., 2021) seek a way to provide supervision regarding nonlocal structures by calculating difference in the Fourier domain. However, all of those losses are auxiliary, and thus cannot eliminate the strong averaging effect of the pixel-wise loss over the RGB space. Instead, we directly *replace* the RGB space with our *aRGB* space that is both color-preserving, pixel-perfect, and contains pixel-grained structural information. Our approach, therefore, can be orthogonally used with the auxiliary losses, such as perceptual loss, for better performance.

Mixture of Experts. Instead of relying on a single model to handle complex large-scale data, a more effective approach is to distribute the workload among multiple workers—the *experts*. Mixture of Experts (MoE) (Jacobs et al., 1991), now a classic paradigm in machine learning, consists of a routing strategy (Shazeer et al., 2017) and multiple expert models, each of which processes a subset of the training data

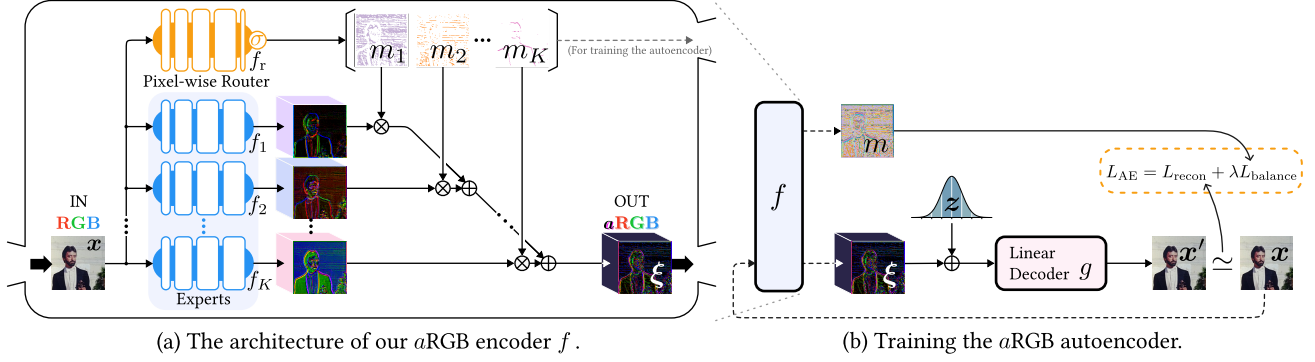


Figure 2. The design and the training of the $aRGB$ autoencoder. Defined with a mixture-of-experts encoder and a linear decoder, the $aRGB$ autoencoder translates an RGB image into $aRGB$ and back. This design allows us to imbue gradient-based supervision from any per-pixel distance loss with rich pixel-grained structural information, while preserving color information. After training the $aRGB$ encoder as an autoencoder fashion, it remains frozen during the training of restoration models.

partitioned and given by the router. Recent studies (Zhou et al., 2022; Fedus et al., 2022) have shown the advantages of MoE in deep learning. Two main challenges arise when working with MoE in deep learning: limited computational resources and training stability (Fedus et al., 2021; He et al., 2021). In response to these challenges, we employ a balancing loss (Fedus et al., 2022) to ensure the stable training of expert networks and incorporate MoE exclusively during the training phase, leaving the testing phase unaffected.

3. Lifting the RGB to $aRGB$

3.1. The $aRGB$ Autoencoder

Our primary goal is to design a representation space for low-level vision tasks in order to facilitate training of image restoration networks. Designing a representation space is achieved by defining the encoder and the decoder to translate images back and forth between the RGB space and the target space. We can split our goal into two parts: (1) the feature at each *pixel* in our space is required to encode its neighboring structure, and (2) the integrity of the color information should be preserved. To fulfill the first requirement, our encoder is a size-preserving ConvNet with nonlinearities to capture the structure among adjacent pixels. For the latter, we employ a per-pixel linear decoder, *i.e.*, a 1×1 convolution, to strongly constrain the embedding of a pixel to include its RGB color information. The overall architecture is illustrated in Figure 2.

We start from an RGB image $x \in \mathbb{R}^{3 \times H \times W}$. Our convolutional encoder f transforms image x into a feature $\xi \in \mathbb{R}^{C \times H \times W}$ of a new representation space. Unlike typical undercomplete autoencoders, which remove information from their inputs, we aim to *add more* information regarding local structures for each pixel $[\xi]_{ij}$ at coordinate (i, j) . Therefore, C must be greater than 3, and the receptive field size R should be greater than unity. Our decoder

$g : \xi \mapsto x$ is effectively a single 1×1 convolution. That is, we can express $g([\xi]_{ij})$ as a per-pixel linear operation: $g([\xi]_{ij}) = A[\xi]_{ij} + b$, where $A \in \mathbb{R}^{3 \times C}$ and $b \in \mathbb{R}^3$. This ensures that each feature $[\xi]_{ij}$ in our representation space extends the color information presented in $[x]_{ij}$, hence the name of our new representation, *augmented RGB*. Additionally, using a linear decoder g offers an interpretability: we can regard the nullspace of A , *i.e.*, the set of undecoded information, as a reservoir of any extra information captured by the encoder f other than local colors.

What is crucial at this juncture is to define our $aRGB$ space to effectively capture the highly varying, complex mixture of information from the color and the neighboring structure at each pixel. To this end, we employ a mixture-of-experts (MoE) architecture (Jacobs et al., 1991; Shazeer et al., 2017; Fedus et al., 2022) within our encoder. We choose this design based on our conjecture that the topology of the space of image patches is disconnected and, therefore, can be more efficiently modeled with a MoE architecture than a single ConvNet. For the set of the smallest images, *i.e.*, a set of pixels, we can argue that their domain is a connected set under absence of quantization, since a pixel can take arbitrary color value. This does not hold in general if the size of the patches becomes large enough to contain semantic structures. In fact, we cannot interpolate between two images of semantically distinct objects *in the natural image domain*, *e.g.*, there is no such thing as a half-cat half-airplane object *in nature*. This implies that topological disconnectedness emerges from the domain of patches as the size of its patches increases. Since a single-module encoder is a continuous function, learning a mapping over a disconnected set may require deeper architecture with a lot of parameters. An MoE encoder, per contra, can model a discontinuous map more effectively through its discrete routing strategy between small, specialized experts. We will revisit our conjecture in Section 5.

Table 1. Results on real image denoising.

Model	Objective	SIDD	
		PSNR \uparrow	SSIM \uparrow
NAFNet-width32	L_{PSNR}	39.9672	0.9599
NAFNet-width32	$L_{\text{PSNR}, a\text{RGB}}$	39.9864	0.9601
NAFNet-width32	$L_{1, a\text{RGB}}$	40.0106	0.9602
NAFNet-width64	L_{PSNR}	40.3045	0.9614
NAFNet-width64	$L_{1, a\text{RGB}}$	40.3364	0.9620

Table 2. Results on motion blur deblurring.

Model	Objective	GoPro		HIDE	
		PSNR \uparrow	SSIM \uparrow	PSNR \uparrow	SSIM \uparrow
DeepDeblur	L_1	27.8333	0.8707	25.7313	0.8400
DeepDeblur	$L_{1, a\text{RGB}}$	27.8409	0.8710	25.8146	0.8424
MPRNet	$L_{\text{Char}} + 0.05L_{\text{Edge}}$	32.6581	0.9589	30.9622	0.9394
MPRNet	$L_{\text{Char}, a\text{RGB}} + 0.05L_{\text{Edge}}$	32.7118	0.9594	31.0248	0.9398
MPRNet-TLC	$L_{\text{Char}} + 0.05L_{\text{Edge}}$	33.3137	0.9637	31.1868	0.9418
MPRNet-TLC	$L_{\text{Char}, a\text{RGB}} + 0.05L_{\text{Edge}}$	33.3886	0.9642	31.2082	0.9421

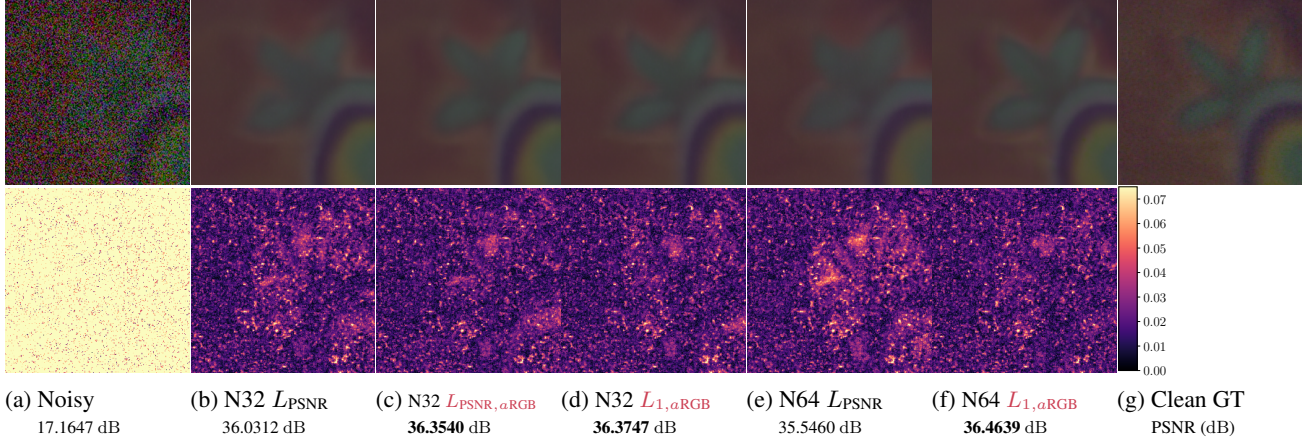


Figure 3. Qualitative comparison of real image denoising models trained with different loss functions. Each column corresponds to each row in Table 1. N32 corresponds to NAFNet-width32 and N64 corresponds to NAFNet-width64. The bottom row shows the maximum absolute difference in color with a range of $[0, 1]$.

In practice, an RGB image $\mathbf{x} \in \mathbb{R}^{3 \times H \times W}$ is fed into the router f_r as well as K encoders f_1, \dots, f_K . The router f_r is a five-layer ConvNet classifier with a softmax at the end. The output of the router $\mathbf{y} = f_r(\mathbf{x}) \in [0, 1]^{K \times H \times W}$ partitions each pixel of \mathbf{x} into K different bins with top-1 policy. This is equivalent to generating mutually exclusive and jointly exhaustive K masks m_1, \dots, m_K of size $H \times W$. The features $\xi_1 = f_1(\mathbf{x}), \dots, \xi_K = f_K(\mathbf{x})$ are aggregated into a single feature $\xi = f(\mathbf{x}) \in \mathbb{R}^{C \times H \times W}$:

$$\xi = \sum_{k=1}^K m_k \odot f_k(\mathbf{x}) \quad (1)$$

$$= \sum_{k=1}^K \mathbb{1}(k = \arg \max_{k'} [f_r(\mathbf{x})]_{k'}) \odot f_k(\mathbf{x}), \quad (2)$$

where \odot is an element-wise multiplication and $\mathbb{1}$ is the indicator function. We ensure that $(g \circ f)(\mathbf{x}) = \mathbf{x}' \simeq \mathbf{x}$ by training f and g jointly in an autoencoder scheme. After the training, the decoder g is discarded and the encoder f is used to generate $a\text{RGB}$ representations from RGB images.

3.2. Training the Autoencoder

Our objective is to ensure that the $a\text{RGB}$ encoder f effectively learns accurate low-level features from clean (or

sharp) and natural images. To achieve this goal, we make use of a dataset D , consisting of clean image patches. With this dataset, the $a\text{RGB}$ autoencoder is trained to minimize the L_1 distance between a patch $\mathbf{x} \in D$ and its reconstruction $(g \circ f)(\mathbf{x})$. In addition, likewise in Switch Transformer (Fedus et al., 2022), a load-balancing loss L_{balance} is applied to encourage the router f_r to distribute pixels evenly across the K experts during training:

$$L_{\text{balance}} = K^2 \sum_{i=1}^H \sum_{j=1}^W \left[\max_k [f_r(\mathbf{x})]_k \right]_{ij}, \quad (3)$$

which is minimized when the distribution is uniform with the value of unity. Furthermore, to increase the sensitivity of the encoder f , we simply add an isotropic Gaussian noise at the output of the encoder only during the training of the $a\text{RGB}$ autoencoder. Therefore, the reconstruction loss for the autoencoder is:

$$L_{\text{recon}} = \|g(f(\mathbf{x}) + \mathbf{z}) - \mathbf{x}\|_1, \quad (4)$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Although the decoder is only informed with three color channels of each pixel during the training, we observe that the latent space does not degenerate into trivial solutions. See Appendix A for more information. Overall, the training loss for the $a\text{RGB}$ autoencoder is:

$$L_{\text{AE}} = L_{\text{recon}} + \lambda L_{\text{balance}}. \quad (5)$$

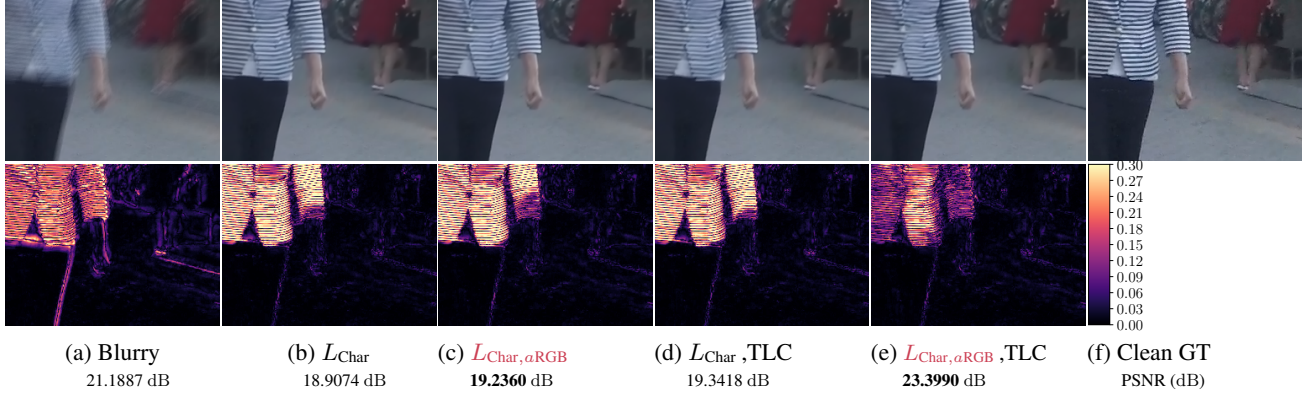


Figure 4. Qualitative comparison of motion blur deblurring models trained with different loss functions. Each column corresponds to each row in Table 2. The bottom row is the maximum absolute RGB difference.

In practice, we choose $\lambda = 0.01$. The final autoencoder achieves 67.21 dB in the reconstruction of the Set5 benchmark (Bevilacqua et al., 2012). In other words, the average RGB color difference is below a tenth of the quantization step. Henceforth, we will consider our $aRGB$ autoencoder lossless in the analysis in Section 5. More implementation details are provided in Appendix B.

3.3. Integration into Existing Restoration Frameworks

Figure 1c illustrates the overall pipeline for integrating our $aRGB$ into the existing restoration framework. The interface is similar to the famous perceptual loss (Johnson et al., 2016), but unlike perceptual loss, the resulting $aRGB$ losses are pixel-grained and color-preserving, and completely replaces the RGB per-pixel loss. Training image restoration models with respect to the $aRGB$ space only requires a few lines of code modification. Typically, an image restoration model is trained to minimize a per-pixel distance L_{pixel} , optionally combined with some auxiliary losses L_{aux} , such as a perceptual loss (Johnson et al., 2016) or an adversarial loss (Ledig et al., 2017). The overall loss function can be represented as:

$$L_{\text{RGB}}(\mathbf{x}, \hat{\mathbf{x}}) = L_{\text{pixel}}(\mathbf{x}, \hat{\mathbf{x}}) + L_{\text{aux}}(\mathbf{x}, \hat{\mathbf{x}}), \quad (6)$$

where \mathbf{x} is the ground-truth image and $\hat{\mathbf{x}}$ is the restoration result. To train the model in the $aRGB$ space, we are only required to modify the input to the per-pixel loss L_{pixel} . That is, the per-pixel distances are now computed between the images in the $aRGB$ space, namely, $f(\mathbf{x})$ and $f(\hat{\mathbf{x}})$.

$$L_{a\text{RGB}}(\mathbf{x}, \hat{\mathbf{x}}) = L_{\text{pixel}}(f(\mathbf{x}), f(\hat{\mathbf{x}})) + L_{\text{aux}}(\mathbf{x}, \hat{\mathbf{x}}). \quad (7)$$

Since what we present is not a specific loss function but the underlying *space* itself, our method can be seamlessly integrated with any existing restoration framework regardless of the type of per-pixel loss it uses.

