DISTILL THE PRIOR, DROP THE DIFFUSION: COMPACT FEED-FORWARD TEXT RESTORATION FROM DIFFUSION PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models provide powerful generative priors for image restoration, but transferring such knowledge beyond the diffusion architecture remains challenging. Can a restoration prior learned by a large diffusion model be absorbed by a fundamentally different feed-forward network, despite the mismatch in internal representations and inference mechanisms? Moreover, under aggressive model compression, can this prior still be retained rather than collapsing with student capacity? We study these questions in text image super-resolution and introduce PixelSTR, a diffusion-prior distillation framework that transfers restoration behavior entirely in output space. Instead of aligning diffusion latents or intermediate features, a lightweight pixel-space student learns the appearance and local structure of teacher outputs through RGB and spatial-gradient supervision, while paired high-resolution images and transcriptions provide reconstruction and recognition constraints. PixelSTR achieves over 100$\times$ parameter compression relative to the teacher restoration module. Under matched training, diffusion-derived supervision improves PSNR by 1.51 dB and OCR accuracy by at least 6.70 percentage points over supervised-only training, while replacing the diffusion targets with additional ground-truth targets does not reproduce these gains. These results suggest that generative restoration priors can be retained without retaining the generative model.

1 INTRODUCTION

Text image super-resolution (TextSR) aims to recover readable high-resolution text from severely degraded observations while preserving fine character structures. Compared with generic image restoration, TextSR often requires stronger restoration priors because low-resolution observations may contain insufficient local evidence to reconstruct missing or merged strokes. Existing methods therefore exploit sequence modeling, recognition guidance, text-focused attention, and generative structure priors to recover visually plausible and recognizable characters (Wang et al., 2020; Chen et al., 2021; Ma et al., 2022; Li et al., 2023). More recently, diffusion-based methods have demonstrated strong restoration capability under severe degradation, making them attractive sources of learned image and text priors (Zhang et al., 2024; Min et al., 2026).

However, using a strong diffusion prior typically ties restoration quality to the generative architecture that provides it. Existing efficiency-oriented methods mainly reduce sampling steps or simplify individual diffusion components, but the deployed model often remains structurally close to the original generative pipeline. This raises a more fundamental question: can restoration knowledge learned by a diffusion model be transferred to a fundamentally different feed-forward architecture? Such transfer is non-trivial because a diffusion teacher and a conventional pixel-space restorer may share neither latent representations nor intermediate features nor inference mechanisms. In our setting, the teacher relies on diffusion-specific latent and conditioning components, whereas the deployed student directly maps an RGB observation to a restored RGB image.

A second difficulty arises when this architectural transfer is combined with aggressive model compression. Even if a different feed-forward architecture can absorb diffusion-derived restoration knowledge, it is unclear how much student capacity is required to retain that knowledge. As the student becomes increasingly small, the transferred prior may gradually weaken or potentially deteriorate sharply once

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

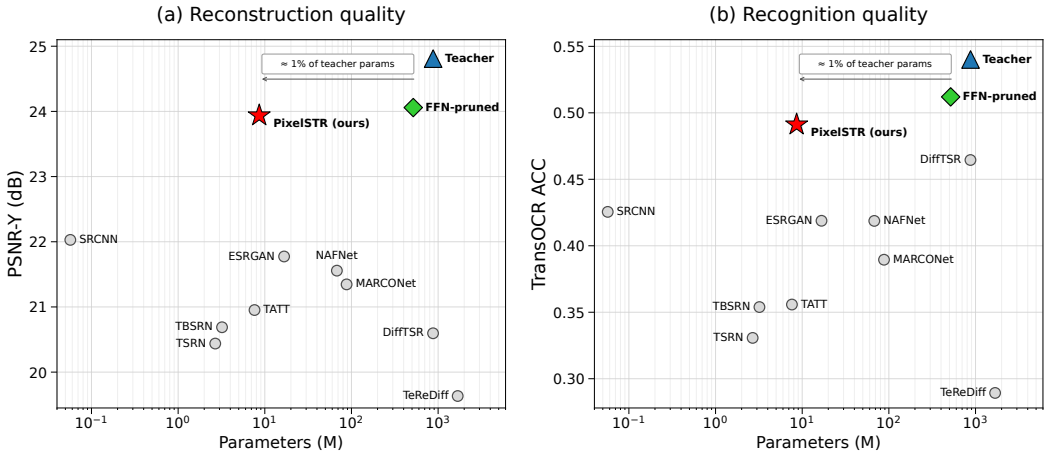

Figure 1: Quality–parameter trade-offs on CTR-TSR-Test ($\times 4$, 8,089 images). Panels show (a) PSNR and (b) TransOCR ACC against parameter count on a logarithmic axis. Teacher and baseline values match table 1; parameter counts follow table 3. The teacher and FFN-pruned counts cover their restoration modules, whereas the PixelSTR count covers the complete student.

the model can no longer represent the restoration behavior of the teacher. This issue is particularly relevant to TextSR, where limited capacity may first compromise subtle character structures and local boundaries rather than coarse image content. We therefore ask whether a diffusion restoration prior remains effective across both a large architectural gap and an extreme capacity gap?

We introduce PixelSTR to investigate these questions by decoupling the source of restoration knowledge from the model used at deployment. A frozen diffusion model serves purely as a training-time teacher, while a lightweight NAFNet performs pixel-space restoration at inference. To bridge their substantial architectural mismatch, PixelSTR avoids latent, feature, or trajectory alignment and instead transfers restoration behavior entirely in output space. RGB matching conveys appearance-level information, while spatial-gradient matching emphasizes local structures such as character strokes and boundaries. Paired high-resolution images and transcriptions further provide reconstruction, perceptual, and recognition supervision, allowing the student to absorb diffusion-derived restoration knowledge without reproducing the internal computation of the teacher.

Despite the large capacity gap, the 8.62M-parameter student uses less than 1% of the 876.23M parameters in the teacher restoration module while removing diffusion inference entirely. It achieves 23.94 dB PSNR and 0.4912 TransOCR accuracy, and a larger 18.98M student brings only marginal PSNR improvement. Diffusion-derived supervision further improves PSNR by 1.51 dB and OCR accuracy by at least 6.70 percentage points over supervised-only training, while additional ground-truth targets do not reproduce these gains. This suggests that diffusion priors can survive substantial architectural and capacity compression without retaining the generative model itself.

Our contributions are threefold:

- We investigate the decoupling of diffusion restoration priors from diffusion inference architectures, transferring learned restoration behavior to a fundamentally different lightweight feed-forward pixel-space network that requires no generative components at deployment.
- We introduce an appearance–structure diffusion-prior distillation framework for TextSR that combines teacher-derived RGB and spatial-gradient targets with paired image and transcription constraints to exploit generative restoration knowledge while preserving character fidelity.
- We provide controlled evidence that diffusion-derived supervision contains useful restoration information beyond additional ground-truth supervision. PixelSTR improves PSNR by 1.51 dB and OCR accuracy by at least 6.70 percentage points over matched supervised-only training, while the GT-target control does not reproduce these gains. The deployed model achieves over $100\times$ parameter compression relative to the teacher restoration module.

2 Related Work

Text-aware image super-resolution. Text-aware SR methods incorporate priors that help preserve character structure and identity. TSRN models text sequences (Wang et al., 2020), while TBSRN and TATT introduce text-focused attention and mechanisms for handling deformation (Chen et al., 2021; Ma et al., 2022). TPGSR and MARCONet use recognition and generative structure priors, respectively (Ma et al., 2023a; Li et al., 2023). More recent methods draw on diffusion priors: DiffTSR couples image and text diffusion (Zhang et al., 2024), TeReDiff incorporates recognition guidance (Min et al., 2026), and a recent one-step diffusion model removes iterative image–text refinement (Xu et al., 2026). PixelSTR investigates whether the resulting restoration behavior can be transferred to a substantially smaller non-diffusion network that discards the diffusion pipeline and auxiliary modules from the deployed system.

