sensors

MDPI

Article

Gram-GAN: Image Super-Resolution Based on Gram Matrix and Discriminator Perceptual Loss

Jie Song, Huawei Yi *, Wenqian Xu, Bo Li and Xiaohui Li

School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou 121001, China
* Correspondence: dxxyyihuawei@lnut.edu.cn

Abstract: The solution of a high-resolution (HR) image corresponding to a low-resolution (LR) image is not unique in most cases. However, single-LR–single-HR supervision is widely adopted in single-image super-resolution (SISR) tasks, which leads to inflexible inference logic of the model and poor generalization ability. To improve the flexibility of model inference, we constructed a novel form of supervision, except for the ground truth (GT). Specifically, considering the structural properties of natural images, we propose using extra supervision to focus on the textural similarity of the images. As textural similarity does not account for the position information of images, a Gram matrix was constructed to break the limitations of spatial position and focus on the textural information. Besides the use of traditional perceptual loss, we propose a discriminator perceptual loss based on the two-network architecture of generative adversarial networks (GAN). The difference between the discriminator features used in this loss and the traditional visual geometry group (VGG) features is that the discriminator features can describe the relevant information from the perspective of super-resolution. Quantitative and qualitative experiments were performed to demonstrate the effectiveness of the proposed method.

Keywords: super-resolution; generative adversarial network; perceptual loss

Citation: Song, J.; Yi, H.; Xu, W.; Li, B.; Li, X. Gram-GAN: Image Super-Resolution Based on Gram Matrix and Discriminator Perceptual Loss. *Sensors* **2023**, *23*, 2098. https://doi.org/10.3390/s23042098

Academic Editor: Jan Cornelis

Received: 6 December 2022
Revised: 9 February 2023
Accepted: 10 February 2023
Published: 13 February 2023

1. Introduction

Single-image super-resolution reconstruction (SISR) is a classic image processing task. Its target is to obtain a corresponding high-resolution (HR) image through some logical inference based on an existing low-resolution image (LR). In the current information age, people have increasingly higher requirements for image resolution (e.g., medical, monitoring and multimedia industries), which makes SISR have high practical value.

In recent years, with the rapid development of deep learning, neural network models related to SISR have emerged in an endless stream. The pioneering work was a proposal of SRCNN [1], which first applied convolutional neural networks to super-resolution (SR) tasks and substantially improved the quality of reconstructed images. Thereafter, a large number of PSNR-oriented methods [2–7] have emerged, which have the uniform property of a loss function consisting of a single mean square error (MSE), pixel-wise loss [8,9]. Although neural networks have strong learning ability, the models so far have focused only on maximizing PSNR, resulting in the problem of over-smooth images.

Perception-driven methods [10–15] are proposed to solve the over-smoothness problem. These methods extract image features through the first half part of the pre-trained partial visual geometry group (VGG) network [16], and then construct the perceptual loss. The loss makes the network have the ability to reason about high-frequency texture details compared with pixel-wise loss, so as to obtain visual effects more in line with human perceptual habits.

Although perception-driven methods substantially improve the visual quality of images, the use of one-to-one supervision in most models is not reasonable. For one thing, LR images are not in a fixed one-to-one relationship with HR images. For another,

multiple HR images may be downsampled to the same LR image (the downsampling method is uncertain). Therefore, one-to-many supervision needs to be constructed to improve the flexibility of model inference.

To solve the above problem, Li et al. [17] implemented one-to-many supervision based on the similarity between patches. However, its similarity measurement standard was the Euclidean distance (i.e., treating all content information within patches equally), which led to the possibility that additional selected supervised patches may differ from the ground-truth (GT) patch in details (cf. the image over-smoothness problem in PSNR-oriented methods), thus affecting the visual quality of the images. In this paper, to produce higher quality images, we use the Gram matrix to develop a supervision that emphasizes textural information. The model can flexibly generate more realistic textures and avoid some distorted structures under this supervision. In addition, we believe that the VGG features used in traditional perceptual loss are not fully adapted to SR models. The original purpose of these features was to be applied to the image recognition task [18,19], which makes the feature type required by the model in the SR task not rich enough. To enhance the richness of the feature types, we propose using the features of the middle layer of the discriminator [20] for the training of the generator. With the combined effect of the discriminator and VGG features, the network can learn richer inference logic, and thus generate more natural textural details.

In this paper, we refer to the models obtained by the above two proposed methods as Gram-GAN. Gram-GAN is compared with a large number of perception-driven methods to demonstrate its advancement, and ablation experiments are conducted to verify the necessity of each method.

The main contributions of this paper are itemized as follows:

1. In order to improve the flexibility of model inference, this paper proposes a method of constructing a Gram matrix for patches to formulate another supervision except for GT. This supervision ignores the position information of images and focuses only on texture information, which can reduce the generation of distorted structures with a large deviation from GT.
2. We propose a discriminator perceptual loss dedicated to the SR task based on the two-network architecture of generative adversarial networks (GAN), which can give the network some additional inference logic from the SR perspective compared with traditional perceptual loss.
3. Massive advanced perception-driven methods are used to compare their performance with Gram-GAN to demonstrate the advancement of the proposed method, and ablation experiments are performed to verify the respective necessity of the constructed extra supervision and discriminator perceptual loss.

2. Related Work

In this section, we introduce the current SR methods from two perspectives, which are PSNR-oriented methods [2–7] and perception-driven methods [10–15,17].

2.1. PSNR-Oriented Methods

With the proposal of SRCNN [1], deep learning in SR tasks have become increasingly mature, and massive models aimed at improving PSNR values have been proposed. In particular, Kim et al. [2] proposed VDSR, which improved the performance of the model by significantly increasing the number of network layers. Ledig et al. [3] combined the ideas of ResNet [21] and proposed the SRResNet. Zhang et al. [4] proposed RCAN, which constructed a channel attention module to focus on improving the PSNR value. Hu et al. [5] proposed Meta-SR to achieve the effect of upsampling images to arbitrary sizes. Li et al. [6] proposed a feedback framework to gradually refine the super-resolved results.