4. Experiments

4.1. Comparison with Previous Methods

To demonstrate our proposed $aRGB$ space, we conduct experiments across various image restoration tasks, including denoising, deblurring, and super-resolution. We also show the versatility of our representation space-based approach by applying diverse types of pixel-wise losses over the $aRGB$ space, *e.g.*, L_1 , PSNR and Charbonnier loss. Furthermore, we empirically show the benefit of combining our per-pixel $aRGB$ loss function with conventional perceptual quality-oriented objectives, including VGG and adversarial losses.

Distortion-oriented: Tables 1 and 2 demonstrate the efficacy of our $aRGB$ representation across various pixel-wise loss objectives. In all comparisons, we substitute the RGB space with our $aRGB$ space in the target domain of pixel-wise loss, as illustrated in Equation 4. Specifically, in Table 1, using NAFNet (Chen et al., 2022) with L_2 loss and L_{PSNR} , our $aRGB$ representation yields better PSNR and SSIM scores than the model trained directly with the original PSNR metric. Here, a per-pixel PSNR loss L_{PSNR} represents a mathematically equivalent form of the L_2 loss. The table indicates that our $aRGB$ representation preserves RGB information and incorporates additional local structural information, enhancing denoising supervision. Furthermore, experiments with different metrics highlight the impact of changing the representation space on training dynamics, offering unexpected benefits, as seen with NAFNets trained for the L_1 metric in our $aRGB$ space (last rows of Table 3).

Moreover, in Table 2, employing a Charbonnier loss (Bruhn et al., 2005), we observe significant improvements in training MPRNet (Waqas Zamir et al., 2021) for motion blur deblurring when the loss is applied in our $aRGB$ space, compared to the RGB space. Furthermore, Table 2 and Figure 4 demonstrate significant improvements, orthogonal to existing enhancement techniques such as test-time local converter (TLC) (Chu et al., 2022). Our experiments indi-

Table 3. **Quantitative results on training $4\times$ super-resolution ESRGAN in the a RGB space.** In our methods using a RGB representation, we modify only the L_1 loss by exchanging it with the $L_{1,aRGB}$ loss. All the other training hyperparameters are left untouched. Better scores in each block are shown in **boldface** text. More results are in Table 10 in Appendix C.

Objective	DIV2K-Val					Urban100				
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	NIQE \downarrow	FID \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	NIQE \downarrow	FID \downarrow
$0.01L_1 + L_{VGG} + 0.005L_{Adv}^\dagger$	26.627	0.7033	0.1154	3.0913	13.557	22.776	0.7033	0.1232	4.2067	20.616
$0.01L_{1,aRGB} + L_{VGG} + 0.005L_{Adv}$	26.845	0.7500	0.1110	2.9615	12.799	23.270	0.7196	0.1183	3.8982	17.739

† The official ESRGAN model (Wang et al., 2018b).

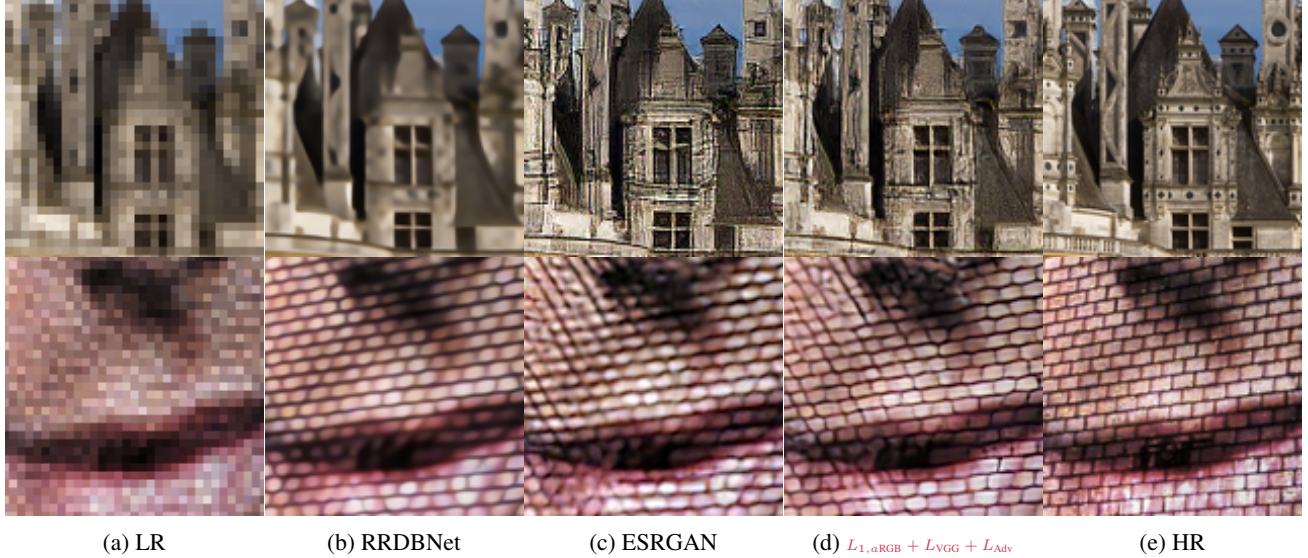


Figure 5. **Qualitative comparison of ESRGAN models trained with different loss functions.** Each column corresponds to each row in Table 3. The loss weights are omitted for brevity, ESRGAN corresponds to the $0.01L_1 + L_{VGG} + 0.005L_{Adv}$ in Table 3.

cate that our a RGB representation enhances the training of image restoration models across various tasks, architectures, and loss functions, synergizing with other techniques like edge loss and test-time local converter.

Perceptual quality-oriented: To demonstrate the strong synergy of our proposed pixel-wise loss with perceptual losses (e.g., VGG loss (Johnson et al., 2016) and adversarial loss (Goodfellow et al., 2014)), we compare our approach on photo-realistic super-resolution. To elaborate, Table 3 shows a comparison using the ESRGAN (Wang et al., 2018b) as a baseline. The consistent improvement in both distortion metrics (e.g., PSNR and SSIM) and perceptual metrics (e.g., LPIPS, NIQE, and FID), deviating from the conventional trade-off between them, underscores the effectiveness of our modified $L_{1,aRGB}$ objective in stabilizing the adversarial training of a super-resolution model. This indicates that the local structural information within our a RGB representation complements supervision from a pre-trained classifier, resulting in superior performance in both distortion-based and perceptual metrics compared to the original ESRGAN. The improvements in PSNR and SSIM scores align with our design philosophy, emphasizing that RGB colors are included as a subspace in our a RGB representation.

Table 4. **The number of parameters of the losses.**

Objective	# params (M)
L_{VGG}	20.2
L_{Adv}	80.2
L_{aRGB}	5.3

Computational overhead: Since our method incorporates an additional a RGB encoder model f during the training of the image restoration model, there is an associated increase in computational overhead. Table 4 provides a comparison of the computational burden with other conventional objectives that utilize additional modules during the training process. Notably, our a RGB encoder f requires fewer parameters compared to the alternatives, demonstrating its efficiency. Specifically, for the adversarial loss L_{Adv} , we quantify the number of parameters based on the discriminator of ESRGAN (Wang et al., 2018b). In comparison to the super-resolution model itself, our overhead is not substantial, accounting for only 9% of the entire model. Also, please note that our method exclusively applies to the training phase of the restoration model, incurring no additional computational overhead during test time.

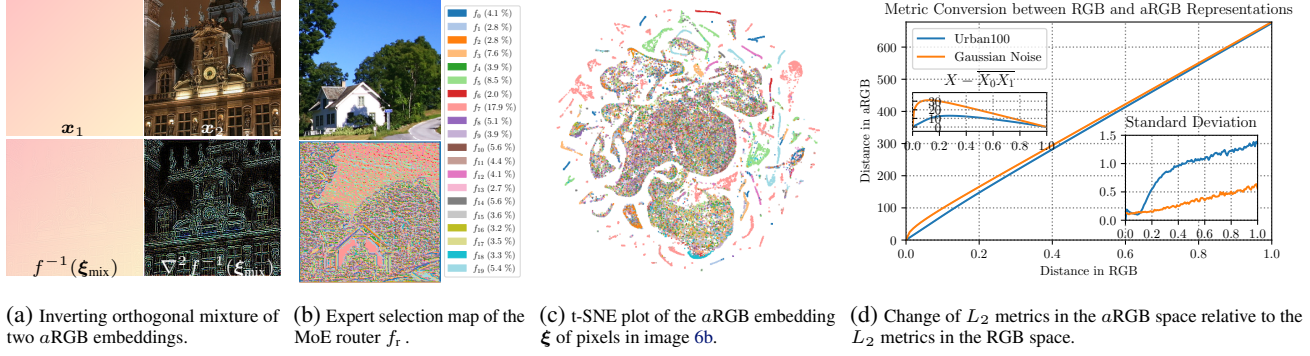


Figure 6. Understanding the learned aRGB representation. Figure 6a show a visual example of aRGB embedding inversion. Figure 6b and 6c reveal clear evidence that the experts of our aRGB encoder f are specialized for a particular type of input structures, and that even the embedding vectors within a single patch are clustered in a complicated manner, justifying our usage of MoE architecture. Figure 6d shows how the distance metric changes in the aRGB space relative to the distance in the RGB space. Mean distances and their standard deviations are measured by MSE losses between an image and the same image with 100 AWGNs with the same standard deviation. Note that the aRGB space slightly exaggerates the distance more outside natural image domain, e.g., Gaussian noise, and the metric’s variance is negligibly small. More examples are in Appendix E.

5. Discussion

5.1. Nullspace of the Decoder

In addition to the design simplicity, our pixel-wise *linear* decoder enjoys an additional benefit: decomposability. Since our autoencoder is almost lossless, we will consider that the RGB $x \in \mathbb{R}^3$ and the aRGB $\xi = f(x) \in \mathbb{R}^C$ representations of any given image equivalent. That is, $x' = g(\xi) = A\xi + b = x$. As a result of the linearity of our decoder g , the aRGB representation ξ can be decomposed into the sum of two orthogonal components:

$$\xi = \xi_{\parallel} + \xi_{\perp}, \quad (8)$$

$$\text{s.t. } f_{\parallel}(x) := \xi_{\parallel} = A^{\dagger}A\xi \quad (9)$$

$$\text{and } f_{\perp}(x) := \xi_{\perp} = (I - A^{\dagger}A)\xi, \quad (10)$$

where A^{\dagger} is the Moore-Penrose pseudoinverse of A . The parallel component ξ_{\parallel} of the aRGB representation lies in the three-dimensional subspace of \mathbb{R}^C that is projected onto the RGB colors by the decoder g , i.e., $A\xi_{\parallel} = AA^{\dagger}A\xi = A\xi$. The remaining perpendicular part ξ_{\perp} can be regarded as the information the aRGB space encodes in addition to the RGB colors. The contribution of the two components can be visualized by inverting the encoder f with respect to a mixed embedding:

$$f^{-1}(\xi_{\text{mix}}) = \arg \min_z \|f(z) - \xi_{\text{mix}}\|_2^2, \quad (11)$$

$$\text{s.t. } \xi_{\text{mix}} = \xi_{1\parallel} + \xi_{2\perp} = A^{\dagger}A\xi_1 + (I - A^{\dagger}A)\xi_2. \quad (12)$$

We use a SGD optimizer with a learning rate of 0.1 for 50 iterations. As shown in Figure 6a and Appendix E, the inversion of the mixed embedding inherits color information from the parallel embedding $\xi_{1\parallel}$, while the perpendicular part $\xi_{2\perp}$ contributes to the high-frequency edge information.

5.2. Expert Specialization and Learned Structures

Figure 6b visualizes how individual pixels of a natural image are distributed to $K = 20$ experts. Unlike in semantic segmentation, where segmentation maps are chunked into large blocks of semantically correlated pixels, our pixel-wise router f_r generates fine-grained distributions of pixels. That is, multiple experts are jointly involved in encoding the same texture, such as the blue sky and the leafy trees. Another salient feature we can observe in the figure is that edges of different orientations are dealt with by different experts, implying their specialization. Visualizing the aRGB embedding space using t-SNE (van der Maaten & Hinton, 2008) provides us with additional insights into the topology of the space. Figure 6c reveals that the aRGB embeddings cluster into multiple disconnected groups in two different types: *common* groups where multiple experts are involved in the encoding process and *specialized* groups where a single expert is exclusively allocated for the embeddings. These observations align well with our initial design principles in Section 3.1, where the feature embeddings occupy a highly complicated, disconnected set, and an MoE architecture effectively deals with this structure by specializing each expert to a subset of the embedding space.

5.3. The aRGB Metric Space and Produced Gradients

The main purpose of our the aRGB space is to provide alternative supervision to the existing image restoration framework. This supervision is realized with a metric defined over the space and its gradients generated from pairs of images. To this end, we first visualize the correlation between L_2 distances defined in the RGB and our aRGB spaces in Figure 6d. We plotted additional figure with title $X - X_0X_1$ to show the deviation of the graph over the straight line,

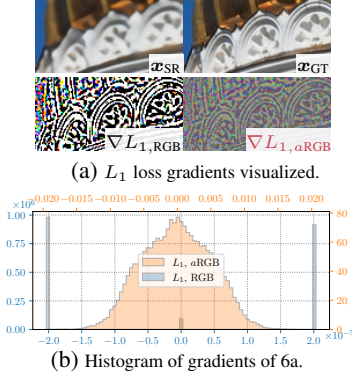


Figure 7. Gradients from $L_{1,RGB}$ and $L_{1,aRGB}$ losses.

showing clear convexity of the graph. This implies that the metrics within $aRGB$ spaces are inflated when the given two images are similar. Figure 7 shows the gradients from two per-pixel L_1 losses between a restored image and its high-quality counterpart defined over both spaces. Unlike RGB L_1 loss which exhibits a highly off-centered, discrete distribution, the $L_{1,aRGB}$ loss shows smooth and centered distribution of gradients. We believe that this allows for the stable training of the image restoration models despite its huge scale of the generated gradients from the $L_{1,aRGB}$ loss, which is more than a hundredfold as shown in the x axis of Figure 7b. In the RGB domain, the same scale of gradient is achievable only through increasing the learning rate, which leads to destabilization of the training. Overall, the analyses show how our $aRGB$ encoder helps the training of image restoration models.

5.4. Ablation Study

Lastly, we provide ablation studies to determine the best hyperparameters for our $aRGB$ autoencoder. We compare the models by the results of training an RRDBNet (Wang et al., 2018b) only on DIV2K dataset. The results are summarized in Table 5. More information is elaborated in Appendix B.

Number of experts: As our encoder employs a Mixture-of-Experts (MoE) architecture, the overall performance of forming the $aRGB$ space can be influenced by the number of experts. To find the optimal number of experts, we conduct an ablation study in Table 5. The first four rows of Table 5 show the effect of the number of experts of the $aRGB$ encoder f on its supervision quality. While applying more experts yields improved results by letting encoder to jointly learn input textures in multiple spaces, as shown in the t-SNE visualization in Figure 6c, using 30 experts shows slightly worse performance compared to 20 experts. This is attributed to the larger number of experts causing input textures to be learned in a too-specified distribution, hindering the encoder’s ability to learn the distribution of

Table 5. Ablation studies on the $aRGB$ autoencoder. RRDBNets (Wang et al., 2018b) are trained with DIV2K (Agustsson & Timofte, 2017) for 300k iterations for $4\times$ SISR tasks with only the L_1 loss between the $aRGB$ embeddings.