Fewer sampling steps and smaller generators. Efforts to reduce diffusion inference cost address both sampling and model architecture. Progressive distillation and consistency models reduce the number of sampling steps (Salimans & Ho, 2022; Song et al., 2023). ResShift, SinSR, OSEDiff, and AddSR pursue efficient diffusion-based restoration through shorter sampling procedures or one-step prediction (Yue et al., 2023; Wang et al., 2024; Wu et al., 2024; Tai et al., 2026). BK-SDM, TinySR, PocketSR, and GenDR-Pix reduce or replace computationally expensive diffusion and VAE components (Kim et al., 2024; Dong et al., 2026; Sun et al., 2025; Wang et al., 2026). PixelSTR transfers supervision from a diffusion model to a compact pixel-space restorer, allowing the entire diffusion pipeline to be removed at deployment.

Distillation through output targets. Knowledge distillation transfers model behavior by matching outputs or intermediate representations (Hinton et al., 2015; Romero et al., 2015). Diffusion2GAN uses paired diffusion outputs to train a conditional GAN student that retains the diffusion U-Net architecture (Kang et al., 2024). PixelSTR uses a structurally different student to map degraded RGB images directly to restored RGB images. The frozen diffusion model supplies online image targets, so the student does not need to reproduce diffusion-specific representations or denoising trajectories. This output-only interface also avoids layer-wise correspondence between the teacher and student.

Structure and recognition supervision. Gradient guidance helps preserve local image structure (Ma et al., 2020), while perceptual losses complement pixel-level reconstruction objectives (Johnson et al., 2016; Zhang et al., 2018). PixelSTR combines teacher-derived RGB and gradient targets with reconstruction, perceptual, and recognition losses based on paired GT images and transcriptions. Adversarial training can improve perceptual quality at the expense of fidelity to paired references (Ledig et al., 2017; Blau & Michaeli, 2018). We therefore report FID alongside similarity and OCR metrics.

3 Method

3.1 Training–inference separation and diffusion targets

Figure 2 illustrates how PixelSTR separates diffusion-based supervision from the deployed restoration architecture. During training, an HR image $\boldsymbol{x}_{\mathrm{HR}}$ is paired with a transcription $\boldsymbol{y}$. An online degradation pipeline transforms the image into a target-resolution input $\boldsymbol{x}_{\mathrm{IN}}$, which is shared by the frozen diffusion model and the student. For an LR test image $\boldsymbol{x}_{\mathrm{LR}}$ evaluated at $\times 4$ magnification, the input is $\boldsymbol{x}_{\mathrm{IN}} = \mathcal{B}_4(\boldsymbol{x}_{\mathrm{LR}})$, where $\mathcal{B}_4$ denotes bicubic upsampling to a 512×128 canvas (width $\times$ height). The training degradation and interpolation procedures are described in appendix A.

We use a frozen TOLA model (Xu et al., 2026) to generate one-step diffusion targets. Its architecture, adaptation procedure, text conditioning, and latent-space inference are adopted without modification from the cited work. At each student training iteration, the teacher performs one-step restoration with freshly sampled noise and without gradient tracking, producing a decoded RGB target $\boldsymbol{x}_{\mathrm{T}}$ online. PixelSTR uses only $\boldsymbol{x}_{\mathrm{T}}$ and its spatial gradients for distillation; diffusion latents, hidden features, predicted tokens, and confidence vectors are not passed to the student. This interface lets the 8.62M-parameter pixel-space network learn from the diffusion prior without adopting its architecture.

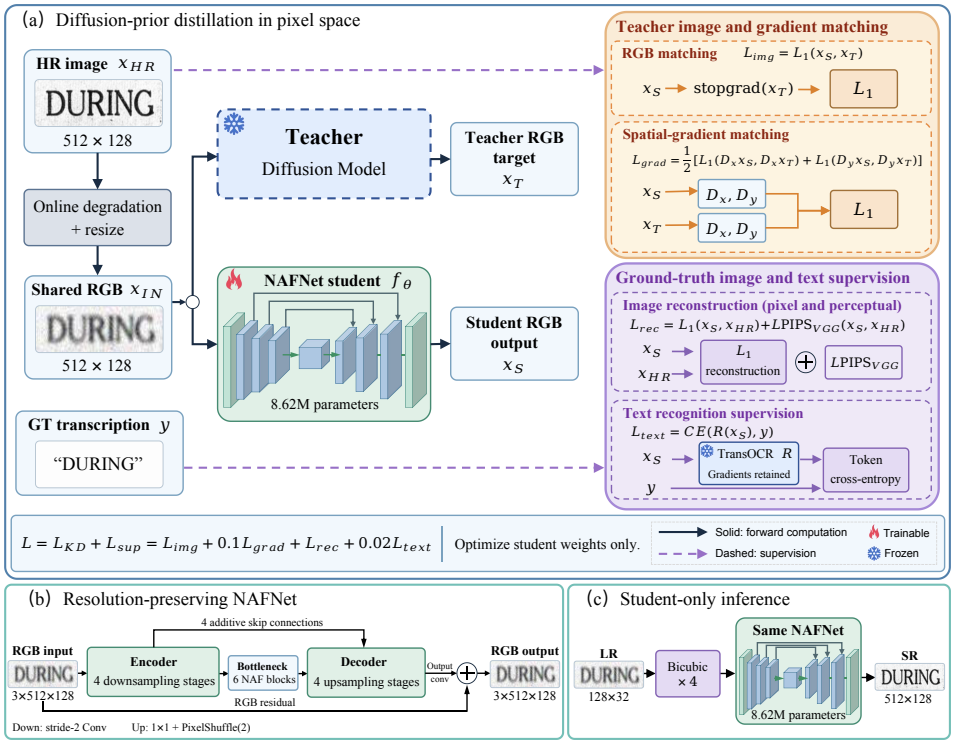

Figure 2: Overview of PixelSTR. The frozen diffusion teacher generates RGB targets online, while the NAFNet student is the only trainable restoration network. Teacher outputs supervise RGB appearance and spatial gradients, and paired HR images and GT transcriptions provide reconstruction, perceptual, and recognition supervision. The OCR network remains frozen but allows gradients to propagate to the student output. Inference uses bicubic upsampling and a single student forward pass.

3.2 Pixel-space student

The student predicts $\boldsymbol{x}_{\mathrm{S}} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{IN}})$ using a resolution-preserving NAFNet (Chen et al., 2022). The default configuration contains 8,624,619 parameters and uses a base width of 24, encoder block counts of $(1, 1, 2, 4)$, six middle blocks, and decoder block counts of $(1, 1, 1, 1)$. It combines stride-two downsampling, pixel-shuffle upsampling (Shi et al., 2016), skip connections, and a global RGB residual connection. At inference, it operates on the bicubic-upsampled canvas and requires no VAE, text conditioning, noise input, or iterative denoising. The larger variant increases the base width to 32 and the number of middle blocks to eight, for a total of 18,981,347 parameters. Full implementation details are provided in appendix A.

3.3 Learning from diffusion and ground-truth targets

Let $d_1(\boldsymbol{a}, \boldsymbol{b})$ denote the mean absolute element-wise difference between two tensors. We define gradient matching in terms of horizontal and vertical differences between adjacent pixels:

$$G(\boldsymbol{a}, \boldsymbol{b}) = \tfrac{1}{2}\big[d_1(\nabla_x \boldsymbol{a}, \nabla_x \boldsymbol{b}) + d_1(\nabla_y \boldsymbol{a}, \nabla_y \boldsymbol{b})\big]. \tag{1}$$

RGB matching encourages the student to reproduce the appearance of the decoded teacher output, while gradient matching constrains local intensity changes. The combined distillation objective is

$$\mathcal{L}_{\mathrm{KD}} = d_1(\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{x}_{\mathrm{T}}) + \lambda_{\mathrm{grad}} G(\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{x}_{\mathrm{T}}), \qquad \lambda_{\mathrm{grad}} = 0.1. \tag{2}$$

To constrain the restored image and character identity, we also supervise the student using the paired HR image and transcription:

$$\mathcal{L}_{\mathrm{sup}} = d_1(\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{x}_{\mathrm{HR}}) + \mathcal{L}^{\mathrm{VGG}}_{\mathrm{LPIPS}}(\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{x}_{\mathrm{HR}}) + \lambda_{\mathrm{ocr}} \mathcal{L}_{\mathrm{OCR}}(\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{y}), \qquad \lambda_{\mathrm{ocr}} = 0.02. \tag{3}$$