2.2. Perception-Driven Methods

It has been found that most PSNR-oriented methods suffer from a severe image over-smoothness problem, which is inextricably linked to the using a single pixel-wise loss. To enable the model to have the ability to reason about texture details, perceptual loss [10] was proposed. The idea was to use a pre-trained VGG model to extract image features and then compare the similarity of deep features between the predicted image and the GT image. With the great success of perceptual loss in SR, a series of perception-driven approaches have emerged. Ledig et al. [3] proposed SRGAN, which applied both GAN [20] and perceptual loss to the SR task to further improve the visual quality of images. Wang et al. [11] made improvements based on SRGAN and proposed ESRGAN. In particular, the modification of the network structure substantially improved the learning ability of the model, and thus reconstructed the finer textures. Rad et al. [12] made adjustments to the composition of perceptual loss and proposed a target perceptual loss based on object, background and boundary labels. Importantly, Li et al. [17] considered that one-to-one supervision was not the most reasonable way, and proposed the Beby-GAN with one-to-many supervision. However, the extra supervision of the method was selected by finding the patches that had the shortest Euclidean distance from the estimated patches. This easily generated texture details that differed significantly from GT patches. In addition, VGG features were oriented towards image recognition tasks, so the current perceptual loss did not enable the model to reason about other details in the images. The construction of additional types of perceptual loss is crucial to enhance inference capability of the model. To this end, we propose using Gram-GAN to solve these problems.

3. Methods

The whole framework of the proposed Gram-GAN is constructed based on GAN, as shown in Figure 1. The overall network consists of a generator and discriminator. The generator uses the RRDB [11], with a strong learning capability to adapt a series of complex loss functions, and the discriminator uses a variant of the VGG network. In this section, we first introduce extra supervision and construct the patch-wise texture loss. Then, we illustrate a novel discriminator perceptual loss. Finally, the other loss functions used in the model are mentioned.

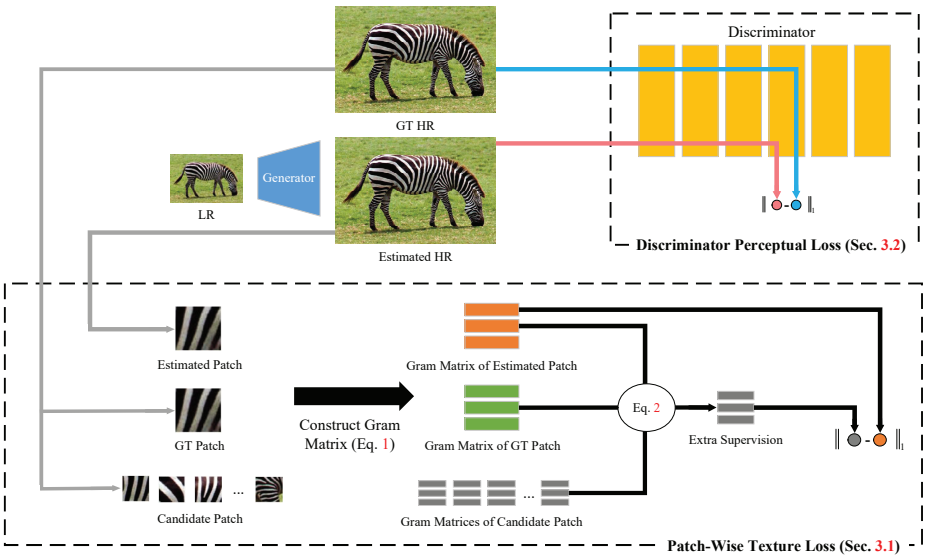

Figure 1. The whole framework of Gram-GAN. The discriminator perceptual loss is constructed by the output of the middle layer of the discriminator. The patch-wise texture loss gives texture-based supervision to the estimated patch.

3.1. Extra Supervision Based on Gram Matrix

In order to enhance the flexibility of model inference through one-to-many supervision, a practical extra supervision needs to be added on top of the original GT image supervision set on the estimated images. Inspired by [17], we set this extra supervision from the patch. To fit the SR task, we considered that texture similarity needed to be given more attention, rather than all content information being treated equally when finding extra supervision. The reason for this was that in most natural images, due to the limitation of location information, it is much harder to find a patch similar to the estimated patch in content than texture, except for the GT patch. Therefore, to construct another kind of supervision more reasonably, this paper proposes to construct the corresponding Gram matrix for each patch to achieve the purpose of ignoring the location information and focusing only on texture information, as follows.

First, the patch is defined as $p \in \mathbb{R}^{S \times C}$, where S represents the dimension after the multiplication of height and width (the two dimensions are combined) and C represents the number of channels. The Gram matrix construction function can be expressed as

$$G(p) = p^T p. \tag{1}$$

The Gram matrix used in this paper was not constructed based on the feature extraction mechanism of the pre-trained network, but was directly constructed from the original features. Considering that every patch carries a small amount of information, the Gram matrix constructed by the original features was sufficient to distinguish between different texture properties. Therefore, the use of a complex feature extraction mechanism was unnecessary.

Then, the selection method for the extra supervision was formulated by joint decision of the GT patch in [17] and the estimated patch. However, the difference is that the measure of similarity between patches is no longer based on the full content, but on the texture. The extra supervision in the i-th iteration can be represented as

$$p_i^* = \arg\min_{p \in \mathcal{O}} [\alpha \|G(p) - G(g_i)\|_2^2 + \beta \|G(p) - G(e_i)\|_2^2], \tag{2}$$

where g_i and e_i represent the GT patch and estimated patch in the i-th iteration, respectively. α and β represent the corresponding weights and $\mathcal{O}$ denotes the candidate database. In particular, besides the patch set composed of the GT patch and downsampled GT patch, we added a patch set with affine transformation [22] to the candidate database to enrich the selectable texture types. The specific affine transformation operation can be formulated as

$$\begin{bmatrix} x^{(1)} \\ y^{(1)} \\ 1 \end{bmatrix} = (\lambda N + I) \begin{bmatrix} x^{(0)} \\ y^{(0)} \\ 1 \end{bmatrix}, \tag{3}$$

where $x^{(0)}$ and $y^{(0)}$ represent the original horizontal and vertical coordinates of the GT patch, respectively. $I \in \mathbb{R}^{3 \times 3}$ is the identity matrix. $x^{(1)}$ and $y^{(1)}$ represent the horizontal and vertical coordinates through affine transformation, respectively. $N \in \mathbb{R}^{3 \times 3}$ is a random matrix conforming to the standard normal distribution, and λ is used to control the magnitude of the affine transformation. Some unconventional distorted patches were added to the candidate dataset after this affine transformation, which could be considered as new texture types for selection.