RRDBNet in $4\times$ SISR				Set14		Urban100		DIV2K-Val	
# experts	Routing	$aRGB$ train set	Reg. noise	PSNR \uparrow	SSIM \uparrow	PSNR \uparrow	SSIM \uparrow	PSNR \uparrow	SSIM \uparrow
1	-	DIV2K	✓	26.87	0.7467	24.75	0.7735	29.08	0.8222
5	MoE	DIV2K	✓	26.87	0.7477	24.83	0.7745	29.12	0.8231
10	MoE	DIV2K	✓	26.89	0.7474	24.84	0.7750	29.11	0.8231
20	MoE	DIV2K	✓	26.91	0.7471	24.87	0.7745	29.14	0.8227
30	MoE	DIV2K	✓	26.89	0.7476	24.84	0.7750	29.11	0.8231
20	MoE	GoPro	✓	26.89	0.7459	24.83	0.7728	29.12	0.8220
20	MoE	SIDD	✓	26.86	0.7420	24.80	0.7691	29.08	0.8186
20	MoE	Noise	✓	26.65	0.7461	24.64	0.7729	28.87	0.8212
20	MoE	DIV2K	✗	26.91	0.7469	24.85	0.7722	29.13	0.8223

input samples properly. From the results, we fix the number of experts to 20 throughout our experiments.

Dataset dependence: As the second part of Table 5 presents, the training data for the $aRGB$ autoencoder decides the quality of supervision the model gives. This implies that our $aRGB$ autoencoder utilizes structural priors of its training data. Appendix 8 provides additional theoretical and empirical evidence that our $aRGB$ autoencoder learns image structures to reconstruct given images.

Regularizers: In the last row of Table 5, we observe that the regularizing noise z added at the end of the encoder during training helps the $aRGB$ encoder to produce stronger supervision for image restoration models. In practice, we observe more than tenfold reduction in the scale of produced gradients when the $aRGB$ autoencoder trained without the regularizing noise is applied. This correlates to our discussion in Section 5.3, that our $aRGB$ encoder helps training image restoration models by stably increasing the scale of gradients.

6. Conclusion

We hypothesized that blurry restoration due to per-pixel losses can be alleviated if the underlying representation space contains enough structural information for each pixel. To prove this, we proposed augmented RGB ($aRGB$) representation space to supply pixel-accurate local structural information and to preserve color information while training the restoration models. Improved performance could be observed across diverse set of restoration tasks such as denoising, deblurring, and perceptual super-resolution by only substituting the underlying representation space for the per-pixel losses to our $aRGB$ space. This suggests that the RGB color space is not the optimal representation space for low-level computer vision tasks. We hope our work spurs more interests in the optimal representation spaces.

Broader Impact

The goal of this paper is to advance the image restoration through representation learning. Better image restoration techniques imply both beneficial and potentially harmful societal consequences. For example, restoring old, compressed, or degraded images to high-quality images not only can help archiving old historical records or personal memories but also can strengthen surveillance systems.

References

- Abdelhamed, A., Lin, S., and Brown, M. S. A high-quality denoising dataset for smartphone cameras. In *CVPR*, 2018.
- Agustsson, E. and Timofte, R. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshop*, 2017.
- Bevilacqua, M., Aline, R., Christine, G., and line Alberi, M. M. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, 2012.
- Bruhn, A., Weickert, J., and Schnörr, C. Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61:211–231, 2005. URL <https://api.semanticscholar.org/CorpusID:15374825>.
- Chen, L., Chu, X., Zhang, X., and Sun, J. Simple baselines for image restoration. In *ECCV*, 2022.
- Cho, S.-J., Ji, S.-W., Hong, J.-P., Jung, S.-W., and Ko, S.-J. Rethinking coarse-to-fine approach in single image deblurring. In *ICCV*, 2021.
- Chu, X., Chen, L., Chen, C., and Lu, X. Improving Image Restoration by Revisiting Global Information Aggregation. In *ECCV*, 2022.
- Cui, Y., Tao, Y., Bing, Z., Ren, W., Gao, X., Cao, X., Huang, K., and Knoll, A. Selective frequency network for image restoration. In *ICLR*, 2023.
- Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:2567–2581, 2020. URL <https://api.semanticscholar.org/CorpusID:215785896>.
- Dong, C., Loy, C. C., He, K., and Tang, X. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, February 2016.
- Fedus, W., Zoph, B., and Shazeer, N. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv e-prints*, art. arXiv:2101.03961, January 2021. doi: 10.48550/arXiv.2101.03961.
- Fedus, W., Zoph, B., and Shazeer, N. Switch Transformers: scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(1):5232—5270, January 2022. ISSN 1532-4435.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NIPS*, 2014.
- Gu, S., Lugmayr, A., Danelljan, M., Fritsche, M., Lamour, J., and Timofte, R. DIV8K: DIVERse 8K resolution image dataset. In *ICCV Workshops*, 2019.
- He, J., Qiu, J., Zeng, A., Yang, Z., Zhai, J., and Tang, J. Fastmoe: A fast mixture-of-expert training system. *arXiv preprint arXiv:2103.13262*, 2021.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NIPS*, 2017.
- Huang, J.-B., Singh, A., and Ahuja, N. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.
- Jiang, K., Wang, Z., Yi, P., Chen, C., Huang, B., Luo, Y., Ma, J., and Jiang, J. Multi-Scale Progressive Fusion Network for Single Image Deraining. In *CVPR*, 2020.
- Johnson, J., Alahi, A., and Li, F.-F. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.
- Kim, J., Lee, J. K., and Lee, K. M. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016a.
- Kim, J., Lee, J. K., and Lee, K. M. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016b.
- Kingma, D. P. and Ba, J. L. Adam: a method for stochastic optimization. In *ICLR*, 2015.
- Lai, W.-S., Huang, J.-B., Ahuja, N., and Yang, M.-H. Deep Laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017.

- Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.
- Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. SwinIR: image restoration using Swin Transformer. In *ICCV Workshops*, 2021.
- Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M. Enhanced deep residual networks for single image super-resolution. In *CVPR Workshop*, 2017.
- Loshchilov, I. and Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In *ICLR*, 2017.
- Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001.
- Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., and Aizawa, K. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, oct 2017. ISSN 1380-7501. doi: 10.1007/s11042-016-4020-z. URL <https://doi.org/10.1007/s11042-016-4020-z>.
- Mittal, A., Soundararajan, R., and Bovik, A. C. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.
- Nah, S., Kim, T. H., and Lee, K. M. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017.
- Park, J., Son, S., and Lee, K. M. Content-aware local gan for photo-realistic super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10585–10594, 2023.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: an imperative style, high-performance deep learning library. In *NeurIPS*, 2019.
- Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F. A., Bengio, Y., and Courville, A. On the Spectral Bias of Neural Networks. In *ICML*, 2019.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.
- Seitzer, M. pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid>, August 2020. Version 0.3.0.
- Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.
- Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.
- Shen, Z., Wang, W., Lu, X., Shen, J., Ling, H., Xu, T., and Shao, L. Human-aware motion deblurring. In *ICCV*, 2019.
- Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- Timofte, R., Agustsson, E., Gool, L. V., Yang, M.-H., and Zhang, L. NTIRE 2017 challenge on single image super-resolution: Methods and results. In *CVPR Workshop*, 2017.
- Tong, T., Li, G., Liu, X., and Gao, Q. Image super-resolution using dense skip connections. In *ICCV*, 2017.
- Ulyanov, D., Vedaldi, A., and Lempitsky, V. Deep image prior. In *CVPR*, 2018.
- van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579 – 2605, November 2008.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NIPS*, 2017.
- Wang, X., Yu, K., Dong, C., and Change Loy, C. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018a.
- Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Loy, C. C. ESRGAN: enhanced super-resolution generative adversarial networks. In *ECCV Workshop*, 2018b.
- Waqas Zamir, S., Arora, A., Khan, S., Hayat, M., Shahbaz Khan, F., Yang, M.-H., and Shao, L. Multi-stage progressive image restoration. *arXiv preprints*, February 2021.
- Waqas Zamir, S., Arora, A., Khan, S., Hayat, M., Shahbaz Khan, F., and Yang, M.-H. Restormer: Efficient Transformer for High-Resolution Image Restoration. In *CVPR*, 2022.

- Zeyde, R., Elad, M., and Protter, M. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, pp. 711–730, 2010.
- Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.
- Zhang, K., Liang, J., Van Gool, L., and Timofte, R. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 2021.
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018a.
- Zhang, W., Liu, Y., Dong, C., and Qiao, Y. RankSRGAN: generative adversarial networks with ranker for image super-resolution. In *ICCV*, 2019.
- Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018b.
- Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A., Chen, Z., Le, Q., and Laudon, J. Mixture-of-experts with expert choice routing. *arXiv preprint arXiv:2202.09368*, 2022.

A. Theoretical study on the structure embedded in a RGB space

The goal of this section is to provide a simple theoretical analysis on how the structure is learned in our a RGB encoder f . This section comprises of two parts. The first part shows that our a RGB encoder is a piecewise linear function over the connected neighborhood regions in the RGB pixel domain. From this, we can equivalently transform our a RGB autoencoder to a coordinate-wise function, which is useful in the upcoming analyses. In the second part, we will show that an autoencoder with a neural network encoder and a linear decoder like ours does not learn structural priors if it is perfectly lossless. In other words, we claim that imperfect, yet almost perfect (> 60 dB PSNR), reconstruction capability of our a RGB autoencoder helps its encoder to learn local structures within natural images. Based on the mathematical analysis, we provide another method to measure how much image structure is captured in the a RGB autoencoder.

A.1. The a RGB encoder is a piecewise linear function over the local neighborhood structures

It is known that a multi-layer perceptron (MLP) with a continuous piecewise linear activation is also a continuous piecewise linear (CPWL) function (Rahaman et al., 2019). That is, given an input $\mathbf{x} \in \mathbb{R}^C$, the network f has an output $\mathbf{y} = f(\mathbf{x}) \in \mathbb{R}^{C'}$, which can be explicitly written as:

$$f(\mathbf{x}) = \sum_{\epsilon} \mathbb{1}_{P_{\epsilon}(\mathbf{x})} (\mathbf{W}_{\epsilon} \mathbf{x} + \mathbf{b}_{\epsilon}), \quad (13)$$

where ϵ is an index of a (connected) region $P_{\epsilon} \subset \mathbb{R}^C$ and $\mathbb{1}_{P_{\epsilon}}$, the region's indicator function. $\mathbf{W}_{\epsilon} \in \mathbb{R}^{C' \times C}$ and $\mathbf{b}_{\epsilon} \in \mathbb{R}^{C'}$ are the effective weight and bias at the region P_{ϵ} , respectively. This interpretation of neural networks is straightforward for multilayer perceptrons. An MLP is a composition of linear layers and rectifiers, which are indeed continuous and piecewise linear, and a composition of a continuous piecewise linear function is also continuous and piecewise linear.

Convolutional neural networks with rectifiers are no different. A linear convolution operation

$$[\mathbf{k} \star \mathbf{x}]_{h,w} := \sum_{i=-\lfloor k_h/2 \rfloor}^{\lfloor k_h/2 \rfloor} \sum_{j=-\lfloor k_w/2 \rfloor}^{\lfloor k_w/2 \rfloor} [\mathbf{k}]_{i,j} [\mathbf{x}]_{h+i,w+j}, \quad (14)$$

where k_h and k_w define the height and width of the discrete kernel, can be viewed as a coordinate-wise linear layer with a flattened weight operating on a concatenation of translations of the input \mathbf{x} , *i.e.*,

$$\mathbf{k} \star \mathbf{x} = \mathbf{W}_{\mathbf{k}} \tilde{\mathbf{x}}, \quad \text{where} \quad (15)$$

$$\mathbf{W}_{\mathbf{k}} = \text{flatten}(\mathbf{k}), \quad \text{and} \quad (16)$$

$$\tilde{\mathbf{x}} = \text{concat}[\text{translate}[\mathbf{x}, (i, j)]]_{i,j}. \quad (17)$$

In other words, with a receptive field of size $R \times R$, an image $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$ is (linearly) transformed into the extended space $\tilde{\mathbf{x}} \in \mathbb{R}^{CR^2 \times H \times W}$, and the ConvNet $f : \mathbb{R}^{C \times H \times W} \rightarrow \mathbb{R}^{C' \times H \times W}$ is equivalent to the coordinate-wise function $\tilde{f} : \mathbb{R}^{CR^2} \rightarrow \mathbb{R}^{C'}$, which is continuous and piecewise linear.

Since our transform from f to \tilde{f} can be applied to all the K experts f_1, \dots, f_K of our a RGB autoencoder, we can abstract away the coordinates for simplicity. Another virtue of this reform is that we do not need to care about the router f_r in our analysis hereafter. From equation 1, for each coordinate $c \in [H] \times [W]$, we have:

$$[\xi]_c = [f(\mathbf{x})]_c = \sum_{k=1}^K [m_k \odot f_k(\mathbf{x})]_c = \sum_{k=1}^K [m_k]_c [f_k(\mathbf{x})]_c = \sum_{k=1}^K [m_k]_c \tilde{f}_k([\tilde{\mathbf{x}}]_c) =: \tilde{f}([\tilde{\mathbf{x}}]_c), \quad (18)$$

and since the coordinate-wise equivalent of each expert \tilde{f}_k is continuous and piecewise linear, its weighted summation is also piecewise linear, yet it may not be continuous. Moreover, the function's argument $[\tilde{\mathbf{x}}]_c \in \mathbb{R}^{CR^2}$ is a reshaping of the receptive field of size $R \times R$ at the locus c . As a result, our mixture of experts encoder f is equivalent to a piecewise linear, yet not generally continuous, function over $R \times R$ neighborhood of each pixel in the RGB representation \mathbf{x} . No matter how each pixel is distributed by the router, each feature vector $[\xi]_c$ at an arbitrary coordinate $c \in [H] \times [W]$ is a piecewise linear function of a local $R \times R$ neighborhood of image \mathbf{x} at the locus c .

A.2. A flawless autoencoder is a bad structure encoder

From the previous analysis, we may omit the spatial coordinate c and assume $\mathbf{x} \in \mathbb{R}^C$ be an image in the RGB color space, $\boldsymbol{\xi} \in \mathbb{R}^{C'}$ be the same image in the a RGB representation. Let $\tilde{\mathbf{x}} = [\mathbf{x} \ B(\mathbf{x})]^\top \in \mathbb{R}^{CR^2}$ be the flattened $R \times R$ neighborhood patch in the RGB domain. $B : \mathbb{R}^C \rightarrow \mathbb{R}^{C(R^2-1)}$ is a structure function that maps a center pixel \mathbf{x} to its peripheral pixels. For the simplicity, we discard the function notation and regard B as a vector in $\mathbb{R}^{C(R^2-1)}$ from now on. The a RGB autoencoder is equivalently reformed coordinate-wise. Let our a RGB encoder be $f : \mathbb{R}^{CR^2} \rightarrow \mathbb{R}^{C'}$ and the a RGB decoder be $g : \mathbb{R}^{C'} \rightarrow \mathbb{R}^C$. Note that we discard tilde from the encoder equivalent \tilde{f} for the sake of simplicity. Again, since the decoder is linear, we can also write $g(\boldsymbol{\xi}) = \mathbf{A}\boldsymbol{\xi} + \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{C \times C'}$ and $\mathbf{b} \in \mathbb{R}^C$.