Algorithm 1 PixelSTR training with online targets from a frozen teacher

Require: HR image $\boldsymbol{x}_{\mathrm{HR}}$, transcription $\boldsymbol{y}$, frozen diffusion model D, student $f_{\boldsymbol{\theta}}$

1: Sample a degradation and an interpolation mode to obtain $\boldsymbol{x}_{\mathrm{IN}}$
2: Draw fresh diffusion noise and compute $\boldsymbol{x}_{\mathrm{T}} \leftarrow D(\boldsymbol{x}_{\mathrm{IN}})$ in one step without gradient tracking
3: Compute student output $\boldsymbol{x}_{\mathrm{S}} \leftarrow f_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{IN}})$
4: $\mathcal{L}_{\mathrm{KD}} \leftarrow d_1(\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{x}_{\mathrm{T}}) + 0.1G(\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{x}_{\mathrm{T}})$
5: $\mathcal{L}_{\mathrm{sup}} \leftarrow d_1(\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{x}_{\mathrm{HR}}) + \mathcal{L}^{\mathrm{VGG}}_{\mathrm{LPIPS}}(\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{x}_{\mathrm{HR}}) + 0.02\mathcal{L}_{\mathrm{OCR}}(\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{y})$
6: Update $\boldsymbol{\theta}$ by minimizing $\mathcal{L} = \mathcal{L}_{\mathrm{KD}} + \mathcal{L}_{\mathrm{sup}}$

During training, LPIPS is computed using a frozen VGG-based network (Zhang et al., 2018). The OCR loss is token-level cross-entropy computed by a frozen TransOCR Transformer, with teacher forcing based on the GT transcription. Recognition targets are the annotated text. The recognizer parameters remain fixed, but gradients propagate through the recognizer and the student output to update the student parameters. The student output is resized to 256×32 before recognition during training. Evaluation uses the benchmark TransOCR and PaddleOCR protocols in section 4.1.

The student minimizes $\mathcal{L} = \mathcal{L}_{\mathrm{KD}} + \mathcal{L}_{\mathrm{sup}}$. The two sources of supervision serve complementary roles: diffusion matching transfers the learned restoration behavior of the teacher, while GT supervision constrains fidelity to the paired image and transcription. Algorithm 1 summarizes one iteration.

Controlled alternatives. Supervised-only training minimizes $\mathcal{L}_{\mathrm{sup}}$, omitting both distillation terms in equation 2. The GT-target control retains the additional matching terms but replaces $\boldsymbol{x}_{\mathrm{T}}$ with $\boldsymbol{x}_{\mathrm{HR}}$, yielding

$$\mathcal{L}_{\text{GT-target}} = 2d_1(\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{x}_{\mathrm{HR}}) + 0.1G(\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{x}_{\mathrm{HR}}) + \mathcal{L}^{\mathrm{VGG}}_{\mathrm{LPIPS}}(\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{x}_{\mathrm{HR}}) + 0.02\mathcal{L}_{\mathrm{OCR}}(\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{y}). \quad (4)$$

GT-target therefore adds both a second GT reconstruction term and a GT gradient constraint to the supervised-only objective. Additional controls individually remove gradient matching or OCR supervision from the full objective. All variants use the same CTR training split, 50K optimization steps, and final-checkpoint evaluation to support comparisons under a matched training budget.

4 Experiments

4.1 Experimental setup

Datasets. Student training follows the CTR-TSR data construction and online degradation pipeline used in DiffTSR (Zhang et al., 2024), with TOLA (Xu et al., 2026) as the frozen diffusion teacher. The preprocessed dataset contains 62,371 training images and 1,273 held-out images. We evaluate final checkpoints on all 8,089 pairs in CTR-TSR-Test for $\times 4$ super-resolution. For cross-dataset evaluation, we use 200 filtered and deduplicated RealCE validation pairs constructed from the official 13 mm and 52 mm camera crops (Ma et al., 2023b). RealCE is excluded from training, loss-weight tuning, and checkpoint selection. We directly evaluate the same CTR-trained checkpoints on RealCE-200 for zero-shot transfer.

Metrics and comparators. On CTR, we report luminance-channel PSNR and SSIM without border cropping (Wang et al., 2004), LPIPS-AlexNet (Zhang et al., 2018), DISTS (Ding et al., 2022), FID (Heusel et al., 2017), and recognition accuracy (ACC) and normalized edit-distance similarity (NED). T and P denote TransOCR and PaddleOCR, respectively (Du et al., 2020). Higher ACC and NED indicate better recognition. TransOCR supplies recognition supervision during training, whereas PaddleOCR is used only for evaluation. The main RealCE table follows the six-metric reference protocol; the appendix reports additional LPIPS and DISTS results for the student variants.

We compare against SRCNN (Dong et al., 2014), ESRGAN (Wang et al., 2018), NAFNet, TSRN, TBSRN, TATT, MARCONet, DiffTSR, and TeReDiff. The non-diffusion baselines are retrained on the CTR split. All methods, including the teacher, the FFN-pruned reference, and the students, are evaluated on the same test manifests using a shared metric implementation.

Implementation. Each student variant is trained from scratch for 50K steps using AdamW (Loshchilov & Hutter, 2019), FP32 precision, a batch size of 16, and a learning rate

Table 1: Results on all 8,089 CTR-TSR-Test pairs for $\times 4$ super-resolution under a shared protocol. T and P denote TransOCR and PaddleOCR; †FFN pruning retains the teacher diffusion architecture.

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	T-ACC↑	T-NED↑	P-ACC↑	P-NED↑
SRCNN	22.0288	0.5860	0.6021	0.3886	156.9406	0.4255	0.6107	0.3303	0.5279
ESRGAN	21.7741	0.5588	0.6455	0.3979	155.7055	0.4188	0.6061	0.3287	0.5273
NAFNet	21.5568	0.5678	0.5924	0.3776	144.9606	0.4187	0.5998	0.2998	0.4870
TSRN	20.4381	0.5840	0.6768	0.3926	140.2070	0.3307	0.5192	0.2238	0.4196
TBSRN	20.6883	0.5840	0.6510	0.3918	137.1033	0.3539	0.5425	0.2574	0.4563
TATT	20.9523	0.5846	0.6614	0.3896	125.6047	0.3558	0.5450	0.2491	0.4466
MARCONet	21.3460	0.6132	0.4386	0.3566	102.7384	0.3895	0.5735	0.2796	0.4600
DiffTSR	20.5958	0.6261	0.3103	0.2428	24.9120	0.4645	0.6519	0.3970	0.5873
TeReDiff	19.6333	0.5125	0.4714	0.3466	77.0387	0.2893	0.4491	0.1972	0.3592
Teacher	24.8095	0.7235	0.2288	0.1953	12.6849	0.5401	0.7370	0.4626	0.6585
FFN-pruned†	24.0587	0.6957	0.2686	0.2124	16.7049	0.5121	0.7060	0.4146	0.6190
PixelSTR (ours)	23.9373	0.6863	0.2914	0.2238	23.8595	0.4912	0.6913	0.3984	0.6000

of 2×10^{-4}. All comparisons use final checkpoints, with the same designated default checkpoint used in every single-checkpoint Full row. We also train Full and supervised-only models in three independent runs and report their means and standard deviations in table 5. Paired bootstrap intervals quantify uncertainty in CTR TransOCR differences for fixed checkpoints, while the independent runs characterize variability across training runs. Optimization uses a 500-step linear warm-up and gradient clipping at a global norm of 1.0. The appendix provides the complete configuration.

4.2 Main results

CTR reconstruction and recognition. As shown in Table 1, the default student achieves 23.9373 dB PSNR and 0.4912 TransOCR accuracy, trailing the teacher by 0.87 dB and 4.89 percentage points, respectively, while using 0.98% of its restoration-module parameters. The teacher achieves the best results on all reported quality metrics, while PixelSTR substantially reduces inference cost.

Compared with the nine standard baselines, PixelSTR achieves the best values on all nine reported metrics. Relative to DiffTSR, it improves PSNR by 3.34 dB and TransOCR ACC by 2.67 percentage points, while reducing LPIPS from 0.3103 to 0.2914 and FID from 24.9120 to 23.8595. PaddleOCR ACC increases from 0.3970 to 0.3984 and NED from 0.5873 to 0.6000, although the ACC gain is only 0.14 percentage points. The matched student controls below assess the contribution of the training objective under a fixed architecture and training budget.