To highlight the benefits of pre-constructing the Gram matrix when selecting extra supervision, in Figure 2, we show the difference between two methods of calculating the Euclidean distance, which include the use of the Gram matrix and direct calculation. p1, p2 and p3 are patches containing the same type of textures, and p4 and p5 are patches containing other types of textures. When measuring the distance between two patches, we expect that the distance between patches with similar textures would be much smaller than

the distance between patches with non-similar textures. From the figure, it can be observed that the direct calculation of Euclidean distance cannot distinguish the patches with similar and non-similar textures well, and the desired effect can be better achieved only after the Gram matrix is constructed. Therefore, the degree of similarity of the different textures can be measured more accurately after the introduction of the Gram matrix, which helps to select the extra supervision that can bring more benefits to the model.

Finally, for each patch pair (e_i, p_i^*) obtained by Equation (2), the corresponding patch-wise texture loss is represented as

$$\mathcal{L}_{PT}(e_i, p_i^*) = \|G(e_i) - G(p_i^*)\|_1. \tag{4}$$

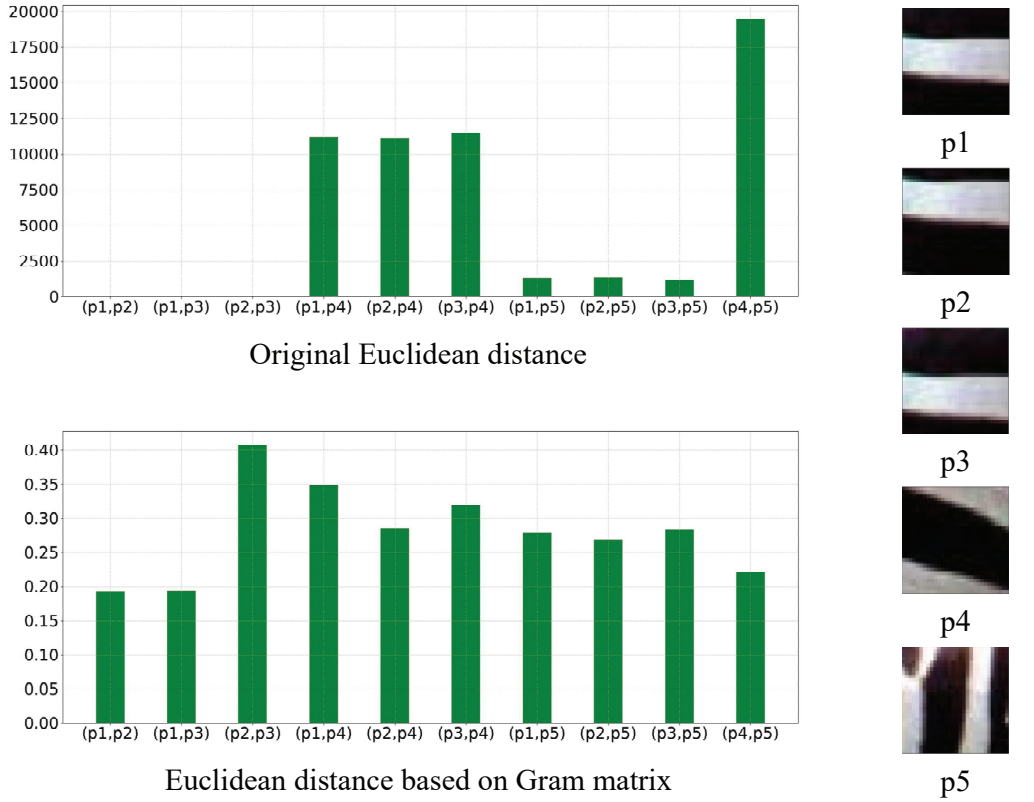

Figure 2. Comparison of the two distance calculation approaches with different texture information. (p1, p2) represents the distance between patch p1 and p2, and so on.

3.2. Discriminator Perceptual Loss

The features extracted by the pre-trained VGG network were initially dedicated to the image recognition task, which made these features focus on the parts that were useful for this task. However, the SR task requires that the richer feature types predict every detail of the images. Therefore, the composition of the perceptual loss should not rely only on VGG features, but also on some additional features extracted for the SR task. Based on the above theory, this paper proposes to use the discriminator in GAN to construct a novel perceptual feature. Specifically, the discriminator in each iteration is used to extract features, and the discriminator perceptual loss corresponding to the i-th iteration can be represented as

$$\mathcal{L}_{DP} = \sum_k \left\| D_k^{(i)}(x_{SR}) - D_k^{(i)}(x_{GT}) \right\|_1, \tag{5}$$

where $D_k^{(i)}$ represents the feature output by the k-th convolutional layer (after activation) of the discriminator at the i-th iteration. x_{SR} and x_{GT} represent the estimated and GT image, respectively.

Figure 3 shows the difference between VGG features and discriminator features, and the difference between the two features is very obvious. VGG features only highlight features useful for image recognition work (e.g., eyes), whereas the discriminator features emphasize features relevant to SR tasks (e.g., textures). Compared with VGG features, the discriminator features can highlight the differences between estimated images and GT images in more detail from different perspectives. Therefore, the network makes inferences based on more types of features after discriminator perceptual loss is added, to further improve the quality of the estimated images.