The piecewise linear characteristic of the encoder f lets us rewrite the autoencoder into a form:

$$f(\tilde{\mathbf{x}}) = \sum_{\epsilon} \mathbb{1}_{P_{\epsilon}(\tilde{\mathbf{x}})} (\mathbf{W}_{\epsilon} \tilde{\mathbf{x}} + \mathbf{b}_{\epsilon}) = \boldsymbol{\xi}, \quad (19)$$

$$g(\boldsymbol{\xi}) = \mathbf{A}\boldsymbol{\xi} + \mathbf{b} = \mathbf{x}' \simeq \mathbf{x}, \quad (20)$$

following equation 13. The autoencoder $h = g \circ f$ can be written as:

$$h(\tilde{\mathbf{x}}) = g(f(\tilde{\mathbf{x}})) = \mathbf{A} \sum_{\epsilon} \mathbb{1}_{P_{\epsilon}(\tilde{\mathbf{x}})} (\mathbf{W}_{\epsilon} \tilde{\mathbf{x}} + \mathbf{b}_{\epsilon}) + \mathbf{b}, \quad (21)$$

$$= \sum_{\epsilon} \mathbb{1}_{P_{\epsilon}(\tilde{\mathbf{x}})} (\mathbf{A}\mathbf{W}_{\epsilon} \tilde{\mathbf{x}} + \mathbf{A}\mathbf{b}_{\epsilon} + \mathbf{b}). \quad (22)$$

Let ϵ' is a subscript for a connected region in the partition P that contains the coordinate c . That is, $\mathbb{1}_{P_{\epsilon'}(\tilde{\mathbf{x}})} = 1$ and $\mathbb{1}_{P_{\epsilon}(\tilde{\mathbf{x}})} = 0$ for $\epsilon \neq \epsilon'$. The summation from equation 22, then, can be simplified:

$$h(\tilde{\mathbf{x}}) = \mathbf{A}\mathbf{W}_{\epsilon'} \tilde{\mathbf{x}} + \mathbf{A}\mathbf{b}_{\epsilon'} + \mathbf{b} = \mathbf{x}' \simeq \mathbf{x}. \quad (23)$$

In other words, we only care about the specific region of the partition generated by the mixture of experts a RGB encoder that includes our pixel of interest. The flattened region $\tilde{\mathbf{x}}$ is decomposed into the center pixel \mathbf{x} and the peripherals B .

$$h(\tilde{\mathbf{x}}) = h\left(\begin{bmatrix} \mathbf{x} \\ B \end{bmatrix}\right) = \mathbf{A}\mathbf{W}_{\epsilon'} \begin{bmatrix} \mathbf{x} \\ B \end{bmatrix} + \mathbf{A}\mathbf{b}_{\epsilon'} + \mathbf{b} \simeq \mathbf{x}. \quad (24)$$

This can be further decomposed if we decompose $\mathbf{W}_{\epsilon'}$ into two matrices,

$$h(\tilde{\mathbf{x}}) = \mathbf{A} [\mathbf{W}_{\text{cen},\epsilon'} \quad \mathbf{W}_{\text{per},\epsilon'}] \begin{bmatrix} \mathbf{x} \\ B \end{bmatrix} + \mathbf{A}\mathbf{b}_{\epsilon'} + \mathbf{b}, \quad (25)$$

$$= \mathbf{A}\mathbf{W}_{\text{cen},\epsilon'} \mathbf{x} + \mathbf{A}\mathbf{W}_{\text{per},\epsilon'} B + \mathbf{A}\mathbf{b}_{\epsilon'} + \mathbf{b} \simeq \mathbf{x}, \quad (26)$$

where $\mathbf{W}_{\text{cen},\epsilon'} \in \mathbb{R}^{C' \times C}$ is the effective weight for the center pixel \mathbf{x} , and $\mathbf{W}_{\text{per},\epsilon'} \in \mathbb{R}^{C' \times C(R^2-1)}$ is the effective weight for the peripheral pixels B .

Let us now assume a perfect lossless autoencoder $h^*(\tilde{\mathbf{x}}) = \mathbf{x}$ for every combination of \mathbf{x} and B . That is, the autoencoder h^* always returns the exact same pixel \mathbf{x} without requiring B to be a function of \mathbf{x} . In this case, equation 26 can be further decomposed into three equations:

$$\mathbf{A}\mathbf{W}_{\text{cen},\epsilon'} \mathbf{x} = \mathbf{x} \in \mathbb{R}^C, \quad (27)$$

$$\mathbf{A}\mathbf{W}_{\text{per},\epsilon'} B = \mathbf{0} \in \mathbb{R}^{C(R^2-1)}, \quad \text{and} \quad (28)$$

$$\mathbf{A}\mathbf{b}_{\epsilon'} + \mathbf{b} = \mathbf{0} \in \mathbb{R}^C. \quad (29)$$

In particular, equation 27 should be satisfied for every combinations of RGB colors \mathbf{x} , and therefore we can conclude that $\mathbf{A}\mathbf{W}_{\text{cen},\epsilon'} = \mathbf{I}$ for the perfect autoencoder. Furthermore, equation 28 should be satisfied for every possible B , signifying that the encoder f^* of the perfect autoencoder h^* should project peripheral pixels to the nullspace of the decoder's weight \mathbf{A} , nullifying the information from the peripherals to propagate into estimating the pixel \mathbf{x} at the center. In other words, the encoder f does not learn to infer either \mathbf{x} from B or B from \mathbf{x} but simply acts as a separation of information between the center pixel and the peripheral pixels.

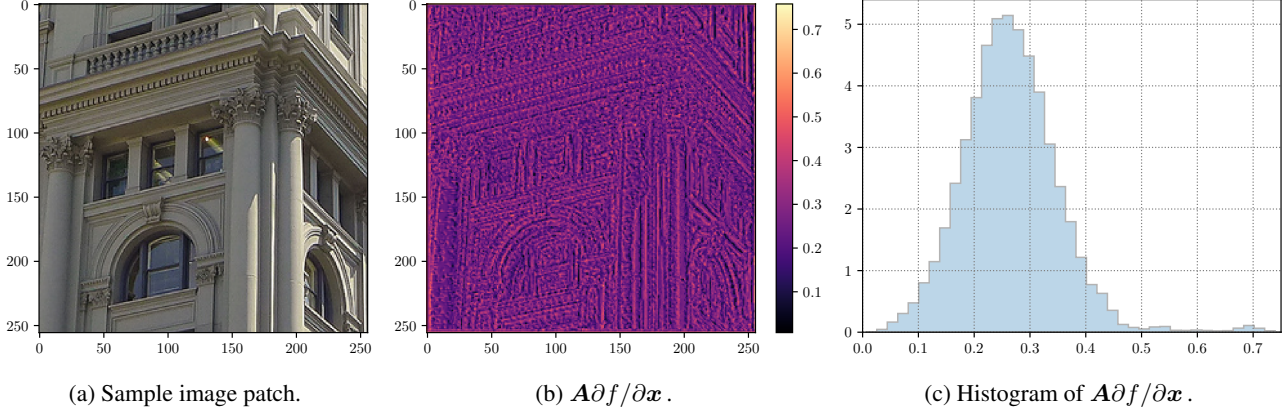


Figure 8. Measurement of the degree of self-reference of the a RGB encoder. The sample image is brought from Urban100 dataset (Huang et al., 2015). For a perfect autoencoder with no structure encoding capability, the values of $A \partial f / \partial \mathbf{x}$ should be 1 for every pixel. However, in spite of our high reconstruction accuracy (62.738dB PSNR for this patch, which corresponds to an average color deviation of a pixel of 0.069/255), the value of $A \partial f / \partial \mathbf{x}$ vary from 0.0066 to 0.7582, with average of 0.2654. This is unexpectedly low regarding the high accuracy, indicating that the reconstruction of our autoencoder heavily relies on the pixel’s neighboring structure. Note that $A \in \mathbb{R}^{C \times C'}$ and $f / \partial \mathbf{x} \in \mathbb{R}^{C' \times H \times W}$, where each element is obtained mutually independently. The heatmap and the histogram is obtained by taking the root-mean-square of the values over the three color channels.

In practice, however, this does not happen. In fact, the matrices $\mathbf{W}_{\text{cen}, \epsilon'}$ and $\mathbf{W}_{\text{per}, \epsilon'}$ are not just mathematical tools to represent the mechanism how the complicated nonlinear encoder f acts upon the given input \mathbf{x} . From equation 25, we can express only the encoder f as:

$$f(\tilde{\mathbf{x}}) = f\left(\begin{bmatrix} \mathbf{x} \\ B \end{bmatrix}\right) = [\mathbf{W}_{\text{cen}, \epsilon'} \quad \mathbf{W}_{\text{per}, \epsilon'}] \begin{bmatrix} \mathbf{x} \\ B \end{bmatrix} + \mathbf{b}_{\epsilon'} = \mathbf{W}_{\text{cen}, \epsilon'} \mathbf{x} + \mathbf{W}_{\text{per}, \epsilon'} B + \mathbf{b}_{\epsilon'}. \quad (30)$$

Taking the derivatives reveals interesting relationship between the gradients of f and the effective weight at a particular pixel.

$$\frac{\partial f}{\partial \mathbf{x}} = \mathbf{W}_{\text{cen}, \epsilon'} \quad \text{and} \quad \frac{\partial f}{\partial B} = \mathbf{W}_{\text{per}, \epsilon'}. \quad (31)$$

Therefore, for the lossless autoencoder, the relationship between the encoder’s gradient and the decoder’s weights are:

$$A \frac{\partial f}{\partial \mathbf{x}} = \mathbf{I} \quad \text{and} \quad A \frac{\partial f}{\partial B} = \mathbf{0}. \quad (32)$$

We can calculate the matrix multiplication $A \partial f / \partial \mathbf{x}$ and see how this differs from the identity to determine how the encoder mixes information between \mathbf{x} and B . Note that $\partial f / \partial \mathbf{x} \in \mathbb{R}^{C' \times C}$ is a gradient only at the locus c . Figure 8 shows a simple experiment to check if the value of $A \partial f / \partial \mathbf{x}$ deviates from unity. As the result presents, our a RGB encoder relies strongly on the neighboring structure to reconstruct a pixel.

B. Implementation detail

This section presents additional details for training the main a RGB autoencoder used throughout the experiments in Section 4.

B.1. Training the a RGB autoencoder

Architecture. The a RGB autoencoder consists of three models: a convolutional router f_r , $K = 20$ convolutional experts f_1, \dots, f_K , and a linear decoder g . The architecture is drawn in Table 6, 7, and 8, where C3 and C1 denote convolutions with kernel size 3 and 1, respectively, and L0.2 is a leaky ReLU with negative slope 0.2. Also, BN is a batch normalization with the same channel width as the output channel of the convolution in the same line. For 3×3 convolutions, we use zero padding of size 1. The router f_r consists of three 3×3 and two 1×1 convolutions, with batch normalization (Ioffe & Szegedy, 2015) and leaky ReLU with negative slope 0.2 between each pair of convolutions, except for after the first

Table 6. Architecture of f_r .

C3 \rightarrow L0.2	3ch \rightarrow 64ch
C3 \rightarrow BN \rightarrow L0.2	64ch \rightarrow 128ch
C3 \rightarrow BN \rightarrow L0.2	128ch \rightarrow 256ch
C1 \rightarrow BN \rightarrow L0.2	256ch \rightarrow 512ch
C1 \rightarrow Softmax	512ch \rightarrow 20ch

 Table 7. Architecture of f_k .

C3 \rightarrow L0.2	3ch \rightarrow 32ch
C3 \rightarrow L0.2	32ch \rightarrow 64ch
C3 \rightarrow L0.2	64ch \rightarrow 128ch
C3	128ch \rightarrow 128ch

 Table 8. Architecture of g .

C1 (no bias)	128ch \rightarrow 64ch
C1 (no bias)	64ch \rightarrow 32ch
C1	32ch \rightarrow 3ch

convolution, where we put only a leaky ReLU. Each expert f_k for all $k \in \{1, \dots, K\}$ has identical architecture, with four 3×3 convolutions and leaky ReLUs with negative slope of 0.2 in between. We do not use normalization layers in the experts. All the layers are initialized by the default PyTorch initializer (Paszke et al., 2019), meaning that each expert is initialized with different weights. Overall, the router f_r has a receptive field of size 7×7 , and each expert f_k has a receptive field of size 9×9 , hence $R = 9$ in Section 3.2. For the linear decoder, we use three 1×1 convolutions without activations and normalization layers, with each layer gradually reducing the channel dimension. Only the last convolution of the decoder has a bias term. We have empirically found that this three linear layer architecture leads to a slightly better convergence. We can easily compute the effective weight of the decoder by multiplying all the internal weights with standard matrix multiplication. The decoder has receptive field of a single pixel and is also linear.

Data preparation. Our training dataset consists of a mix of image patches obtained from several sources, including DIV2K (Agustsson & Timofte, 2017), Flickr2K (Timofte et al., 2017), DIV8K (Gu et al., 2019), and ImageNet-1k (Rusakovsky et al., 2015) datasets. The DIV2K, Flickr2K, and DIV8K datasets are high quality high resolution image datasets designed for training image super-resolution models. Those are selected in order to provide our autoencoder with rich structural information in clean, visually pleasing natural images. The DIV2K training set contains 800 high quality images, the Flickr2K set has 2,650 images, and the DIV8K dataset consists of 1,500 very high quality images up to 8K resolution. We also added the ImageNet-1k dataset to the training data to increase the diversity of image structures for supervision. Since the DIV2K, Flickr2K, and DIV8K datasets have much higher image size than one typically used for training a network, we have preprocessed these three dataset by cropping into patches of size 480×480 with stride 240. The results are 32,592 patches for the DIV2K, 107,403 patches for the Flickr2K, and 551,136 patches for the DIV8K dataset. The patch datasets are then concatenated with 1,281,167 images from the ImageNet-1k training data to build up the training data with 1,972,298 images in total. For each image fetched for training, a random crop of 256×256 is applied first, and then random horizontal and vertical flips, followed by a random 90 degrees rotation are applied consecutively.

Finding the optimal hyperparameters. We searched for the optimal architecture and other training hyperparameters by ablation studies mentioned in Section 5.4. For the ease of comparison, we unify the training process for the ablation studies in a simpler setting. Only using DIV2K dataset (Agustsson & Timofte, 2017) except for the second batch of experiments in Table 5, we vary the hyperparameters for training. Because the training dataset and the task become simple, we reduce the patch size to 192×192 and train our autoencoder for 300k iterations. Furthermore, since the per-pixel router f_r of our a RGB encoder distributes each pixel to a single expert, the number of iterations required for each expert to be trained using the same amount of data scales linearly to the number of experts. To compensate the effect, we changed the number of training iterations for the autoencoder accordingly. All the results in Section 5.4 is quantified by the accuracy of $\times 4$ image super-resolution task using RRDBNets (Wang et al., 2018b) trained to minimize the L_1 loss defined over the a RGB space. All the RRDBNet models are trained only with DIV2K dataset for 300k iterations from scratch.

Training hyperparameters. The weight of the load-balancing loss is selected to be $\lambda = 10^{-2}$ based on empirical observations. We found this to be the minimum value that ensures uniform expert assignment throughout the training process. This was chosen from a parameter sweep in the ranges from 10^0 to 10^{-5} in powers of 10. The network is trained with a batch size of 16. An Adam (Kingma & Ba, 2015) optimizer is used with its default hyperparameters and an initial learning rate of $5e-4$ is used. We use cosine learning rate schedule (Loshchilov & Hutter, 2017), starting from a period of 1k iteration, increased doubly up to 256k iterations. The training ends at 511k iterations.

Table 9. **Reconstruction accuracy of the α RGB autoencoder on out-of-distribution datasets.** We measured the PSNR scores *without* quantization to 255 scale. Average RGB difference in the second row stands for the average absolute difference in the pixel’s color value out of the maximum range of 255. All the values are significantly below the quantization gap of 0.5, indicating almost perfect reconstruction of the input.

	Set5	Set14	Urban100	DIV2K-Val	SIDD-Val	GoPro-Val
PSNR [dB]	67.206	64.531	65.556	70.812	72.007	72.853
Avg. RGB diff.	0.0477	0.0602	0.0669	0.0418	0.0301	0.0266

Autoencoding accuracy. For completeness, we provide the autoencoding accuracy for the validation datasets used throughout this work in Table 9. Because the output of the α RGB autoencoder deviates from the original image only less than a single quantization step ($1/255$) on average, we regard the α RGB autoencoder as almost lossless in our analysis except for Appendix A, where we find that the nonideal reconstruction capability helps the autoencoder to learn structural prior of images.