CTR visual comparisons. Figure 3 presents aligned examples of numeric, Chinese, English, and mixed-language text. The student recovers character structure without diffusion inference, although errors remain in narrow strokes and heavily degraded regions. These examples complement the image-quality and OCR results on the complete test set.

RealCE cross-dataset performance. On RealCE-200, the same CTR-trained checkpoint achieves 19.2071 dB PSNR, 0.3850 T-ACC, and 0.5000 P-ACC (table 2). It outperforms DiffTSR on all six reported metrics, with gains of 1.58 dB PSNR and 3.5 and 7.0 percentage points in TransOCR and PaddleOCR accuracy, respectively. Using 0.98% of the teacher restoration-module parameters, the student trails the teacher by 0.0169 dB PSNR and 1.0 percentage point under both recognizers.

Compared with the nine standard baselines, PixelSTR achieves the best or tied-best values on five of the six RealCE metrics. Its T-ACC matches SRCNN and ESRGAN, and its P-ACC matches ESRGAN; its PSNR advantage over ESRGAN is only 0.0009 dB. MARCONet achieves higher SSIM (0.6198 versus 0.6102). Bicubic upsampling also performs well on this benchmark: Full improves PSNR by only 0.0159 dB, matches its T-ACC, and increases P-ACC by 0.5 percentage points, while reducing LPIPS and DISTS. The complete component, capacity, GAN, and bicubic comparisons appear in appendix E, with qualitative examples in figure 4.

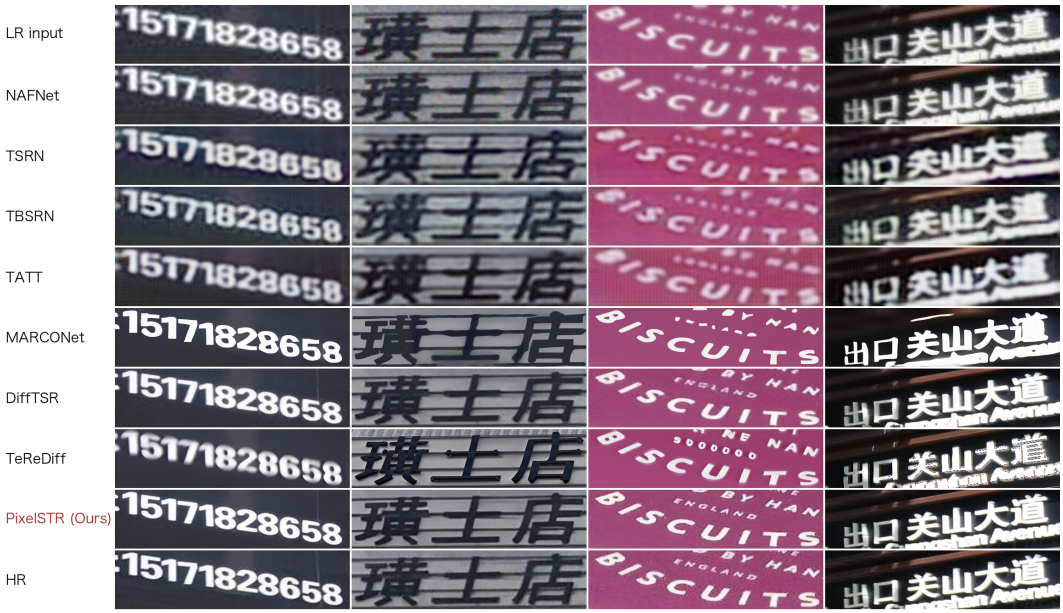

Figure 3: Selected qualitative comparisons on CTR-TSR-Test ($\times 4$), covering numeric, Chinese, English, and mixed-language text. Rows show the LR input, baseline outputs, PixelSTR, and the HR reference. Appendix F compares student, teacher, and FFN-pruned outputs on matched inputs.

Table 2: Results on 200 RealCE-200 camera-image pairs. All learned models use CTR-trained checkpoints without adaptation to RealCE; Full uses the same checkpoint as in table 1.

Method	PSNR↑	SSIM↑	T-ACC↑	T-NED↑	P-ACC↑	P-NED↑
SRCNN	19.1609	0.6024	0.3850	0.5837	0.4950	0.6670
ESRGAN	19.2062	0.5971	0.3850	0.5799	0.5000	0.6710
NAFNet	18.9038	0.5920	0.3650	0.5704	0.4750	0.6406
TSRN	18.6767	0.6009	0.2850	0.5103	0.3850	0.5649
TBSRN	18.9903	0.5947	0.3400	0.5364	0.3950	0.5980
TATT	18.9725	0.5954	0.3300	0.5291	0.4000	0.5970
MARCONet	18.3506	0.6198	0.3500	0.5408	0.3800	0.5972
DiffTSR	17.6287	0.5970	0.3500	0.5402	0.4300	0.6188
TeReDiff	17.8180	0.5272	0.2900	0.4826	0.3350	0.5508
Teacher	19.2240	0.6199	0.3950	0.5841	0.5100	0.6720
PixelSTR (ours)	19.2071	0.6102	0.3850	0.5839	0.5000	0.6717

4.3 Inference cost and latency

PixelSTR uses 8.62M parameters and 4.95 GMACs versus 876.23M and 297.50 GMACs for the teacher restoration module (table 3) and removes its VAE and text modules at inference. Under the shared forward protocol, PixelSTR takes 8.36 ms versus 131.9 ms for the teacher, giving approximately $101.7\times$ fewer restoration parameters, $60.1\times$ fewer MACs, and a $15.8\times$ forward speedup. FFN pruning retains the diffusion pipeline; measurement boundaries are in appendix H.

4.4 Ablations and diagnostic analyses

Diffusion supervision and the GT-target alternative. Table 4 evaluates diffusion-derived targets under a fixed architecture and 50K-step budget. Removing both matching terms reduces PSNR by 1.51 dB and TransOCR and PaddleOCR accuracy by 6.70 and 8.03 percentage points, respectively.

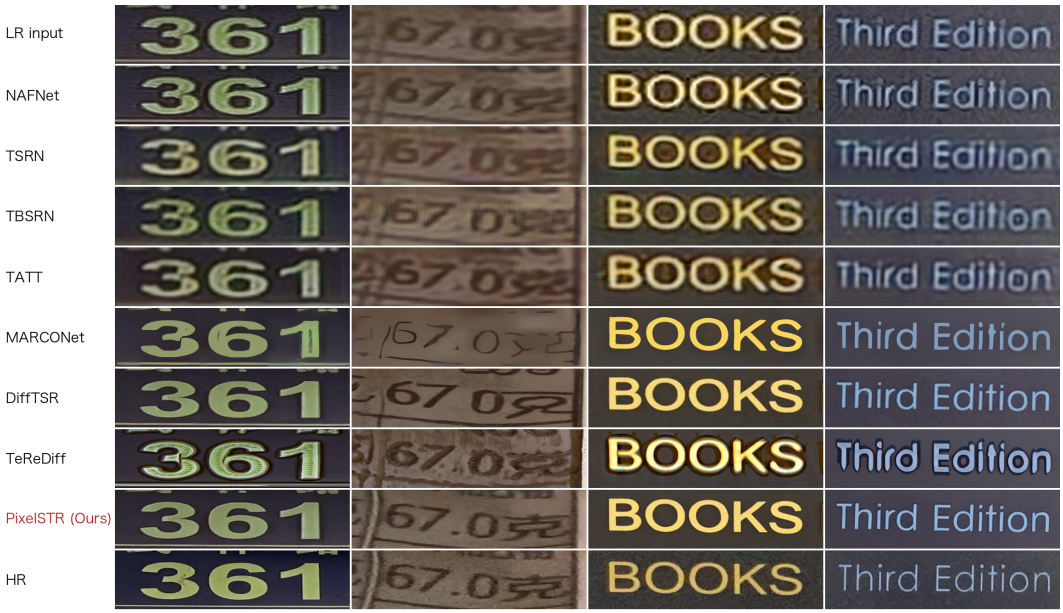

Figure 4: RealCE-200 qualitative comparisons. Rows show LR, baselines, PixelSTR, and HR; PixelSTR uses the CTR-trained checkpoint without RealCE adaptation. These examples illustrate zero-shot transfer from CTR training to real camera images under unseen capture conditions and domain-specific degradations.