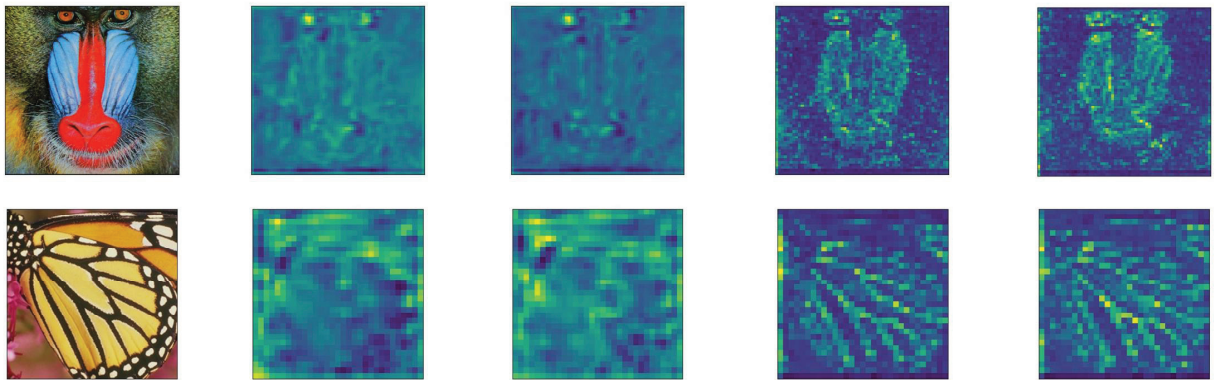

Figure 3. Comparison of VGG features and discriminator features on the estimated image and GT image. The estimated image is generated by RRDB.

3.3. Other Loss Functions

3.3.1. Perceptual Loss

In addition to the use of discriminator perceptual loss proposed in Section 3.2, traditional perceptual loss [10] is also considered in this paper, which is represented as

$$\mathcal{L}_P = \sum_i \lambda_i \|\Phi_i(x_{SR}) - \Phi_i(x_{GT})\|_1, \tag{6}$$

where Φ_i represents the i-th activation layer in the pre-trained VGG19 network. λ_i represents the coefficient of balance loss. Following [17], the layers we considered included $conv_{3\_4}$, $conv_{4\_4}$ and $conv_{5\_4}$, and the corresponding scaling coefficients were $\frac{1}{8}$, $\frac{1}{4}$ and $\frac{1}{2}$, respectively.

3.3.2. Adversarial Loss

For adversarial training under the GAN [20] mechanism, we used relativistic average GANs (RaGANs) with region-perceptual ability based on the ideas proposed in [17,22]. The loss functions of RaGANs can be represented as

$$\mathcal{L}_D = -\mathrm{E}_{x_r^{\text{mask}} \sim \mathbb{P}}\left[\log\left(D_{Ra}\left(x_r^{\text{mask}}\right)\right)\right] - \mathrm{E}_{x_f^{\text{mask}} \sim \mathbb{Q}}\left[\log\left(1 - D_{Ra}\left(x_f^{\text{mask}}\right)\right)\right], \tag{7}$$

$$\mathcal{L}_G = -\mathrm{E}_{x_r^{\text{mask}} \sim \mathbb{P}}\left[\log\left(1 - D_{Ra}\left(x_r^{\text{mask}}\right)\right)\right] - \mathrm{E}_{x_f^{\text{mask}} \sim \mathbb{Q}}\left[\log\left(D_{Ra}\left(x_f^{\text{mask}}\right)\right)\right], \tag{8}$$

where

$$D_{Ra} = \begin{cases} \text{Sigmoid}\left(C(x) - \mathrm{E}_{x_f^{mask} \sim \mathbb{Q}} C\left(x_f^{\text{mask}}\right)\right), & x \text{ is real} \\ \text{Sigmoid}\left(C(x) - \mathrm{E}_{x_r^{\text{mask}} \sim \mathbb{P}} C\left(x_r^{\text{mask}}\right)\right), & x \text{ is fake}, \end{cases} \tag{9}$$

where $C(\cdot)$ is the discriminator used to determine the true or false image, x_r^{mask} represents real data that is sampled from distribution $\mathbb{P}$ and partially masked and x_f^{mask} represents the fake data that is sampled from distribution $\mathbb{Q}$ and partially masked. The binary mask that masks the true and false data can be represented as

$$M_{i,j} = \begin{cases} 1, & \text{std}(B_{i,j}) \geq \delta \\ 0, & \text{std}(B_{i,j}) \leq \delta, \end{cases} \tag{10}$$

where $B_{i,j}$ represents the patch with coordinates (i, j) obtained by unfolding the image (length and width are fixed). δ is the predefined threshold, and $std(\cdot)$ is the operation of calculating the standard deviation. The value of δ and size of the patch were set to 0.005 and 11×11, respectively [17].

3.3.3. Content Loss

The content loss was used to evaluate the $\ell 1$-norm distance between the estimated and GT image, and was formulated as

$$\mathcal{L}_C = \|x_{SR} - x_{GT}\|_1. \tag{11}$$

The reason for using $\ell 1$-norm instead of $\ell 2$-norm is as follows: The advantage of the $\ell 1$-norm over the $\ell 2$-norm is that the $\ell 1$-norm is insensitive to outliers. As this paper uses a GAN-based network architecture, the model is trained in an adversarial way. This adversarial training inevitably results in some outliers. Therefore, we needed to use $\ell 1$-norm to reduce the impact of outliers as much as possible to enhance the stability of the model training.

3.3.4. Overall Loss

Based on the above sections, the overall loss of the generator is

$$\mathcal{L} = \eta_1 \mathcal{L}_{PT} + \eta_2 \mathcal{L}_{DP} + \eta_3 \mathcal{L}_P + \eta_4 \mathcal{L}_G + \eta_5 \mathcal{L}_C, \tag{12}$$

where $\eta_1 = 1.0$, $\eta_2 = 1.0$, $\eta_3 = 1.0$, $\eta_4 = 0.005$ and $\eta_5 = 1.0$. In particular, the reason why the weight of $\mathcal{L}_G$ was taken as 0.005 is as follows: For the weights in Equation (12), the purpose was to ensure the consistency of magnitudes among the losses and to prevent the phenomenon that some losses do not bring gains to the model. Regarding the acquisition of specific values, our strategy was to take the value of the initial state of the loss as a benchmark and calculate the corresponding weight values with the goal of balancing the magnitude differences. As the value of $\mathcal{L}_G$ in the initial state is larger compared with other losses, setting the weight to 0.005 can better balance the effect of each loss.