B.2. Notes on training image restoration models

Change in optimal training hyperparameters As emphasized throughout our manuscript, we have not changed the hyperparameters except for the very loss function in every experiment in Section 4. However, we also note that our demonstration does not necessarily mean that those setups are optimal. For instance, as mentioned in Section 5.3, altering the representation space leads to dramatic change in the scale and the shape of the gradients fed into the image restoration model during its training. Under stochastic gradient descent algorithms, this increase in the size of gradients of more than a hundredfold leads to significant changes in the training dynamics of those models. It is, therefore, less likely that the original set of hyperparameters is still optimal in the new representation space. Likewise, replacing the representation space may also change the optimal architecture for the image restoration task defined over the new space. Although we strongly believe that searching through the new possibilities allowed by our α RGB representation should be a fascinating research topic, this is beyond the scope of our paper. One example close to this direction is our demonstration of the NAFNet trained for $L_{1,\alpha\text{RGB}}$ loss, reported in Section 4. In this experiment, because the replacement of metric causes the change in scale and shape of its gradients, we have doubled the weight of the metric for better convergence. As a result, this new setup has led us to better denoisers than both the original one and the one trained with only the representation space being altered. We leave further exploration of this topic for future work.

Recommendation regarding gradient clipping As a final remark, it is highly recommended to remove gradient clipping to maximize the advantage of using α RGB-based losses. Section 3.2 attributes the performance gain caused by the additive noise to the sensitivity increase in the α RGB encoder f . In practice, the effect can be observed as an increment of two orders of magnitude in the size of gradients of the image restoration models being trained. The same scale of optimizer’s step size can only be achieved by increasing the learning rate a hundredfold, which quickly leads to training instability. We may safely conclude that the per-pixel distance losses in our α RGB space helps training of image restoration models by stably increasing the internal gradients. However, in recent image restoration techniques (Waqas Zamir et al., 2021; Chen et al., 2022), especially for the models with attention layers (Vaswani et al., 2017), gradient clipping is a common practice to stabilize the training of the model. To take advantage of our method, in Section 4, we changed gradient clipping mechanism to clamp at a inf norm of 20 for every experiment. This value is barely touched throughout the training process.

C. More quantitative results on perceptual image super-resolution

Table 10 extends Section 4 to provide full evaluation results on the $4\times$ perceptual super-resolution task using ESRGAN (Wang et al., 2018b). As mentioned in Section 4, our training process exactly follows the official implementation (Wang et al., 2018b). We first pre-train the network, RRDBNet, with DIV2K (Agustsson & Timofte, 2017) and Flickr2K (Timofte et al., 2017) combined, for 1M iterations using RGB L_1 loss. Then, the weights are fine-tuned with the loss written in the first column for 400k iterations using DIV2K, Flickr2K, and OSTv2 (Wang et al., 2018a) datasets combined. Preparation of the datasets are done with the same code the authors has provided. In this task, the α RGB loss is a simple L_1 loss over the range of the α RGB encoder. We leave all the other training hyperparameters untouched.

In addition to LPIPS (Zhang et al., 2018a), NIQE (Mittal et al., 2013), and FID (Heusel et al., 2017) metrics, we added

Table 10. Complete quantitative results for training ESRGAN $\times 4$ in the a RGB space. Improved results are highlighted in boldface characters.

Training objective	Set14						B100					
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	NIQE \downarrow	FID \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	NIQE \downarrow	FID \downarrow
$0.01L_1 + L_{VGG} + 0.005L_{Adv}^\dagger$	24.494	0.6543	0.1341	0.06374	3.8774	56.700	23.909	0.6205	0.1617	0.09603	3.6636	51.521
$0.01L_{1,aRGB} + L_{VGG} + 0.005L_{Adv}$	24.796	0.6623	0.1281	0.06052	3.7230	52.797	24.138	0.6306	0.1595	0.09578	3.4185	47.665
	Manga109						Urban100					
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	NIQE \downarrow	FID \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	NIQE \downarrow	FID \downarrow
$0.01L_1 + L_{VGG} + 0.005L_{Adv}^\dagger$	26.441	0.8170	0.0646	0.01036	3.5758	11.282	22.776	0.7033	0.1232	0.02432	4.2067	20.616
$0.01L_{1,aRGB} + L_{VGG} + 0.005L_{Adv}$	26.651	0.8186	0.0630	0.00863	3.4245	10.907	23.270	0.7196	0.1183	0.02050	3.8982	17.739
	DIV2K-Val						OST300					
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	NIQE \downarrow	FID \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	NIQE \downarrow	FID \downarrow
$0.01L_1 + L_{VGG} + 0.005L_{Adv}^\dagger$	26.627	0.7033	0.1154	0.00545	3.0913	13.557	23.249	0.6166	0.1688	0.05734	3.3834	21.851
$0.01L_{1,aRGB} + L_{VGG} + 0.005L_{Adv}$	26.845	0.7500	0.1110	0.00351	2.9615	12.799	23.649	0.6338	0.1662	0.05484	3.5247	19.952

[†]The official ESRGAN models (Wang et al., 2018b).

DISTS (Ding et al., 2020), a pairwise perceptual metric for image restoration. Moreover, four common benchmarks for the perceptual super-resolution, *i.e.*, Set14 (Zeyde et al., 2010), BSD100 (Martin et al., 2001), Manga109 (Matsui et al., 2017), and OutdoorSceneTest300 (Wang et al., 2018a) datasets are added for comparison. The updated results are shown in Table 10. The reported values are calculated in the RGB domain, cropping 4 pixels from the outer sides for PSNR, SSIM, and NIQE (Mittal et al., 2013) scores. For DISTS (Ding et al., 2020) and LPIPS (Zhang et al., 2018a), we use the code from official repo, and for FID (Heusel et al., 2017), we report scores from the PyTorch reimplementation (Seitzer, 2020) of the original Tensorflow code.

The data shows that $L_{1,aRGB}$ loss without the VGG perceptual loss yields comparable results to the L_1 loss without the perceptual loss in adversarial training. However, when equipped with VGG perceptual loss, the results show a general increase in performance over the original ESRGAN in all the distortion-based metrics (PSNR, SSIM), the pairwise perceptual metrics (LPIPS, DISTS), and the unpaired quality assessment metrics (NIQE, FID).

D. More qualitative results

This section provides more visual results from the main experiments. Figure 9 through 14 show results from the ESRGAN models (Wang et al., 2018b) trained with and without a RGB representation. Empirically, the supervision given by our a RGB encoder helps the model avoid generating visual artifacts and color inconsistency induced by adversarial training. Figure 15 shows additional results from the real image denoising task solved with NAFNet (Chen et al., 2022). Lastly, image deblurring using MPRNet (Waqas Zamir et al., 2021) trained from our a RGB representation is further demonstrated in Figure 16 and 17. As visualized in Appendix E, the additional information embodied in the extra dimensions of the a RGB representation resembles edginess information of an image. We conjecture that this allows a pairwise per-pixel distance defined over our a RGB space to provide image restoration models with stronger supervision leading to the reconstruction of sharper edges. The results reveal that this effect is realized as suppression of artifacts in the perceptual image super-resolution task, sharper produced images in the image denoising task, and better reconstruction of edges and more accurate alignments in the image deblurring task.

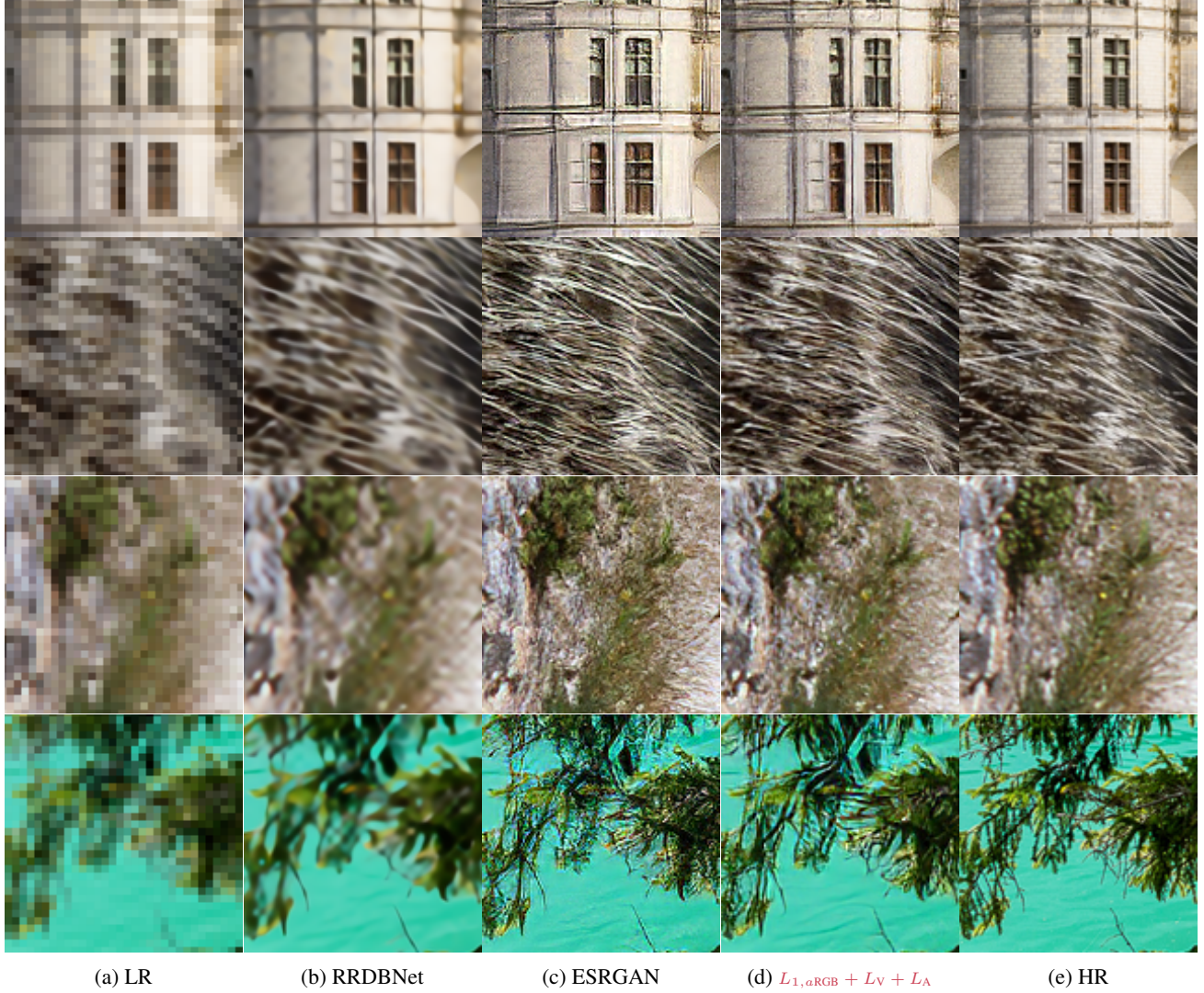


Figure 9. Qualitative comparison of ESRGAN models trained with different loss functions on DIV2K-Val (Agustsson & Timofte, 2017) benchmark.

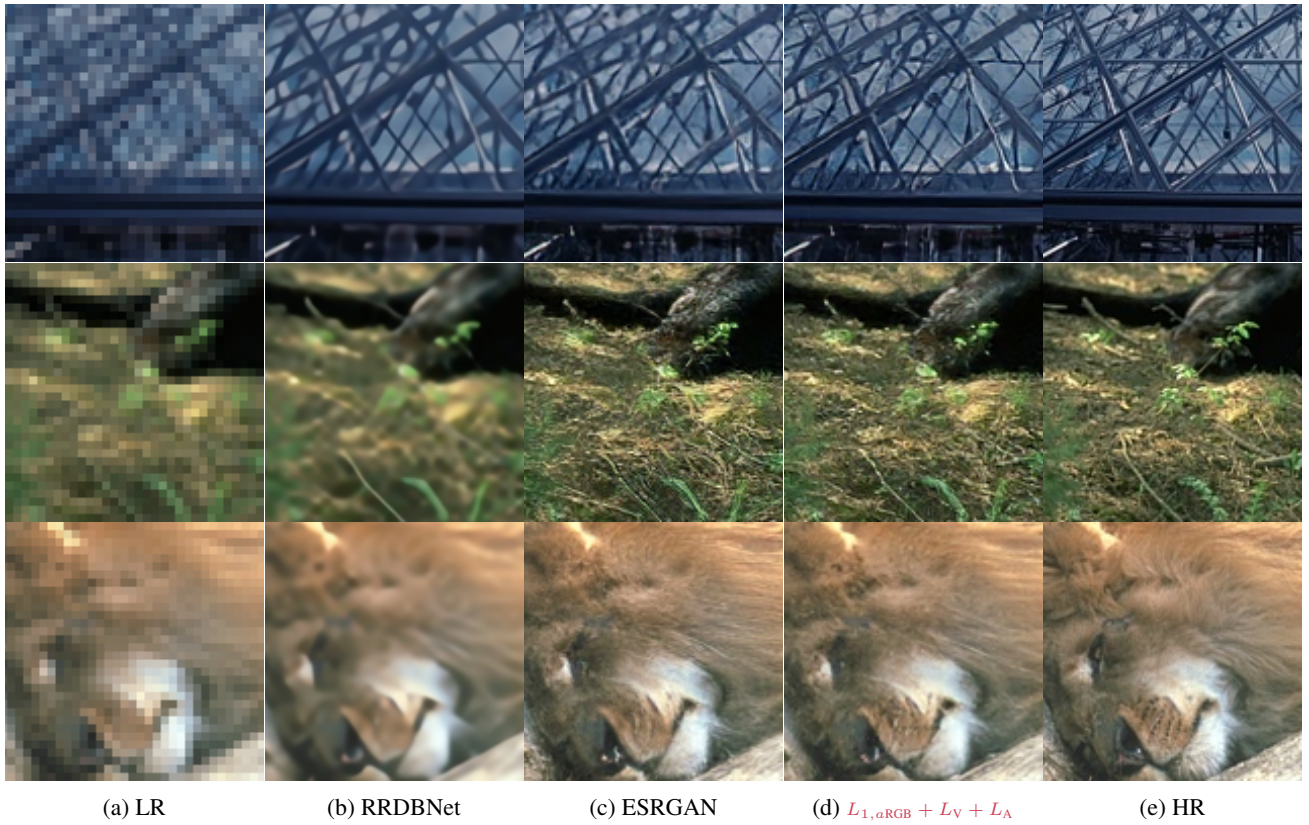


Figure 10. Qualitative comparison of ESRGAN models trained with different loss functions on B100 (Martin et al., 2001) benchmark.