Table 3: Efficiency at 512×128, batch size one, FP32 on an RTX PRO 6000 under the shared forward protocol. Teacher/FFN-pruned counts cover restoration modules; student counts cover the complete NAFNet. Timing excludes file I/O and external bicubic resizing; full boundaries are in appendix H.

Method	Params (M)	MACs (G)	Time (s)
SRCNN	0.057	3.748	0.0119
ESRGAN	16.70	73.453	0.0335
NAFNet	67.89	63.195	0.0324
TSRN	2.68	0.904	0.0148
TBSRN	3.21	2.535	0.0209
TATT	7.61	1.269	0.0287
MARCONet	87.90	466.178	0.1986
DiffTSR	874.00	59213.970	9.9548
TeReDiff	1682.54	26839.909	3.8499
Teacher	876.23	297.497	0.1319
FFN-pruned	516.73	197.683	0.07029
Larger student	18.98	9.495	0.01108
PixelSTR (ours)	8.62	4.952	0.00836

Replacing them with GT reduces PSNR by 1.42 dB and the two accuracies by 6.49 and 8.13 points. Thus, additional GT image and gradient penalties do not reproduce the diffusion-supervision gains, supporting the joint use of diffusion-derived RGB and structural targets beyond GT penalties.

Full minus supervised-only is 6.70 pp in TransOCR ACC (paired 95% CI [5.97, 7.44]) and 8.03 pp in PaddleOCR ACC; Table 5 shows consistent gains across independent runs.

Training repeatability. Across three independent 50K-step runs, Full reaches 23.9287 dB mean PSNR versus 22.3619 dB for supervised-only and also has higher mean accuracy under both OCR evaluators; table 5 reports the corresponding standard deviations.

Table 4: CTR ablations: 8,089 images, 50K steps each; Full matches table 1 and GT-target uses GT.

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	T-ACC↑	T-NED↑	P-ACC↑	P-NED↑
PixelSTR (Full)	**23.9373**	**0.6863**	**0.2914**	**0.2238**	**23.8595**	**0.4912**	**0.6913**	**0.3984**	**0.6000**
Supervised only	22.4290	0.6033	0.4722	0.3161	86.2946	0.4242	0.6105	0.3181	0.5104
GT-target	22.5162	0.6074	0.4708	0.3146	85.1284	0.4263	0.6141	0.3171	0.5091
w/o Gradient	22.5012	0.6059	0.4656	0.3078	82.8581	0.4243	0.6129	0.3182	0.5129
Gradient-only	22.4352	0.6022	0.4704	0.3134	84.5313	0.4254	0.6099	0.3172	0.5088
w/o OCR	22.5824	0.6138	0.4426	0.2984	73.2668	0.4204	0.6091	0.3145	0.5106

Table 5: CTR repeatability: mean $\pm$ standard deviation over three independent 50K-step runs.

Model	Runs	PSNR (dB)↑	T-ACC↑	P-ACC↑
PixelSTR (Full)	3	23.9287 ± 0.06	0.4881 ± 0.0056	0.3970 ± 0.0013
Supervised only	3	22.3619 ± 0.08	0.4232 ± 0.0023	0.3178 ± 0.0003

Gradient and OCR supervision. Without gradient or OCR supervision, CTR PSNR/T-ACC falls to 22.5012/0.4243 or 22.5824/0.4204. Gradient-only reaches 22.4352/0.4254, near supervised-only, versus Full at 23.9373/0.4912. Neither diffusion-derived target alone reproduces the gains; appendix D examines TT/TG/GT/GG interactions.

On RealCE, Full leads core controls in PSNR and recognition. Without OCR supervision, LPIPS falls from 0.4337 to 0.3991 while T-ACC/P-ACC falls from 0.3850/0.5000 to 0.3500/0.4500, showing a recognition–LPIPS trade-off (appendix E).

Student capacity. The 18.98M student raises T-ACC/P-ACC from 0.4912/0.3984 to 0.5017/0.4007 with only a 0.0088 dB PSNR gain. It uses 9.50 versus 4.95 GMACs, mainly improving perceptual and recognition metrics.

Adversarial objective. At $\lambda = 0.05$, patch-GAN lowers CTR FID from 23.8595 to 23.4797 but also PSNR and both OCR accuracies. The paired TransOCR interval and a 1.47-point PaddleOCR lead favor Full; smaller weights have higher FID. We retain non-adversarial Full; see Tables 11–12.

5 Conclusion and Limitations

We presented PixelSTR, which transfers a diffusion restoration prior from a one-step diffusion teacher to a compact feed-forward network. Teacher-derived RGB and spatial-gradient targets convey appearance and character structure, while paired HR images and transcriptions constrain reconstruction and recognition. The resulting 8.62M-parameter student uses 0.98% of the teacher restoration-module parameters and achieves a $15.8\times$ speedup under the measured forward-pass protocol, with inference reduced to bicubic upsampling and a single pixel-space forward pass. Under matched CTR training, diffusion supervision improves PSNR by 1.51 dB and TransOCR and PaddleOCR accuracy by 6.70 and 8.03 percentage points over supervised-only training. Component ablations and the GT-target control show that jointly matching teacher appearance and structure provides gains that additional GT supervision alone does not reproduce. Repeated CTR training, independent PaddleOCR evaluation, and zero-shot transfer to RealCE-200 further support the reconstruction and recognition benefits of this transfer. These findings show that useful diffusion-derived restoration behavior can survive substantial architectural and capacity compression, separating the model that learns the restoration prior from the model that applies it at inference.

This study is limited to one diffusion-teacher family, two student capacities, repeated runs for the principal CTR comparison, and 200 RealCE crops from 73 source images. Training also requires the frozen teacher and recognizer; broader validation across teachers, capacities, and real-capture settings remains necessary.

AI USE STATEMENT

Generative AI tools were used to aid and polish manuscript writing and for literature retrieval and discovery. The authors verified the resulting edits and cited sources and take responsibility for the final manuscript.

ETHICS STATEMENT

This work focuses on text image super-resolution and does not involve human-subject studies or the collection of personally identifiable information.

REPRODUCIBILITY STATEMENT

Sections 3 and 4.1 describe the method and evaluation protocol. Appendices A–H provide details of implementation, target generation, evaluation, uncertainty analysis, and efficiency measurement, including results across independent training runs. Student inference does not require the diffusion model.

REFERENCES

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018.

Jingye Chen, Bin Li, and Xiangyang Xue. Scene text telescope: Text-focused scene image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12026–12035, 2021.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision*, volume 13667, pp. 17–33, 2022.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, volume 8692, pp. 184–199, 2014.

Linwei Dong, Qingnan Fan, Yuhang Yu, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. TinySR: Shallow diffusion transformers for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 5075–5085, 2026.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. PP-OCR: A practical ultra lightweight OCR system. arXiv preprint arXiv:2009.09941, 2020. URL https://arxiv.org/abs/2009.09941.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015. URL https://arxiv.org/abs/1503.02531.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, volume 9906, pp. 694–711, 2016.

Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling diffusion models into conditional GANs. In *European Conference on Computer Vision*, pp. 428–447, 2024.

Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. BK-SDM: A lightweight, fast, and cheap version of Stable Diffusion. In *European Conference on Computer Vision*, volume 15112, pp. 381–399, 2024.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4681–4690, 2017.

Xiaoming Li, Wangmeng Zuo, and Chen Change Loy. Learning generative structure prior for blind text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10103–10113, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7769–7778, 2020.

Jianqi Ma, Zhetong Liang, and Lei Zhang. A text attention network for spatial deformation robust scene text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5911–5920, 2022.

Jianqi Ma, Shi Guo, and Lei Zhang. Text prior guided scene text image super-resolution. *IEEE Transactions on Image Processing*, 32:1341–1353, 2023a.

Jianqi Ma, Zhetong Liang, Wangmeng Xiang, Xi Yang, and Lei Zhang. A benchmark for Chinese-English scene text image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19452–19461, 2023b.

Jaewon Min, Jin Hyeon Kim, Paul Hyunbin Cho, Jaeeun Lee, Jihye Park, Minkyu Park, Sangpil Kim, Hyunhee Park, and Seungryong Kim. Text-aware image restoration with diffusion models. In *International Conference on Learning Representations*, 2026.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883, 2016.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252, 2023.