4. Experiments

4.1. Datasets and Similarity Measures

The training set was from 800 high-resolution images of the widely used dataset DIV2K [23]. All images were cropped by sliding window and expanded to obtain 44,226 non-overlapping sub-images with the size of 192×192. The test sets were Set5 [24], Set14 [25], BSD100 [26] and Urban100 [27], which had 5 images, 14 images, 100 images and 100 images, respectively.

In this paper, we used four evaluation metrics. The ones with reference objects were peak signal-to-noise ratio (PSNR), structure similarity (SSIM) [28], learned perceptual image patch similarity (LPIPS) [29], and the one without reference objects was natural image quality evaluator (NIQE) [30]. Among these, a higher PSNR and SSIM mean better resolution, and lower LPIPS and NIQE mean better resolution.

4.2. Training Details

All experiments were performed at 4× scaling factor and NVIDIA GeForce RTX 2080Ti GPUs were used. In order to make a fair performance comparison between our proposed Gram-GAN and the baseline model BebyGAN [17], we referred to the basic experimental configuration of Beby-GAN. Specifically, the optimizer was Adam with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The size of the input image for the training set was 48×48, and the data was enhanced by random rotation and flip. The size of the mini-batch was set to 8. In Section 3.1, the size of each candidate patch was 4×4 and the magnitude of random affine transform was 0.003. In Section 3.2, we used the features under the 5-th and 11-th convolutional layers to obtain the discriminator perceptual loss. The total number of iterations was 600 k, and every 200 k iterations was a period. The initial learning rate for each period was 1×10^{-4} and accompanied with the warm-up and cosine decay.

4.3. Comparison with State-of-the-Art Technologies

We compare the proposed Gram-GAN with other state-of-the-art perception-driven methods, including SRGAN [3], ESRGAN [11], SFTGAN [13], ESRGAN+ [15] and Beby-GAN [17]. In this paper, we evaluate model performance based on both quantitative and qualitative results, and the details are described in the following section.

4.3.1. Quantitative Results

Table 1 shows the score comparison of the proposed method and other perception-driven methods on each evaluation metric. The proposed Gram-GAN had the highest PSNR and SSIM values among all the methods and also had excellent LPIPS values. The Beby-GAN with single-LR–multiple-HR supervision also had high PSNR, SSIM and LPIPS values. However, its NIQE values were much worse than those of Gram-GAN, which indicated that the visual quality of images generated by Beby-GAN was much lower than Gram-GAN. ESRGAN had low NIQE values, but its PSNR, SSIM and LPIPS values were poor, indicating that the model focused on optimizing the visual quality of the predicted images at the expense of their facticity. The PSNR values reported by SFTGAN were relatively high, whereas its SSIM and LPIPS values were significantly worse than those of Gram-GAN. In conclusion, Gram-GAN showed better improvements in the disadvantage of perception-driven methods generally lacking facticity, and also retained the advantage of the high visual quality of perception-driven methods.

Table 1. Quantitative evaluation of state-of-the-art perception-driven methods.

DataSet	Metric	Bicubic	SRGAN [3]	ESRGAN [11]	SFTGAN [13]	ESRGAN+ [15]	Beby-GAN [17]	Gram-GAN (Ours)
Set5	PSNR	26.69	26.69	26.50	27.26	25.88	27.82	27.97
	SSIM	0.7736	0.7813	0.7565	0.7765	0.7511	0.8004	0.8021
	LPIPS	0.3644	0.1305	0.1080	0.1028	0.1178	0.0875	0.0867
	NIQE	29.56	24.58	18.75	26.87	19.45	25.40	21.34
Set14	PSNR	26.08	25.88	25.52	26.29	25.01	26.86	26.96
	SSIM	0.7467	0.7480	0.7175	0.7397	0.7159	0.7691	0.7710
	LPIPS	0.3870	0.1421	0.1254	0.1177	0.1362	0.1009	0.1003
	NIQE	25.22	18.60	15.19	16.71	16.09	18.45	17.27
BSD100	PSNR	26.07	24.65	24.95	25.71	24.62	26.13	26.32
	SSIM	0.7177	0.7063	0.6785	0.7065	0.6893	0.7347	0.7376
	LPIPS	0.4454	0.1622	0.1428	0.1357	0.1446	0.1192	0.1202
	NIQE	24.35	19.64	16.27	17.23	17.76	21.05	18.53
Urban100	PSNR	24.73	24.04	24.21	25.04	23.98	25.72	25.89
	SSIM	0.7101	0.7209	0.7045	0.7314	0.7182	0.7652	0.7679
	LPIPS	0.4346	0.1534	0.1354	0.1259	0.1334	0.1066	0.1076
	NIQE	20.63	14.93	12.52	13.12	13.38	15.76	14.28

The best performance is highlighted in red (best) and blue (second best).

4.3.2. Qualitative Results

Figures 4–7 show the comparison of Gram-GAN and other perception-driven methods in terms of visual effects. Gram-GAN was able to reconstruct texture details closer to GT than other methods. Specifically, Figure 4 highlights the pattern on the tiger. Gram-GAN generates the closest pattern to GT, whereas the other methods either have poor effects in pattern shape or have too many non-realistic artifacts. Figure 5 highlights the ground pattern in the distance. It can be found that the textures generated by all the methods except for Gram-GAN and Beby-GAN were distorted to some extent, and the advantage of Gram-GAN over Beby-GAN was that Gram-GAN could make the pattern in the backward position clear. Figure 6 highlights the lines on the ceiling, and all the methods except for Gram-GAN generated some pseudo lines. Figure 7 highlights the cross-stripes on the chair. Only the proposed Gram-GAN was able to generate dense and clear cross-stripes; the other methods could not achieve this effect.