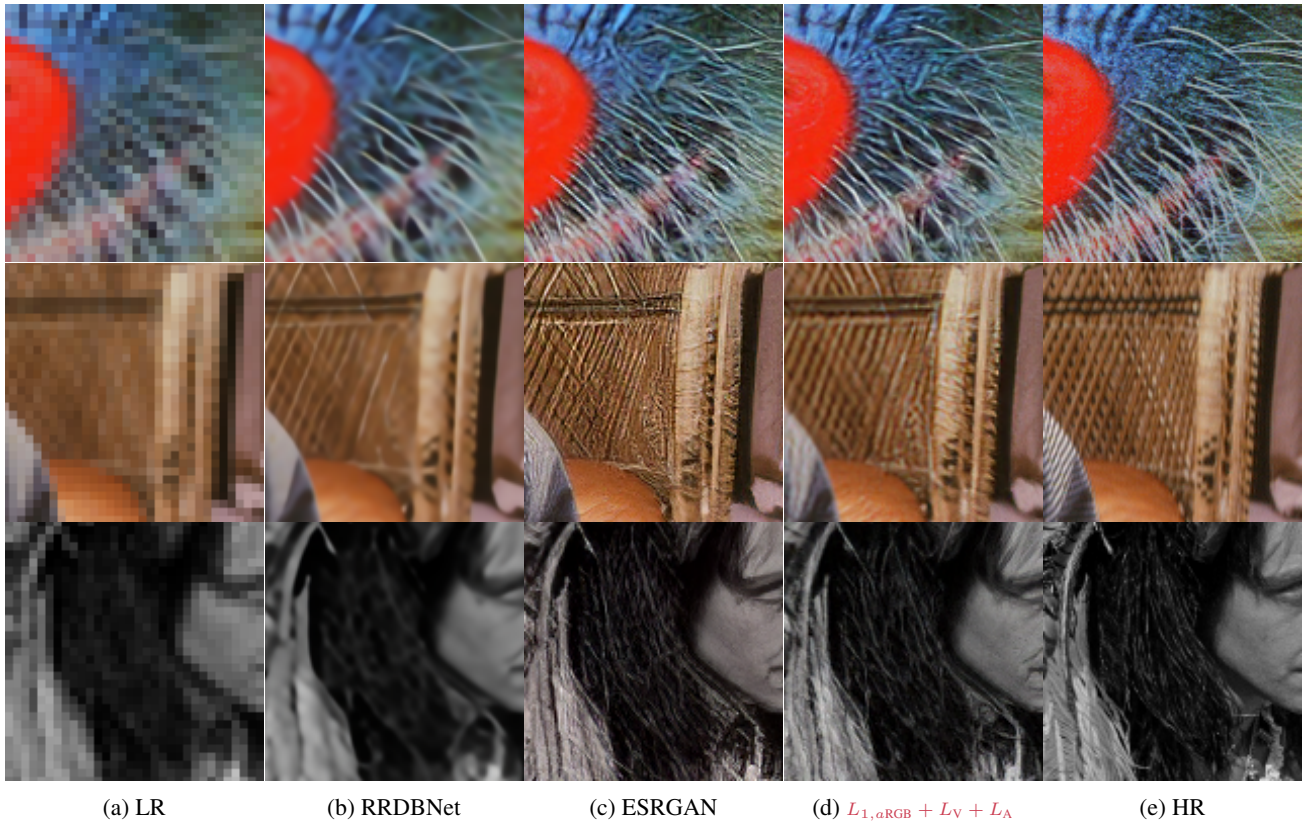


Figure 11. Qualitative comparison of ESRGAN models trained with different loss functions on Set14 (Zeyde et al., 2010) benchmark.

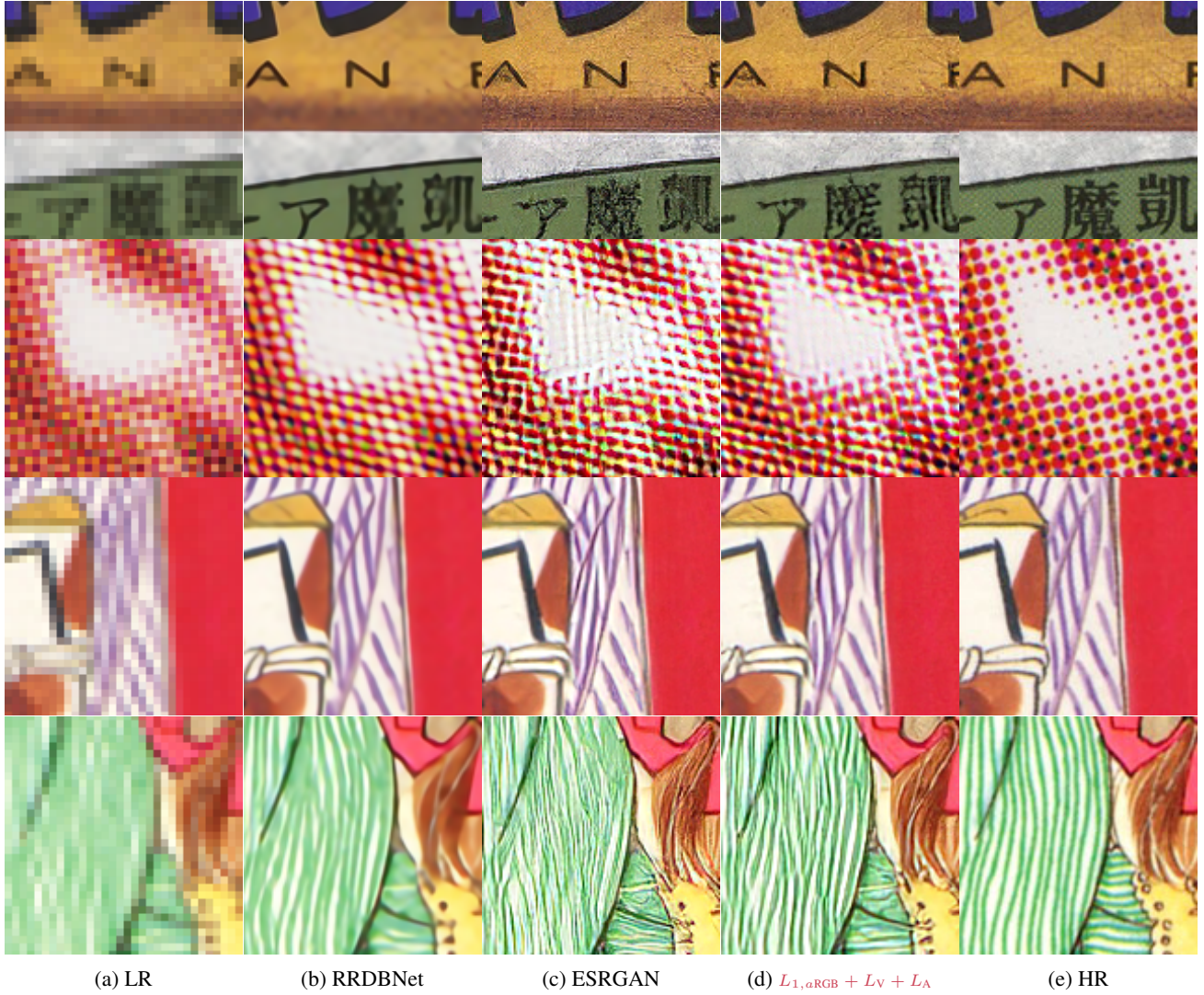


Figure 12. Qualitative comparison of ESRGAN models trained with different loss functions on Manga109 (Matsui et al., 2017) benchmark.

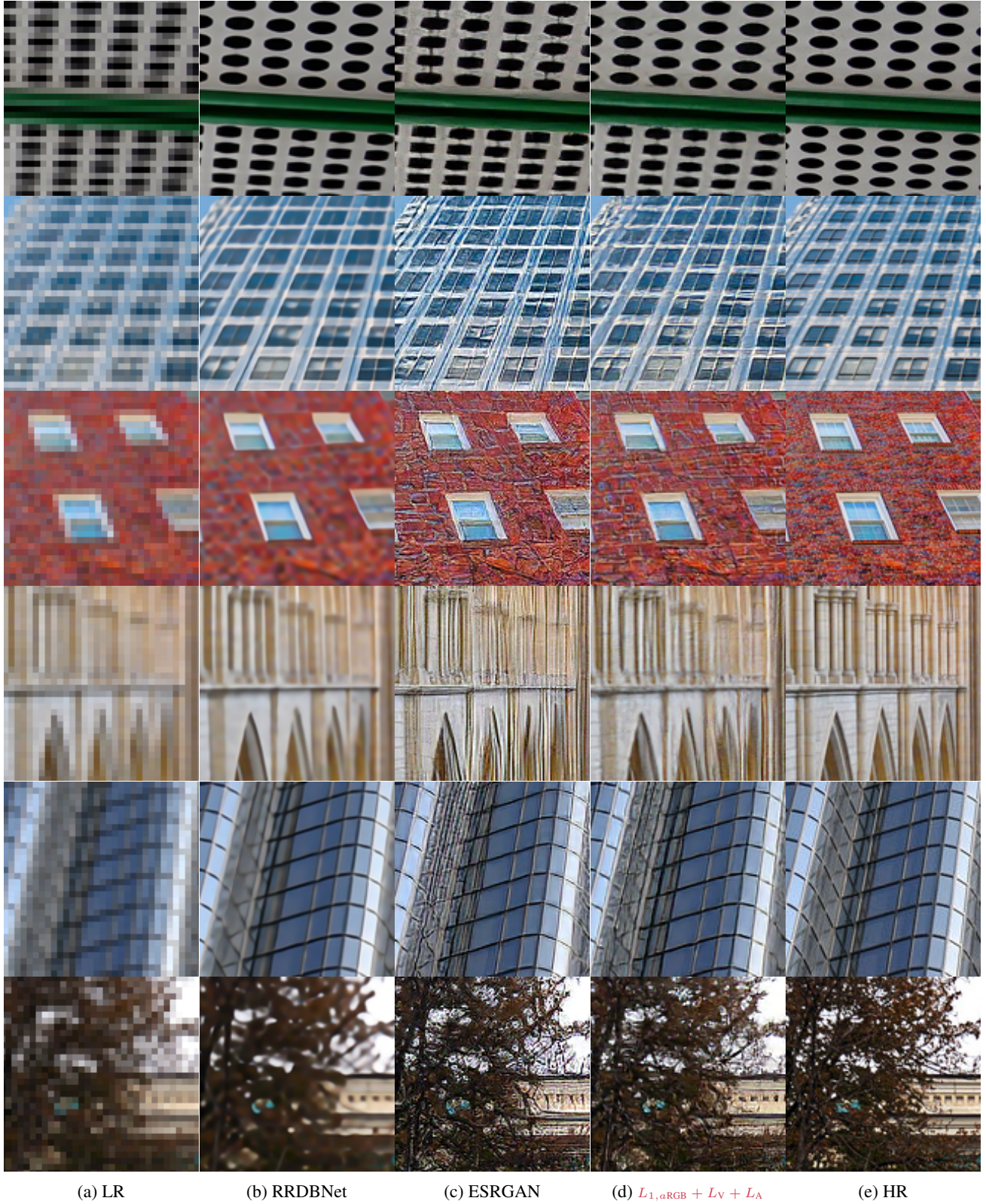


Figure 13. Qualitative comparison of ESRGAN models trained with different loss functions on Urban100 (Huang et al., 2015) benchmark.

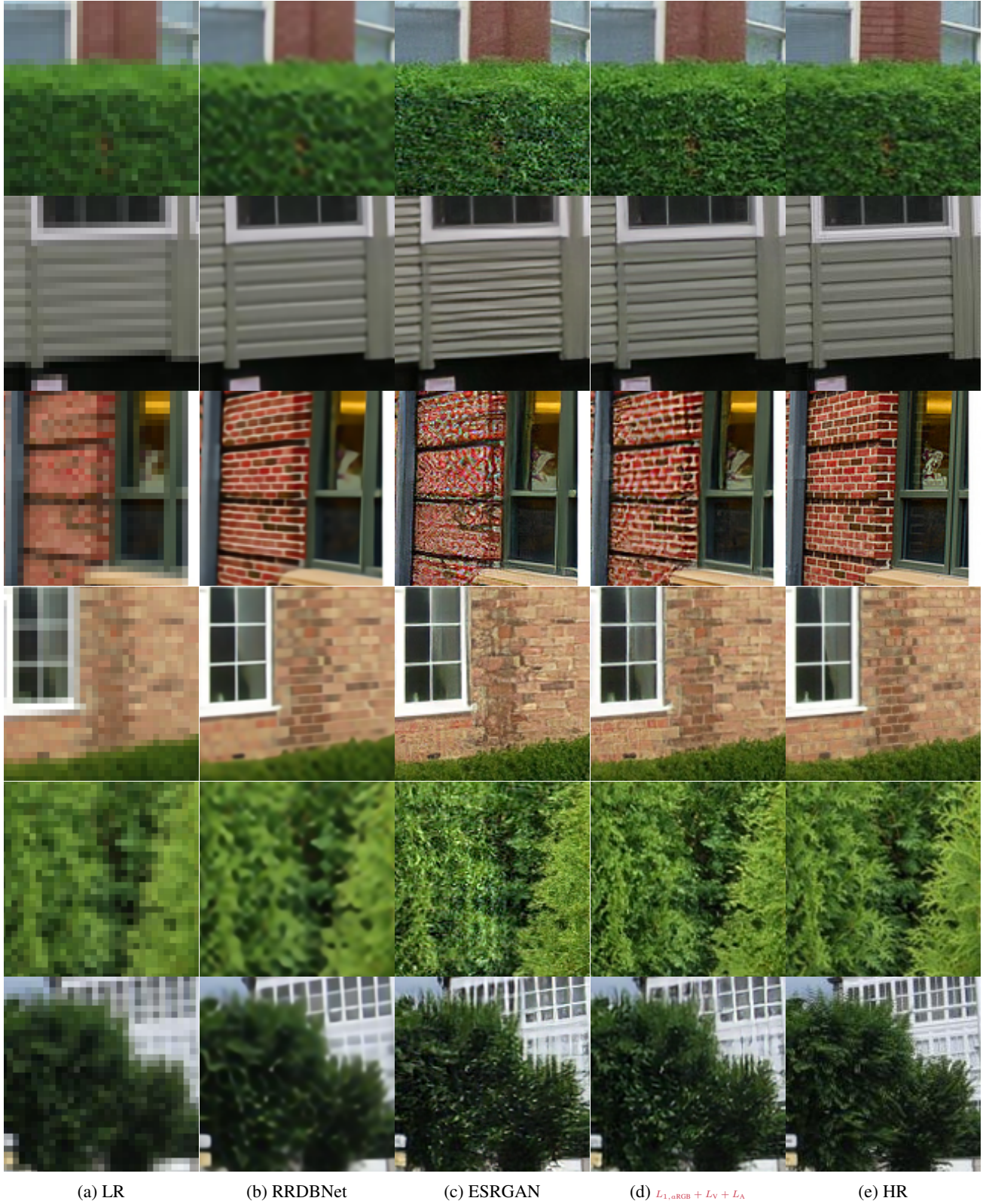


Figure 14. Qualitative comparison of ESRGAN models trained with different loss functions on OutdoorSceneTest300 (Wang et al., 2018a) benchmark.

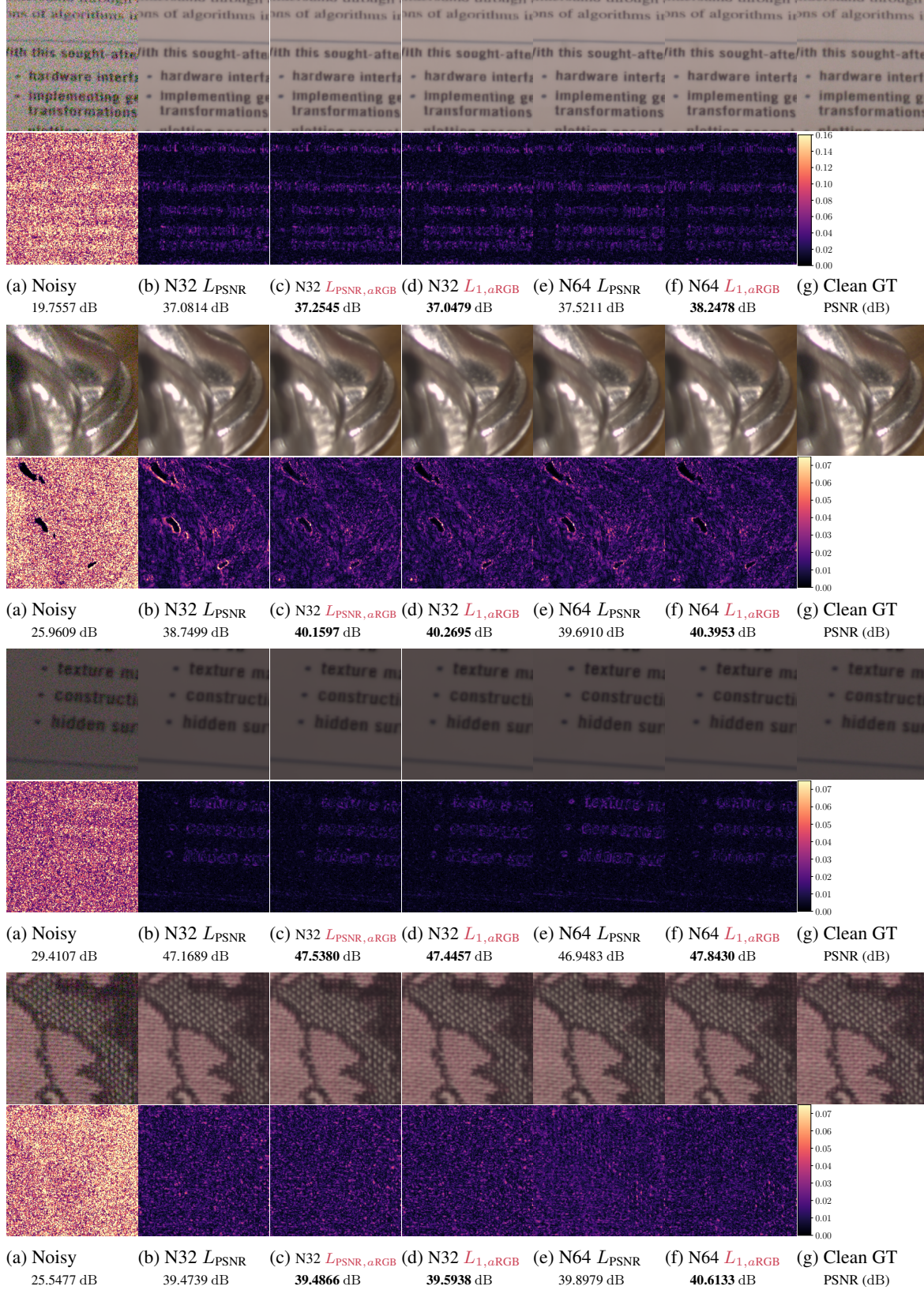


Figure 15. Qualitative comparison of real image denoising models on SIDD benchmark (Abdelhamed et al., 2018). Each column corresponds to each row in Table 1. N32 corresponds to NAFNet-width32 and N64 corresponds to NAFNet-width64. The bottom rows show the maximum absolute difference in color with a range of [0, 1].