Haoze Sun, Linfeng Jiang, Fan Li, Renjing Pei, Zhixin Wang, Yong Guo, Jiaqi Xu, Haoyu Chen, Jin Han, Fenglong Song, Yujiu Yang, and Wenbo Li. PocketSR: The super-resolution expert in your pocket mobiles. In *Advances in Neural Information Processing Systems*, volume 38, pp. 103450–103468, 2025.

Ying Tai, Rui Xie, Chen Zhao, Kai Zhang, Zhenyu Zhang, Jun Zhou, and Jian Yang. AddSR: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. *Pattern Recognition*, 175:113012, 2026.

Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *European Conference on Computer Vision*, pp. 650–666, 2020.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 63–79, 2018.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1905–1914, 2021.

Yan Wang, Shijie Zhao, Junlin Li, and Li Zhang. Eliminating VAE for fast and high-resolution generative detail restoration. In *International Conference on Learning Representations*, 2026.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. SinSR: Diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25796–25805, 2024.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems*, volume 37, pp. 92529–92553, 2024.

Yike Xu, Yue Shi, Yong Guo, and Jiezhang Cao. TOLA: Text-aware one-step latent adaptation for diffusion-based text image super-resolution. arXiv preprint arXiv:2609.29240, 2026. URL `https://arxiv.org/abs/2609.29240`.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. ResShift: Efficient diffusion model for image super-resolution by residual shifting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 13294–13307, 2023.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4791–4800, 2021.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Yuzhe Zhang, Jiawei Zhang, Hao Li, Zhouxia Wang, Luwei Hou, Dongqing Zou, and Liheng Bian. Diffusion-based blind text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25827–25836, 2024.

A TRAINING AND IMPLEMENTATION DETAILS

A.1 STUDENT OPTIMIZATION AND DATA PROCESSING

Table 6 summarizes the student training configuration. Full, supervised-only, GT-target, the gradient and OCR ablations, the GAN variants, and the 19M model are each trained for 50K optimization steps. Single-checkpoint results use designated final checkpoints without test-set selection. The same default checkpoint is used for CTR, RealCE, and every single-checkpoint Full result. Results across independent runs of Full and supervised-only are reported separately in table 5.

Table 6: Training settings for the PixelSTR student variants.

Setting	Value
Training budget / batch size	50,000 steps / 16 images
Optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01
Learning rate	2×10^{-4}; 500-step linear warmup, then constant
Precision / gradient clipping	FP32 / global norm 1.0
Data-loader workers	4
Default student	Width 24; encoder $(1, 1, 2, 4)$; middle 6; decoder $(1, 1, 1, 1)$
Larger student	Width 32; middle 8; same encoder/decoder block counts
Diffusion targets	Online frozen one-step model, fixed timestep 999, fresh noise
Teacher loss weights	RGB 1.0 / gradient 0.1
GT loss weights	L1 1.0 / LPIPS 1.0 / OCR 0.02
GAN λ / D LR	$0.01, 0.025, 0.05$ / 10^{-4}
Student adversarial warmup	First 1,000 steps without the student adversarial term
Checkpoint selection	Final checkpoint; no validation-based selection

Data construction. The diffusion model and student use the same preprocessed CTR dataset and training split. A seeded split reserves 2% of the 63,644 images, yielding 1,273 held-out images and 62,371 training images. The held-out partition is not used to select student checkpoints. Each HR image is resized to a 512×128 canvas. The online pipeline selects Real-ESRGAN or BSRGAN degradation with equal probability and a scale from $\{1, 2, 4\}$ (Wang et al., 2021; Zhang et al., 2021). The degraded image is then resized to the target canvas using a sampled linear, cubic, or Lanczos interpolation mode. Thus, the diffusion model and student receive the same tensor at the target resolution, with training interpolation sampled from these three modes.

During benchmark inference, LR crops are loaded as RGB images and resized to 512×128 using bicubic interpolation. The student receives a tensor with values in $[0, 1]$. Its output is clamped, rounded, and saved as an 8-bit PNG before image-quality and OCR evaluation. Paired-image metrics are computed for each image and then averaged; FID is computed across the dataset. Training losses are computed from the unclamped student output. The CTR and RealCE manifests contain 8,089 and 200 matched pairs, respectively, and every entry is included in evaluation.

Randomness and repeatability. Comparisons between fixed checkpoints use the same training split, degradation procedure, and optimization settings. Network construction, discriminator initialization, and changes to the OCR branch can alter the sequence of random-number draws. Initializing random-number generators identically therefore does not guarantee identical minibatches or diffusion noise across variants. Full and supervised-only are each trained in three independent runs with the same budget. Variation across these runs is reported separately from bootstrap analyses of the designated fixed checkpoints.

Loss reduction and gradients. Image and finite-difference losses use mean absolute differences. The horizontal and vertical gradient losses are averaged and then multiplied by 0.1. Training LPIPS uses a VGG backbone with inputs mapped to $[-1, 1]$. The frozen OCR module receives the student output without detachment, allowing recognition-loss gradients to propagate through the image to the student. Individual loss values are logged before weighting, while the KD total and full objective

include the configured weights. Supervised-only removes both diffusion-matching terms. GT-target uses the GT image for both image and gradient matching. The OCR ablation removes the recognition loss and retains all remaining terms.

A.2 Adversarial and structural-compression references

The adversarial variants use a pix2pix-style patch discriminator (Isola et al., 2017) with 4×4 convolutions, channel widths $(64, 128, 256, 512)$, strides $(2, 2, 2, 1)$, batch normalization in all blocks except the first, and a LeakyReLU negative slope of 0.2. A final stride-one convolution gives each output a 70×70 receptive field. The discriminator receives output images without concatenated LR conditioning. During discriminator updates, detached diffusion outputs serve as real samples and detached student outputs serve as fake samples. The discriminator and non-saturating generator objectives use binary cross-entropy with logits. A separate AdamW optimizer updates the discriminator with a learning rate of 10^{-4} and $\boldsymbol{\beta} = (0.9, 0.999)$. The student adversarial term is enabled after an initial 1,000-step warmup. The diffusion model remains frozen throughout training.

FFN pruning provides a structural-compression reference within the diffusion teacher architecture. The pruned model retains VAE processing and text conditioning and is evaluated at its designated checkpoint. It contains 516.73M restoration parameters and requires 197.68 GMACs under the IDM/LRRB accounting used for the diffusion model. This experiment evaluates a single pruning configuration, so it does not characterize the full quality–compression trade-off or provide a parameter-matched comparison with the 8.62M student.

B Diffusion-model interface

We use TOLA (Xu et al., 2026) as the frozen one-step diffusion model throughout distillation. Its architecture, adaptation procedure, confidence-aware conditioning, latent conventions, and one-step restoration path follow the cited work without modification. At each training iteration, the diffusion model receives the same target-resolution input as the student and generates an online target using fresh noise. Only the decoded RGB restoration $\boldsymbol{x}_{\mathrm{T}}$ is passed to PixelSTR for target construction.

Algorithm 2 Online teacher-target construction

Require: Target-resolution input $\boldsymbol{x}_{\mathrm{IN}}$, frozen TOLA diffusion model D

1: Draw fresh noise for the one-step stochastic restoration path
2: Compute the decoded restoration $\boldsymbol{x}_{\mathrm{T}} \leftarrow D(\boldsymbol{x}_{\mathrm{IN}})$ without gradient tracking
3: Compute the spatial-gradient target from $\boldsymbol{x}_{\mathrm{T}}$
4: **return** $\boldsymbol{x}_{\mathrm{T}}$ and its gradient target

Distillation operates entirely in decoded RGB space. The student does not receive diffusion latents, text tokens, confidence scores, intermediate features, or internal denoising states. This interface allows the student architecture to differ from the diffusion architecture and removes the complete diffusion pipeline from deployment. The 876.23M parameters and 297.497 GMACs reported for the Teacher reference in table 3 follow the restoration-module accounting used by TOLA (Xu et al., 2026). These counts exclude the VAE and auxiliary text modules, all of which are also absent from PixelSTR inference.