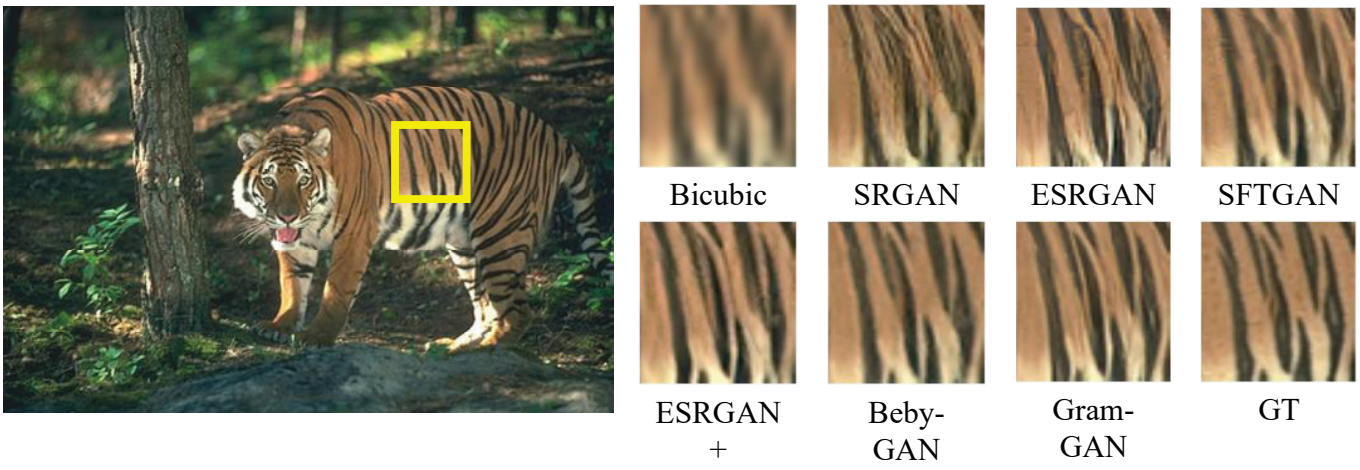

Figure 4. Visual evaluation of state-of-the-art perception-driven methods. Image "108005" from BSD100.

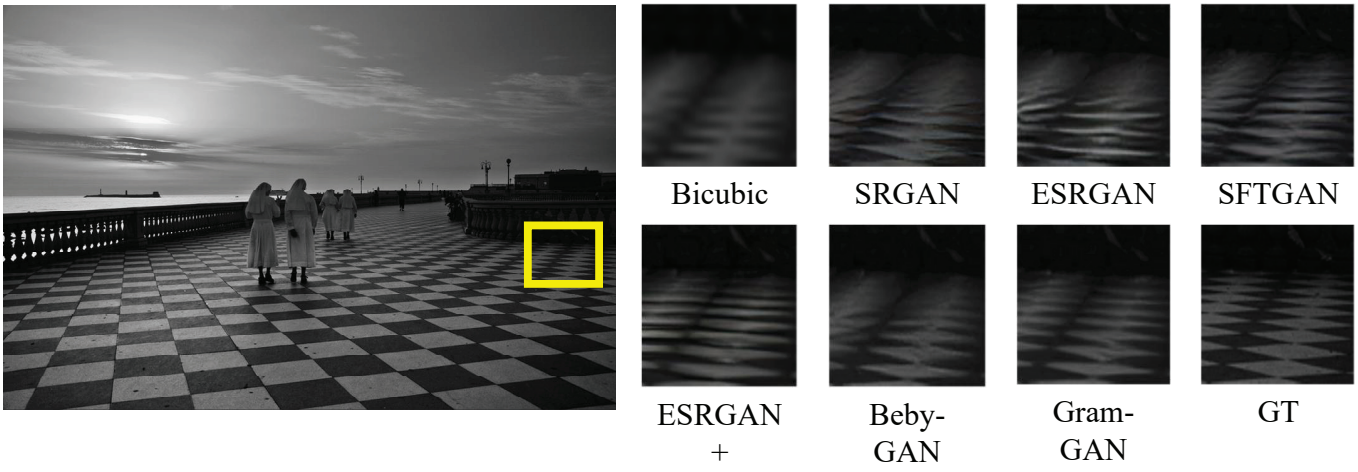

Figure 5. Visual evaluation of state-of-the-art perception-driven methods. Image "img_028" from Urban100.

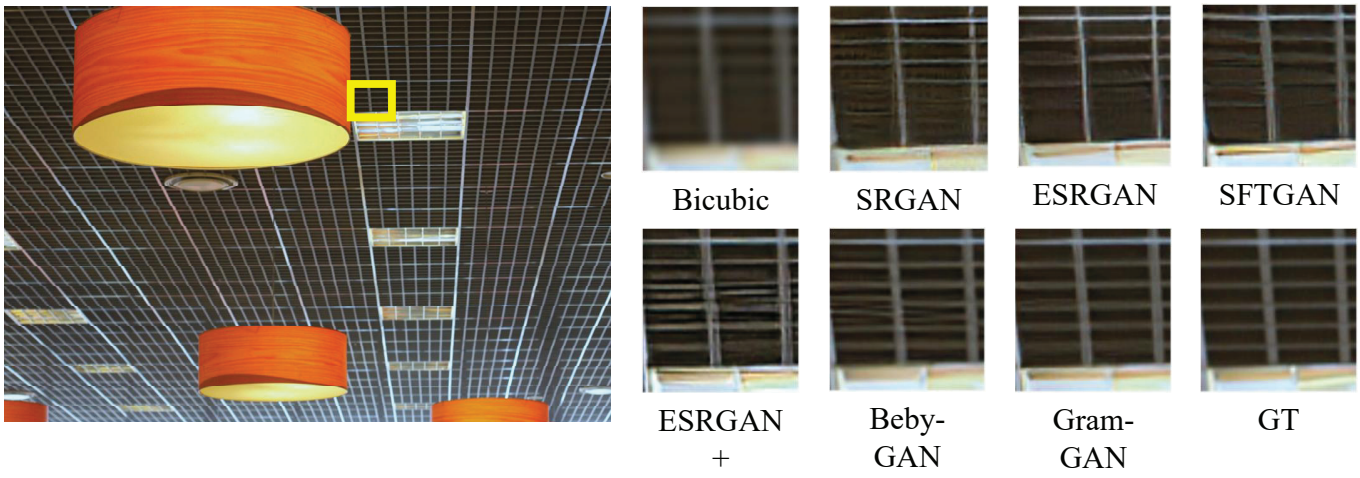

Figure 6. Visual evaluation of state-of-the-art perception-driven methods. Image "img_044" from Urban100.