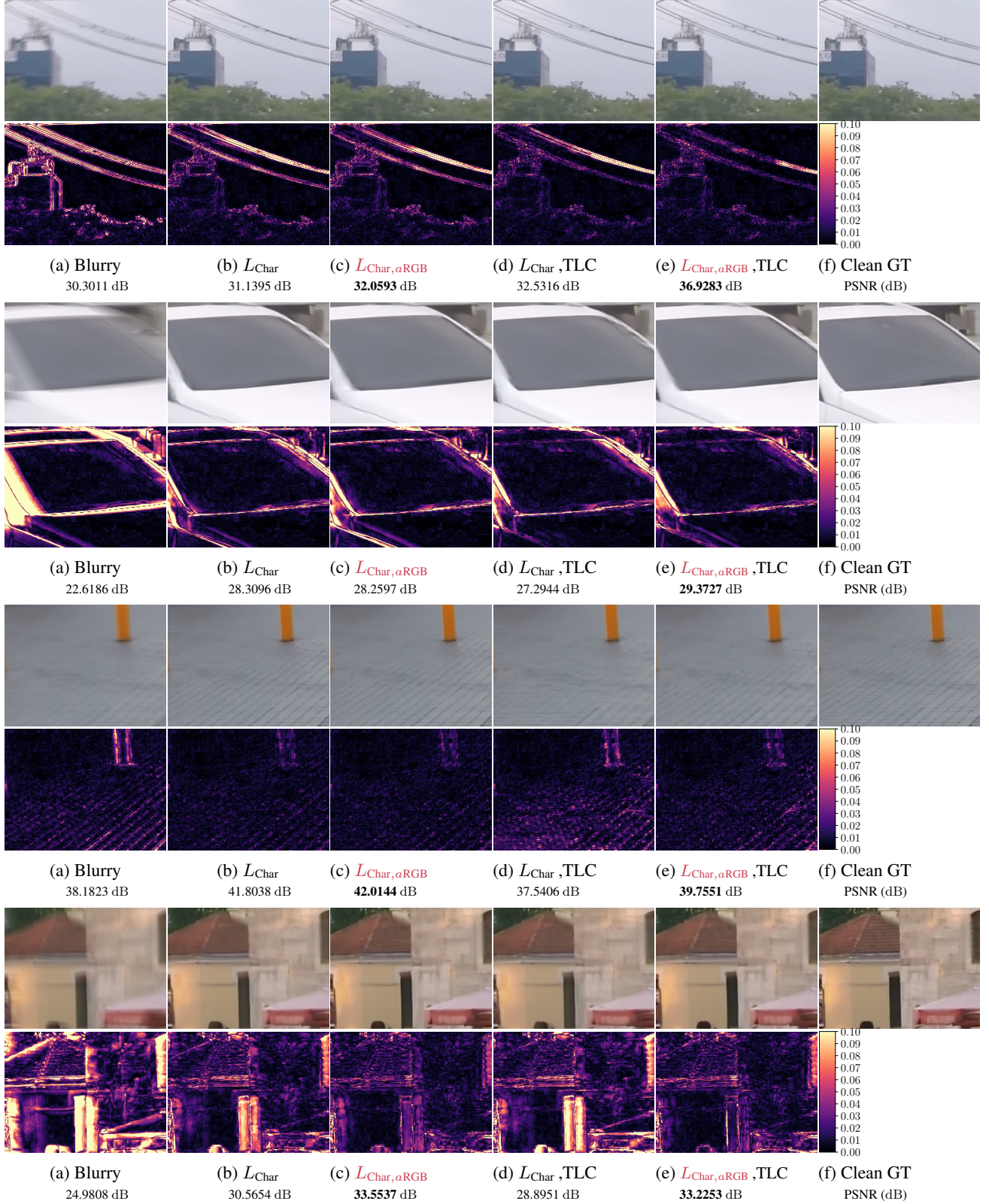


Figure 16. Qualitative comparison of motion blur deblurring models in GoPro benchmark (Nah et al., 2017). Each column corresponds to each row in Table 2. The bottom rows show the maximum absolute difference in color with a range of $[0, 1]$.

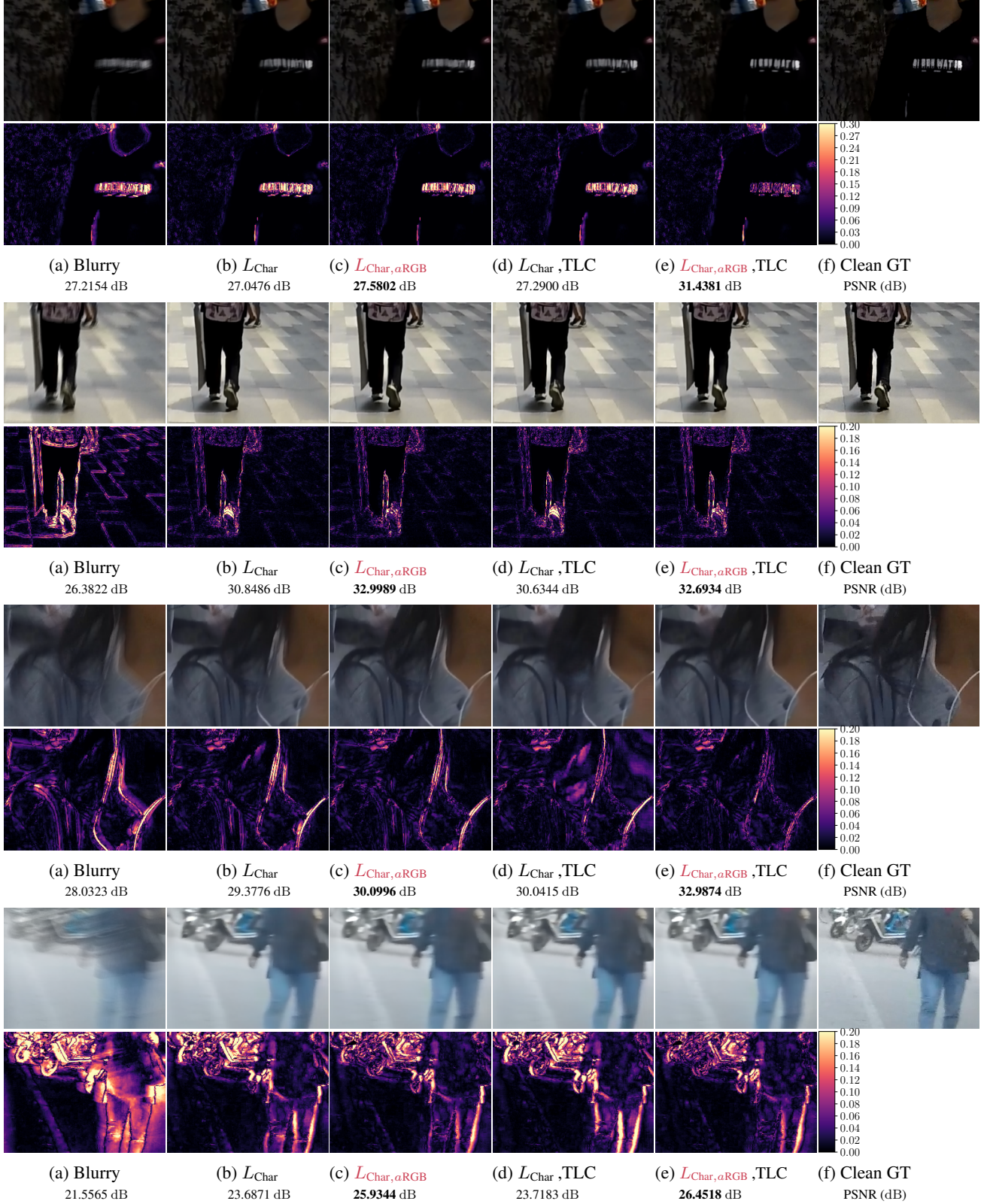


Figure 17. Qualitative comparison of motion blur deblurring models in HIDE benchmark (Shen et al., 2019). Each column corresponds to each row in Table 2. The bottom rows show the maximum absolute difference in color with a range of $[0, 1]$.

E. Understanding the a RGB representation space

This section is divided into three parts: In the first part of this section, we discuss further on the embedding decomposition test conducted in Section 5.1. Mixing different types of images in the same way as we did in Section 5.1 reveals how our a RGB representation encodes additional information in the extra dimensions. Next, we show more examples of the t-SNE (van der Maaten & Hinton, 2008) visualization of our learned embeddings in addition to Figure 6c. Unlike segmentation maps typically generated and assessed for semantic tasks, the ones produced by our a RGB encoder are extremely fine-grained and complicated, yet we find some common structural features in the results. In the last part, we conduct another experiment to visualize how each expert in the a RGB encoder f learns to specialize.

E.1. Decomposition of a RGB representation

In Section 5.1, we have discussed how the linearity of our decoder g helps understanding the learned representation space. In particular, given its linear weight \mathbf{A} and bias \mathbf{b} , we can decompose any given embedding in the a RGB representation $\xi \in \mathbb{R}^C$ into a sum of two orthogonal vectors:

$$\xi = \mathbf{A}^\dagger \mathbf{A} \xi + (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A}) \xi, \quad (33)$$

where we denote the matrix \mathbf{A}^\dagger as the Moore-Penrose pseudoinverse of \mathbf{A} . We can regard the multiplicand of the second term $\mathbf{I} - \mathbf{A}^\dagger \mathbf{A}$ as a linear projection operator onto the nullspace of \mathbf{A} , and the multiplicand of the first term $\mathbf{A}^\dagger \mathbf{A}$ as a linear projection onto the orthogonal complement of the nullspace of \mathbf{A} . Mathematically, and also empirically, it is easy to see that summation of any vector projected onto the nullspace of \mathbf{A} leads to no change in the decoded image $\mathbf{x}' = g(\xi)$. That is,

$$g(\xi + (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A}) \zeta) = \mathbf{A}(\xi + (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A}) \zeta) + \mathbf{b} \quad (34)$$

$$= \mathbf{A} \xi + \mathbf{b} + \mathbf{A}(\mathbf{I} - \mathbf{A}^\dagger \mathbf{A}) \zeta \quad (35)$$

$$= \mathbf{A} \xi + \mathbf{b} + (\mathbf{A} - \mathbf{A} \mathbf{A}^\dagger \mathbf{A}) \zeta \quad (36)$$

$$= \mathbf{A} \xi + \mathbf{b} + (\mathbf{A} - \mathbf{A}) \zeta \quad (37)$$

$$= \mathbf{A} \xi + \mathbf{b} \quad (38)$$

$$= g(\xi). \quad (39)$$

The identity in the fourth row is from the equality $\mathbf{A} = \mathbf{A} \mathbf{A}^\dagger \mathbf{A}$ of Moore-Penrose pseudoinverse. Therefore, the subspace to which the projection operator $\mathbf{I} - \mathbf{A}^\dagger \mathbf{A}$ is mapped is the allowed degree of freedom that additional information can be embedded.

Inversion of mixed embedding uncovers what information lies in this particular subspace. In addition to the results in Figure 6a, Figure 18 and 19 show visual results of a RGB embedding inversion. The flat image source is manually synthesized by a single color gradient. The real image source is a patch brought from DIV2K validation dataset (Agustsson & Timofte, 2017), and the Gaussian noise source is sampled from $\mathcal{N}(0.5, 0.5)$, where the RGB range is $[0, 1]$. Each patch has an equal size of 256×256 . First, we obtain the two a RGB embeddings from both sources. Then, we take the parallel part f_{\parallel} from the first source and the perpendicular part f_{\perp} from the second source. An image is optimized to produce the synthesized embedding as its a RGB representation. Due to the simple training setup, all the optimization quickly converges after 50 iterations of SGD with learning rate 0.1, as shown in Figure 20. We show the result of the optimization in Figure 18. The edge structure is extracted and highlighted by applying the Laplacian operator on the final image and displayed in Figure 19.

From the results, we can conclude the followings: Firstly, the parallel part of the a RGB embedding dominates the color information. This is predictable since this parallel component is the information that is decoded back to the RGB image with our linear decoder g . Secondly, the perpendicular part of the a RGB embedding conveys edge structure from the source image. As a consequence, we can retrieve this information from the a RGB representation following the same process used to train the inversion in this section. Moreover, this edgeness is the additional information brought with our a RGB representation to serve as a performance boost for training the image restoration models. As our main results in Section 4 suggests, the reduction of visual artifacts in perceptual image super-resolution and the enhancement of edge structures in image denoising and deblurring are attributed to this particular information carried in the perpendicular embeddings. Lastly, we can clearly observe that this additional information is diminished away if the underlying image is highly noisy, far from the manifold of clean, natural images. From this observation, we can conclude that the information learned by the a RGB autoencoder from its training dataset gives the model a structural prior in order to process images similar to its original

training data. This argument aligns with our finding in the ablation studies in Section 5.4, where a network trained from different dataset generally produce poorer restoration models.

E.2. Topology of the learned a RGB space

We have conjectured in Section 3.1 that the underlying distribution of small image patches are disconnected, exhibiting very complex structures. We can indirectly check the validity of our argument by visualizing the learned embeddings of an image using dimension reduction techniques. Specifically, we use t-SNE (van der Maaten & Hinton, 2008) algorithm to embed $H \times W = 65,536$ vectors of size $C = 128$ into a two dimensional plane to visualize the structure of our learned embedding for a particular image patch.

In this small experiment, we expect two outcomes: First, we expect perceptually distinct groups of embedding vectors to be appeared in the t-SNE results. Although the geometric information of the 2D projection carries little meaning in depicting the exact structure of the underlying feature space, visually distinguishable clustering in this region will signify that the structure of underlying representation is not connected into a single manifold, but consists of multiple disconnected regions with distinct characteristic values. Second, we expect that the experts are well specialized, requiring that clusters of a RGB *pixels* each assigned to a single expert to be appeared in the embeddings. Figure 21 shows three more examples in addition to Figure 6b and 6c from our main manuscript. The sample images are picked from different datasets, *i.e.*, DIV2K (Agustsson & Timofte, 2017), Urban100 (Huang et al., 2015), and Manga109 (Matsui et al., 2017), having different color distributions, contents, and styles. However, as the results show that the embeddings generated from this set of images have several commonalities: First, the embeddings are clustered in a well-separated regions. We can observe two distinct types of clusters: *common* groups where multiple experts are involved in generating similar embeddings and *expert-specific* groups where a single expert dominates in encoding the information of pixels. Existence of the first type of groups indicates the existence of common subspace between the feature spaces of each specialized expert. The latter type of groups show a clear evidence of expert specialization in our a RGB encoder. From the observations, we conclude that our initial design philosophy of the network serves its original purpose.