C Recognition and image evaluation

Recognition training. Training transcriptions are converted from full-width to half-width characters, lowercased, and appended with an end-of-sequence symbol. The frozen TransOCR loss wrapper resizes the student output to 256×32 using the default nearest-neighbor interpolation mode and computes token-level cross-entropy against the GT sequence. Targets use the native recognition labels rather than the IDM vocabulary or transcriptions predicted by the diffusion model. This loss has a coefficient of 0.02 in the student objective. Recognition weights remain fixed, while gradients propagate from the loss through the student output.

Recognition evaluation. TransOCR evaluation uses a fixed 7,934-character alphabet, greedy decoding, and a maximum sequence length of 24. Input images are resized using bicubic interpolation. Text normalization converts full-width characters to half-width characters and text to simplified Chinese, and removes whitespace while preserving letter case. Evaluation therefore preserves case, whereas training labels are lowercased. PaddleOCR processes only the supplied text crops, with detection and angle classification disabled. We use PaddleOCR 2.7.3, PaddlePaddle 2.6.2, and the Chinese recognition model on CPU, with the same text normalization as TransOCR evaluation. PaddleOCR is used exclusively for evaluation.

Let p_i and y_i denote the normalized prediction and reference strings, respectively, N the number of evaluated images, and d_{edit} the Levenshtein edit distance. The recognition metrics are

$$\text{ACC} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[p_i = y_i], \qquad \text{NED} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{d_{\text{edit}}(p_i, y_i)}{\max(|p_i|, |y_i|)}\right). \tag{5}$$

All evaluated GT strings are nonempty. Accuracy differences in percentage points equal differences on the $[0, 1]$ scale multiplied by 100. NED measures string similarity, with higher values indicating better recognition.

Image metrics and result sources. Saved outputs are matched to reference images using the dataset manifest, converted to RGB tensors, and evaluated at the target resolution. PSNR and SSIM are computed on the luminance channel without border cropping. LPIPS uses AlexNet for evaluation and VGG for training. FID compares the output and reference distributions across all 8,089 CTR images. Diffusion-model and baseline scores follow the unified benchmark protocol described in section 4.1. Student and pruning scores are computed from their respective outputs using pyiqa for image metrics and supplemental FID evaluation. FID complements the paired-image and recognition metrics, but small FID differences alone do not establish improved character fidelity. Image resizing and saving conventions also affect FID (Parmar et al., 2022).

D Additional supervision diagnostics

Isolating the two teacher-matching terms. The main ablations remove individual loss terms from Full. We additionally train a teacher-gradient-only control for 50K steps, retaining the GT reconstruction, LPIPS, and OCR terms and setting the teacher-image coefficient to zero. Table 7 compares this control with supervised-only, teacher-image-only (the "without gradient matching" model), and Full. Both controls with a single teacher-matching term perform similarly to supervised-only, while Full achieves substantially higher PSNR and recognition accuracy. These results support the joint use of teacher-derived image and gradient targets.

Table 7: Isolated teacher-matching controls on CTR-TSR-Test. Each model is trained for 50K steps. Teacher-image-only uses the checkpoint reported as "w/o Gradient" in table 4.

Method	PSNR↑	LPIPS↓	FID↓	T-ACC↑	P-ACC↑
Supervised only	22.4290	0.4722	86.2946	0.4242	0.3181
Teacher image only	22.5012	0.4656	82.8581	0.4243	0.3182
Teacher gradient only	22.4352	0.4704	84.5313	0.4254	0.3172
PixelSTR (Full)	**23.9373**	**0.2914**	**23.8595**	**0.4912**	**0.3984**

On RealCE-200, the teacher-gradient-only control achieves 18.6067 dB PSNR, 0.4102 LPIPS, 0.3700 T-ACC, and 0.4300 P-ACC. Its PSNR and both recognition accuracies are again lower than those of Full.

Target-source factorial diagnostic. To assess the effects of the RGB and gradient target sources separately, we evaluate four target-source configurations under the same student architecture and GT reconstruction, LPIPS, and OCR supervision. TT uses teacher RGB and teacher gradients and corresponds to the designated Full model; TG uses teacher RGB and GT gradients; GT uses GT

RGB and teacher gradients; and GG uses GT RGB and GT gradients. GG is the same all-GT control reported as GT-target in table 4. Table 8 shows that TG, GT, and GG perform similarly, while TT achieves substantially better reconstruction and recognition results. Within these configurations, teacher-derived RGB and gradient targets are most effective when used together.

Table 8: Target-source diagnostic on CTR-TSR-Test. All variants are trained for 50K steps. "Teacher" and "GT" indicate the source of the additional RGB and gradient matching targets. All variants retain the base GT reconstruction, LPIPS, and OCR terms.

ID	RGB target	Gradient target	PSNR↑	LPIPS↓	T-ACC↑	P-ACC↑
TT (Full)	Teacher	Teacher	**23.9373**	**0.2914**	**0.4912**	**0.3984**
TG	Teacher	GT	22.5194	0.4719	0.4268	0.3193
GT	GT	Teacher	22.5195	0.4692	0.4298	0.3204
GG	GT	GT	22.5162	0.4708	0.4263	0.3171

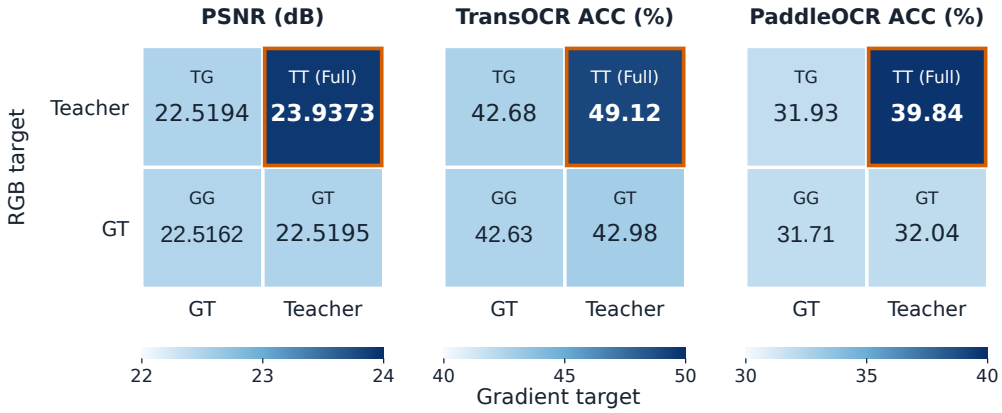

Figure 5: Target-source interactions on CTR-TSR-Test. Using both teacher-derived targets (TT) yields the highest PSNR and OCR accuracy, while the mixed-source and all-GT configurations perform similarly. This pattern supports the complementary roles of teacher-derived RGB and gradient supervision.

With teacher RGB fixed, replacing GT gradients with teacher gradients increases PSNR by 1.4179 dB and T-ACC by 6.44 percentage points (TT versus TG). With GT RGB fixed, the corresponding increases are 0.0033 dB and 0.35 percentage points (GT versus GG). Similarly, teacher RGB yields a large gain when paired with teacher gradients (TT versus GT), but little change when paired with GT gradients (TG versus GG). A separate control trained for 50K steps retains teacher RGB matching and removes both gradient matching and OCR supervision. It achieves 22.5488 dB PSNR, 0.4402 LPIPS, 0.4160 T-ACC, and 0.3144 P-ACC, further supporting the complete objective for recognition.

The same target-source comparison on RealCE-200 is reported in table 9. Full (TT) achieves the highest PSNR and both recognition accuracies among the four configurations, but also has higher LPIPS than the mixed-source and all-GT variants. The cross-dataset results therefore favor the complete teacher-derived target pair for PSNR and recognition, with a trade-off in LPIPS.

Teacher correctness and student behavior. We also group CTR samples according to whether the teacher restoration is recognized correctly by the same TransOCR evaluator used for the student. The teacher-correctness groups are defined using a separate cached teacher evaluation, for which 4,381 of 8,089 outputs are recognized correctly by TransOCR (ACC = 0.5416). Table 1 retains the benchmark Teacher T-ACC of 0.5401; the cached evaluation is used only for stratification and does not replace the benchmark score. As shown in table 10, the improvement of Full over supervised-only is substantially larger on samples with correct teacher recognition, although a smaller positive difference remains on

Table 9: Target-source diagnostic on RealCE-200 using the same 50K CTR-trained checkpoints as table 8.