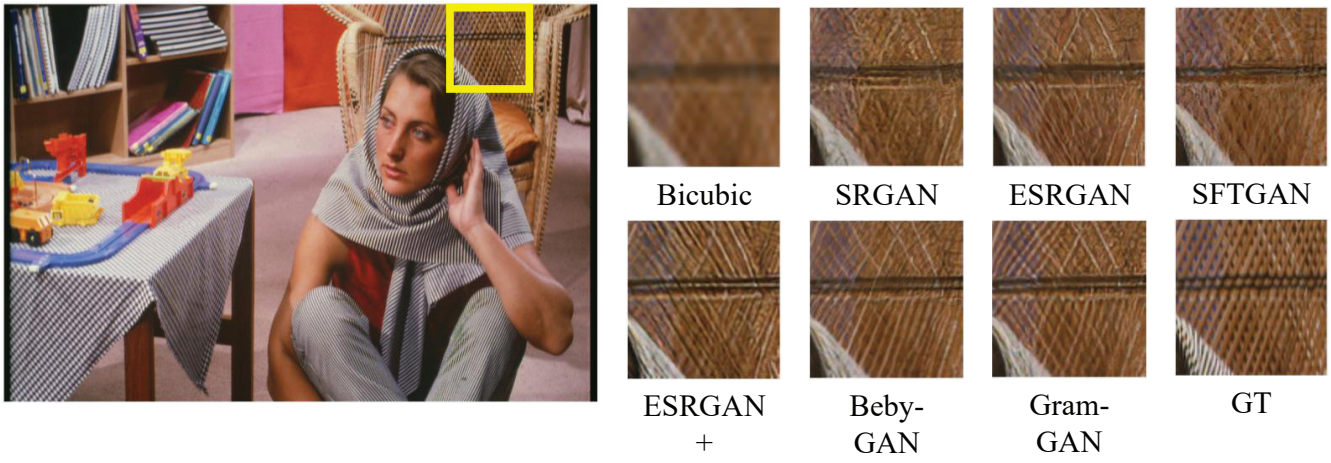

Figure 7. Visual evaluation of state-of-the-art perception-driven methods. Image "barbara" from Set14.

4.4. Ablation Study

In Table 2, we perform the ablation study by superimposing the losses. The initial loss function combination $\mathcal{L}_{PGC}$ references to [11], that is, the perceptual loss, content loss and adversarial loss were used. Both of the proposed losses brought some positive effects to the model. PSNR and SSIM values of the model were significantly improved and LPIPS values were reduced after the addition of the patch-wise texture loss. NIQE values of the model were substantially reduced and the values of PSNR and SSIM were further improved after the addition of discriminator perception loss. Figure 8 shows the related visual effects. With the superposition of the proposed loss, the structure of the reconstructed images became gradually close to that of GT images and artifacts arewere eliminated, resulting in higher facticity and visual quality.

Table 2. Quantitative evaluation of the proposed method under the ablation experiment.

Metric	$\mathcal{L}_{PGC}$	$L_{PGC} + \mathcal{L}_{PT}$	$\mathcal{L}_{PGC} + \mathcal{L}_{PT} + \mathcal{L}_{DP}$	Set5	Set14	BSD100	Urban100
PSNR	✓			27.72	26.69	26.06	25.59
	✓	✓		27.96	26.97	26.24	25.79
	✓	✓	✓	27.97	26.96	26.32	25.89
SSIM	✓			0.7967	0.7647	0.7290	0.7593
	✓	✓		0.8016	0.7709	0.7353	0.7654
	✓	✓	✓	0.8021	0.7710	0.7376	0.7679
LPIPS	✓			0.0891	0.1031	0.1215	0.1099
	✓	✓		0.0883	0.1021	0.1205	0.1079
	✓	✓	✓	0.0867	0.1003	0.1202	0.1076
NIQE	✓			22.21	18.34	19.32	14.70
	✓	✓		24.36	20.32	19.17	14.65
	✓	✓	✓	21.34	17.27	18.53	14.28

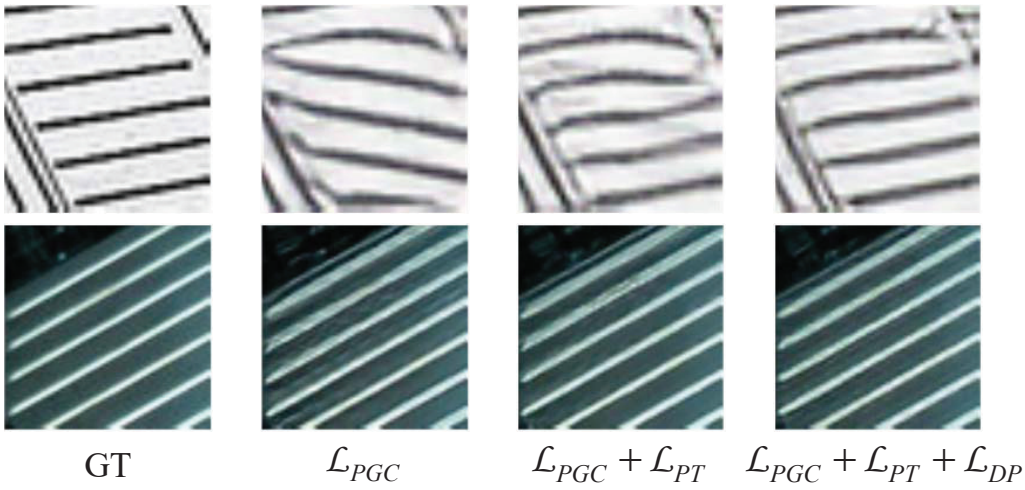

Figure 8. Visual evaluation of the proposed method under the ablation experiment.