E.3. Visualization of the learned features of the a RGB encoder

In this last section of our paper, we provide another visualization results to facilitate understanding of the behavior of our a RGB encoder. In order to visualized the learned features for individual expert of our encoder, we simply maximize the activation of a single last channel of one of the experts. Starting from a random image of size 32×32 , which is more than three times larger than the receptive field of our network, we run a simple maximization of the average activation of each channels to produce a single feature image for each channel of an expert. Figure E.3 shows some of the results. First, we observe that channels of the same index at each of the experts are maximally activated at the similar color distribution. This similarity comes from the shared linear decoder of our autoencoder. However, we also notice that the same set of filters are maximally stimulated at different *patterns*. The results uncover another evidence of expert specialization in our a RGB autoencoder.

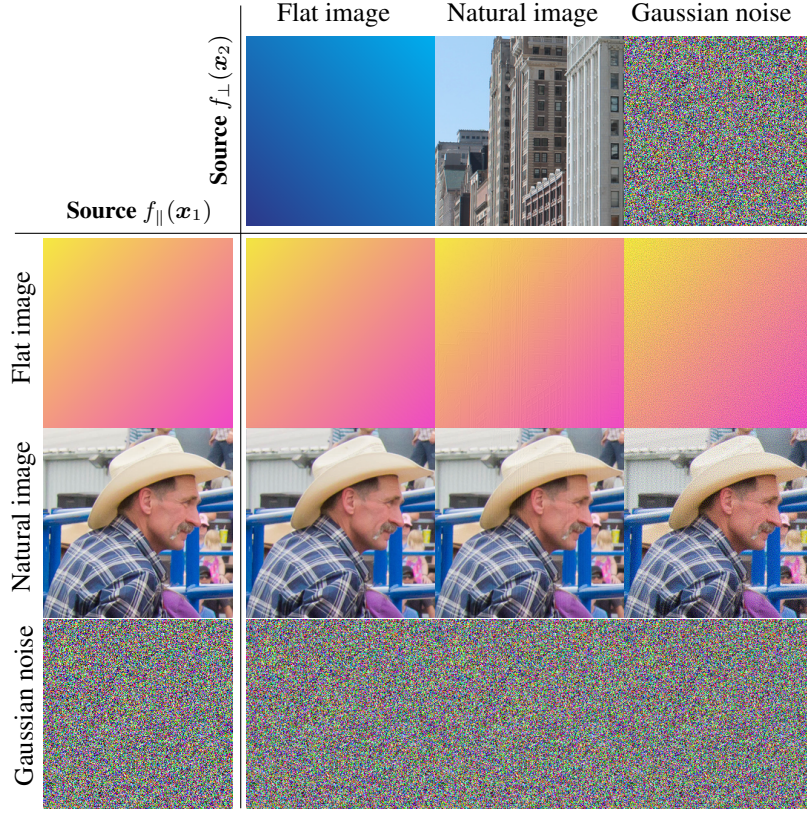


Figure 18. **Decomposition of the a RGB representation space.** The a RGB embeddings of the images on the sides are decomposed into orthogonal components and mixed $\xi_{\text{mix}} = f_{\parallel}(x_1) + f_{\perp}(x_2)$. Each cell is the image corresponding to the mixed embedding $f^{-1}(\xi_{\text{mix}})$.

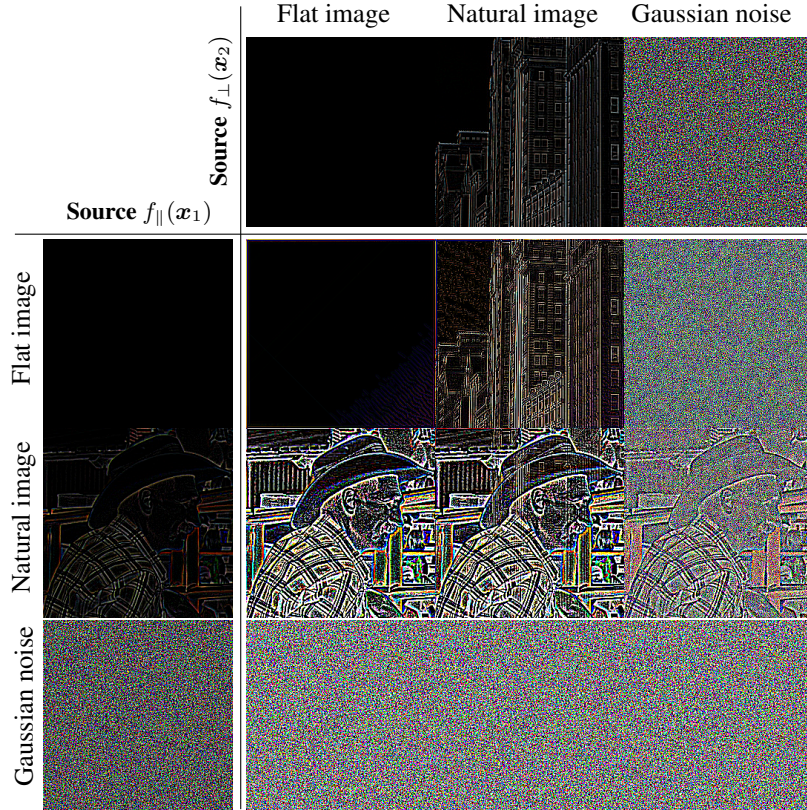


Figure 19. **Edge-enhanced inversion results of Figure 18.** A discrete Laplacian operator is applied to the same images in Figure 18 to enhance the high-frequency structures for clearer understanding. The results reveal that the perpendicular component of the a RGB embedding f_{\perp} contributes to high-frequency structures.

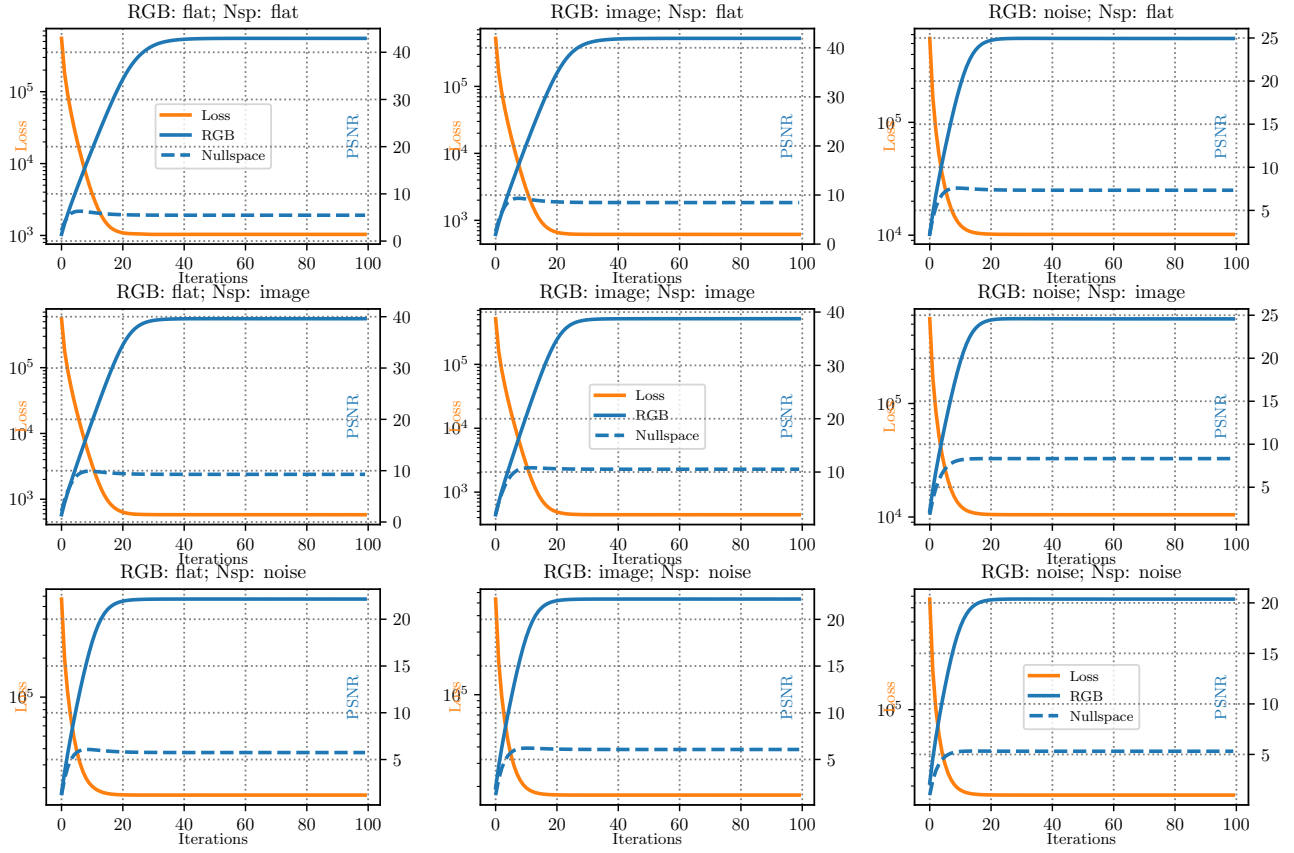


Figure 20. Training curve for the decomposition test. All the embedding inversion test quickly converge after 50 iterations. *RGB* corresponds to the source image \mathbf{x}_1 used for the parallel component $\xi_{\parallel} = \mathbf{A}^\dagger \mathbf{A} f(\mathbf{x}_1) = \mathbf{A}^\dagger \mathbf{A} \xi_{\text{mix}}$ of the target *aRGB* embedding ξ_{mix} , and *Nullspace* corresponds to the source image \mathbf{x}_w used for the perpendicular component $\xi_{\perp} = (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A}) f(\mathbf{x}_2) = (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A}) \xi_{\text{mix}}$, where $m\mathbf{A}$ is the weight of the linear decoder g . As shown in Figure 18, The low-frequency color distribution of the resulting inversion follows that of the parallel component’s source \mathbf{x}_1 , resulting in high PSNR scores. Although the PSNR scores between the inversions $f^{-1}(\xi_{\text{mix}})$ and the corresponding source images \mathbf{x}_2 of the perpendicular component $f_{\perp}(\mathbf{x}_2)$ are low, Figure 19 reveals that the perpendicular components encode high frequency information of the image.

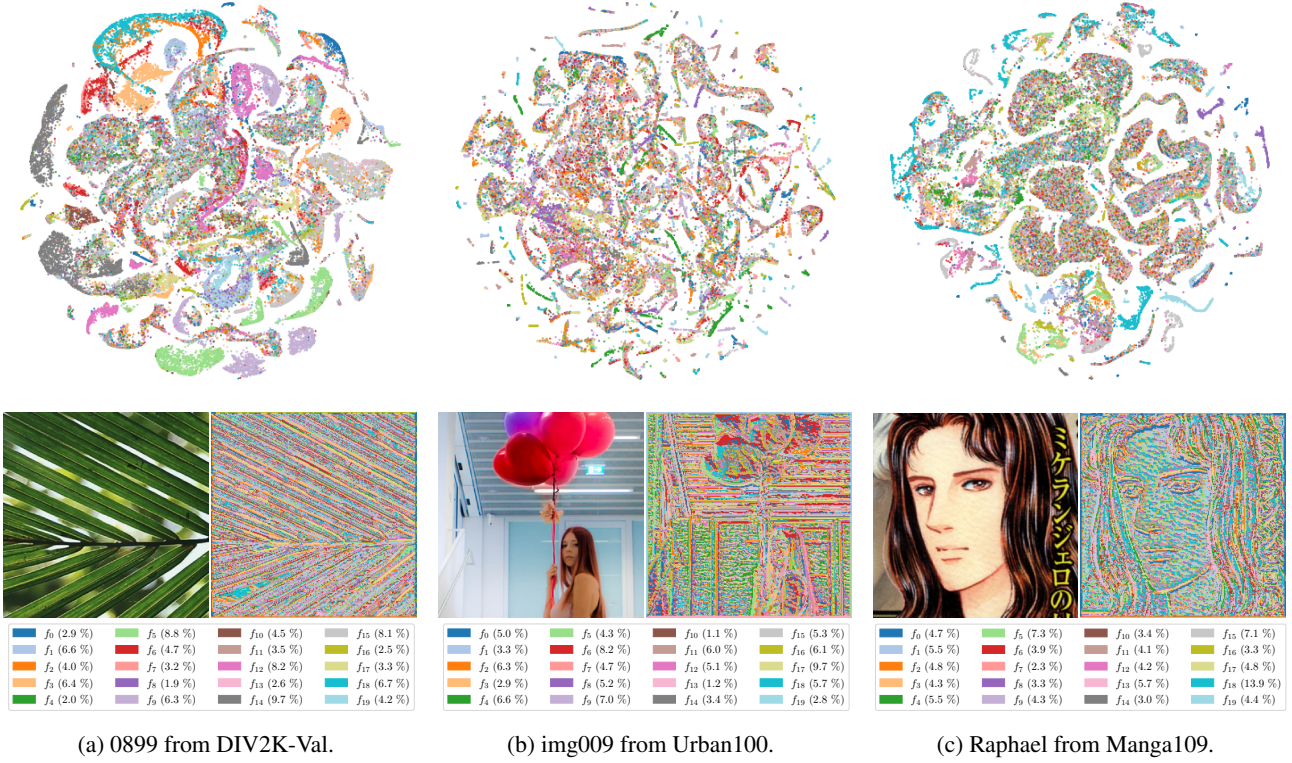


Figure 21. More examples on expert specialization using t-SNE and segmentation map. Sample images are brought from three well-used super-resolution benchmark datasets, *i.e.*, DIV2K (Agustsson & Timofte, 2017), Urban100 (Huang et al., 2015), and Manga109 (Matsui et al., 2017). Although the content and the style of each patches are widely different, the distribution of the learned $aRGB$ embeddings in these patches exhibit similar pattern: the distributions are decomposed into *common* groups, where multiple experts are involved in the encoding, and *expert-specific* groups.

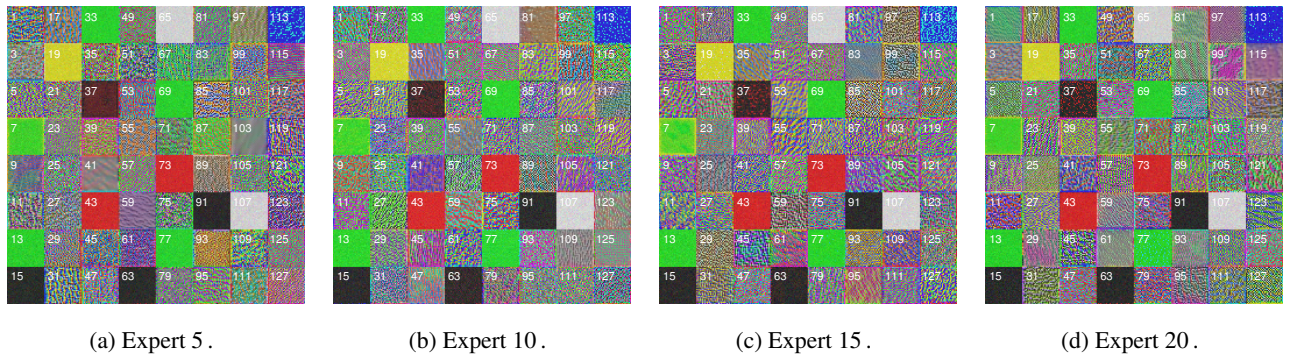


Figure 22. Visualization of the output filters of the experts. Randomly initialized 32×32 images are trained to maximize a specific filter at the last convolutional layer of the selected expert. The ID of each filter is annotated with white numbers. Note that the $aRGB$ representation space has a dimension of 128, the same as the number of filters in the last layer of each experts. The results show that while filters of different experts encoding the same channel is maximally activated at a similar average color, the high-frequency patterns each filter maximally attends to vary significantly.