5. Conclusions

For SR tasks with high visual quality requirements, we first constructed a novel supervision based on the Gram matrix to enhance the flexibility of model inference. Then, a discriminator perceptual loss specifically for SR tasks was proposed to enrich the feature types required for network inference. Finally, a large number of quantitative and qualitative experiments were conducted to verify the effectiveness of the proposed methods, and the necessity of each proposed loss was demonstrated through ablation studies. In future work, considering the high complexity of RRDB, we will focus on optimizing the computational complexity of networks and try to build a high-performance lightweight network.

Author Contributions: Conceptualization, J.S. and H.Y.; methodology, J.S. and H.Y.; software, J.S.; validation, W.X., X.L. and B.L.; formal analysis, J.S.; investigation, J.S.; resources, H.Y.; data curation, J.S.; writing—original draft preparation, J.S.; writing—review and editing, W.X.; visualization, J.S.; supervision, X.L.; project administration, W.X.; funding acquisition, H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the Natural Science Foundation of Liaoning Province, China (No. 2020-MS-292), the Scientific Research Foundation of Liaoning Provincial Education Department, China (No. JZL202015402), the Applied Foundation Research Project of Liaoning Province (No. 2022JH2/101300278), the Foundation Research Project of the Educational Department of Liaoning Province (No. LJKZZ20220085), and the Cooperation Innovation Plan of Yingkou for Enterprise and Doctor (No. 2022-13).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Our training DIV2K datasets can be obtained online: https://data.vision.ee.ethz.ch/cvl/DIV2K/ (accessed on 21 November 2022). Set5, Set14, BSD100 and Urban100 can be obtained from: https://arxiv.org/abs/1909.11856/ (accessed on 21 November 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chao, D.; Chen, C.L.; He, K.; Tang, X. Learning a Deep Convolutional Network for Image Super-Resolution. In Proceedings of the ECCV, Zurich, Switzerland, 6–12 September 2014.
2. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
3. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the IEEE Computer Society, Las Vegas, NV, USA, 27–30 June 2016.
4. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. *arXiv* **2018**, arXiv:1807.02758.
5. Hu, X.; Mu, H.; Zhang, X.; Wang, Z.; Tan, T.; Sun, J. Meta-SR: A Magnification-Arbitrary Network for Super-Resolution. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
6. Li, Z.; Yang, J.; Liu, Z.; Yang, X.; Wu, W. Feedback Network for Image Super-Resolution. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
7. Hussein, S.A.; Tirer, T.; Giryes, R. Correction Filter for Single Image Super-Resolution: Robustifying Off-the-Shelf Deep Super-Resolvers. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
8. Zhou, W.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In Proceedings of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 9–12 November 2003.
9. Gupta, P.; Srivastava, P.; Bhardwaj, S.; Bhateja, V. A modified PSNR metric based on HVS for quality assessment of color images. In Proceedings of the 2011 International Conference on Communication and Industrial Application, Kolkata, India, 26–28 December 2012.
10. Johnson, J.; Alahi, A.; Fei-Fei, L. *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*; Springer: Cham, Switzerland, 2016.
11. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Loy, C.C. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. *arXiv* **2018**, arXiv:1809.00219.
12. Rad, M.S.; Bozorgtabar, B.; Marti, U.V.; Basler, M.; Ekenel, H.K.; Thiran, J.P. SROBB: Targeted Perceptual Loss for Single Image Super-Resolution. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
13. Wang, X.; Yu, K.; Dong, C.; Loy, C.C. Recovering Realistic Texture in Image Super-resolution by Deep Spatial Feature Transform. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
14. Soh, J.W.; Gu, Y.P.; Jo, J.; Cho, N.I. Natural and Realistic Single Image Super-Resolution With Explicit Natural Manifold Discrimination. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
15. Rakotonirina, N.C.; Rasoanaivo, A. ESRGAN+: Further Improving Enhanced Super-Resolution Generative Adversarial Network. In Proceedings of the ICASSP 2020—IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3637–3641.
16. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
17. Li, W.; Zhou, K.; Qi, L.; Lu, L.; Jiang, N.; Lu, J.; Jia, J. Best-Buddy GANs for Highly Detailed Image Super-Resolution. *Proc. AAAI* **2022**, *36*, 1412–1420. [CrossRef]
18. Alex Krizhevsky, I.S.; Hinton, G.E. Best-Buddy GANs for Highly Detailed Image Super-Resolution. In Proceedings of the NeuriPS, Lake Tahoe, NV, USA, 3–6 December 2012.
19. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A Discriminative Feature Learning Approach for Deep Face Recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
20. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
22. Jolicoeur-Martinea, A. Deep residual learning for image recognition. In Proceedings of the ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.

23. Agustsson, E.; Timofte, R. Ntire 2017 challenge on single image super-resolution: Dataset and study. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 126–135.
24. Bevilacqua, M.; Roumy, A.; Guillemot, C.; Alberi-Morel, M.L. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In Proceedings of the 23rd British Machine Vision Conference (BMVC), Surrey, UK, 3–7 September 2012.
25. Zeyde, R.; Elad, M.; Protter, M. On single image scale-up using sparse-representations. In Proceedings of the International Conference on Curves and Surfaces, Avignon, France, 24–30 June 2010; Springer: Cham, Switzerland, 2010; pp. 711–730.
26. Martin, D.; Fowlkes, C.; Tal, D.; Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In Proceedings of the Eighth IEEE International Conference on Computer Vision, ICCV 2001, Vancouver, BC, Canada, 7–14 July 2001; Volume 2, pp. 416–423.
27. Huang, J.B.; Singh, A.; Ahuja, N. Single image super-resolution from transformed self-exemplars. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5197–5206.
28. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
29. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
30. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.* **2012**, *20*, 209–212. [CrossRef]