ID	PSNR↑	LPIPS↓	T-ACC↑	P-ACC↑
TT (Full)	**19.2071**	0.4337	**0.3850**	**0.5000**
TG	18.7007	0.4145	0.3600	0.4450
GT	18.6047	**0.4119**	0.3600	0.4350
GG	18.6253	0.4114	0.3600	0.4400

samples with incorrect teacher recognition. This diagnostic identifies an association between teacher recognition correctness and the magnitude of student improvement.

Table 10: CTR TransOCR accuracy grouped by whether the re-evaluated teacher output is recognized correctly. Differences are Full minus supervised-only; intervals are paired image-level 95% bootstrap confidence intervals.

Teacher group	Images	Full ACC	Supervised ACC	Difference (pp), 95% CI
Correct	4,381	0.8667	0.7558	11.09 [9.91, 12.30]
Incorrect	3,708	0.0475	0.0324	1.51 [0.76, 2.27]

Among samples with nonempty, incorrect teacher predictions, Full produces the same erroneous TransOCR string as the teacher more often than supervised-only does (15.10% versus 9.60%). Thus, some student errors also coincide with teacher errors.

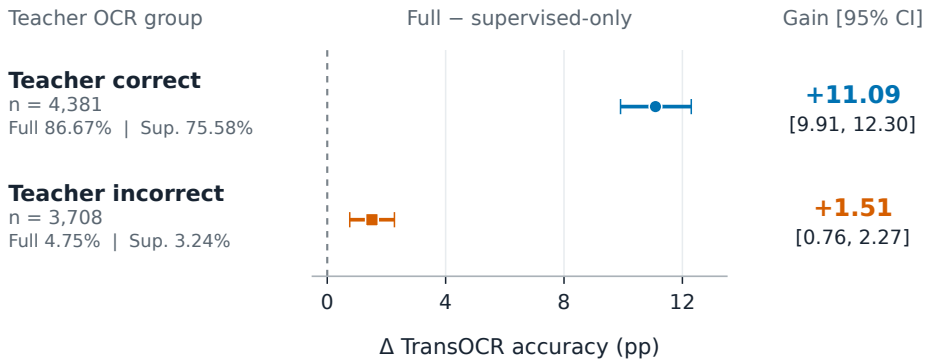

Figure 6: CTR-TSR-Test results grouped by teacher recognition correctness. Full improves TransOCR accuracy over supervised-only in both groups, with a substantially larger gain when the teacher restoration is recognized correctly. Error bars show paired image-level 95% bootstrap confidence intervals.

E Complete cross-dataset ablations

Table 11 compares the core controls, all three GAN weights, the larger student, and bicubic interpolation on the same 200 RealCE pairs. Every learned model is trained on CTR for 50K steps and evaluated without RealCE tuning, and Full uses the default checkpoint reported in table 2. The larger student does not improve recognition accuracy over Full on RealCE. GAN $\lambda = 0.05$ reduces LPIPS and DISTS, but also reduces both OCR accuracies. Bicubic interpolation already provides competitive recognition performance on this dataset: its T-ACC matches Full, and its P-ACC is only 0.5 percentage points lower. Full nevertheless achieves lower perceptual distances.

Relative to supervised-only training, Full improves RealCE PSNR by 0.5681 dB, T-ACC by 3.0 percentage points, and P-ACC by 4.5 percentage points. The corresponding improvements over

Table 11: RealCE-200 controls and additional comparisons. Learned models are trained for 50K steps on CTR and evaluated without RealCE tuning. Full uses the default checkpoint from table 2. Bicubic provides a reference without learned restoration. Bold values indicate the best result in each column among the models shown.

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	T-ACC↑	T-NED↑	P-ACC↑	P-NED↑
PixelSTR (Full)	**19.2071**	**0.6102**	0.4337	0.3083	**0.3850**	**0.5839**	**0.5000**	**0.6717**
Sup-only	18.6390	0.5898	0.4088	0.2935	0.3550	0.5553	0.4550	0.6242
GT-target	18.6253	0.5870	0.4114	0.2949	0.3600	0.5626	0.4400	0.6254
w/o Grad	18.6170	0.5872	0.4099	0.2910	0.3650	0.5622	0.4250	0.6173
w/o OCR	18.6666	0.5903	**0.3991**	**0.2907**	0.3500	0.5534	0.4500	0.6204
GAN .01	18.6770	0.5938	0.4188	0.2965	0.3600	0.5641	0.4700	0.6376
GAN .025	18.7905	0.5988	0.4200	0.3003	0.3700	0.5681	0.4500	0.6404
GAN .05	18.8138	0.5987	0.4177	0.3006	0.3750	0.5752	0.4850	0.6540
Larger	18.8080	0.5872	0.4459	0.3185	**0.3850**	0.5741	0.4850	0.6534
Bicubic	19.1912	0.5982	0.4927	0.3367	**0.3850**	0.5813	0.4950	0.6689

Table 12: Consolidated CTR-TSR-Test ablations on 8,089 images, with 50K training steps per model. Results reproduce the main-text ablation measurements and do not represent additional runs. Full uses the default checkpoint from the main table.

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	T-ACC↑	T-NED↑	P-ACC↑	P-NED↑
PixelSTR (Full)	23.9373	0.6863	0.2914	0.2238	23.8595	0.4912	0.6913	0.3984	0.6000
Supervised only	22.4290	0.6033	0.4722	0.3161	86.2946	0.4242	0.6105	0.3181	0.5104
GT-target	22.5162	0.6074	0.4708	0.3146	85.1284	0.4263	0.6141	0.3171	0.5091
w/o Gradient	22.5012	0.6059	0.4656	0.3078	82.8581	0.4243	0.6129	0.3182	0.5129
w/o OCR	22.5824	0.6138	0.4426	0.2984	73.2668	0.4204	0.6091	0.3145	0.5106
GAN ($\lambda = 0.01$)	22.5819	0.6125	0.4806	0.3178	89.9906	0.4310	0.6191	0.3209	0.5159
GAN ($\lambda = 0.025$)	22.5834	0.6138	0.4894	0.3261	97.1003	0.4358	0.6210	0.3217	0.5145
GAN ($\lambda = 0.05$)	23.7138	0.6787	0.2957	0.2285	23.4797	0.4810	0.6796	0.3837	0.5944
Larger student	23.9461	0.6846	0.2831	0.2186	23.0076	0.5017	0.6978	0.4007	0.6072

GT-target are 0.5818 dB, 2.5 percentage points, and 6.0 percentage points. Full has higher LPIPS and DISTS than both controls, indicating a trade-off between recognition accuracy and perceptual similarity. The aggregate accuracy differences are reported in table 14. Paired intervals for these scores would require the corresponding per-image predictions. With 200 crops, one additional exact-match prediction changes accuracy by 0.5 percentage points. Dependence among crops from the same source image further limits the interpretation of small differences.

Table 12 consolidates the CTR component, capacity, and GAN comparisons discussed in the main text. These results reuse the same measurements; no additional runs are introduced. Every comparison uses the same default Full checkpoint and the complete test set.

F Additional qualitative comparison

Figure 7 compares the teacher, the FFN-pruned reference, and the student on the same CTR inputs. All displayed restorations are unmodified experimental outputs.

G Training repeatability and paired uncertainty

Table 5 summarizes variability across independent training runs of Full and supervised-only. The analyses below report aggregate accuracy differences and, where the corresponding per-image outputs



Figure 7: Selected CTR-TSR-Test comparisons. Columns show LR, Teacher (876.23M), FFN-pruned (516.73M), PixelSTR (8.62M), and HR. Parameter counts follow table 3.

are available, uncertainty due to test-sample variation for fixed trained checkpoints. These bootstrap intervals and the standard deviations across training runs quantify distinct sources of variation.

Table 13: CTR accuracy differences between Full, the default checkpoint, and each variant. TransOCR intervals are unadjusted 95% percentile intervals from 100,000 image-level bootstrap replicates and exclude variation across training runs. PaddleOCR entries are differences between reported aggregate scores; confidence intervals are unavailable. Positive differences favor Full.

Full minus variant	ΔT-ACC (pp), 95% CI	ΔP-ACC (pp)
Sup-only	6.70 [5.97, 7.44]	8.03
GT-target	6.49 [5.76, 7.22]	8.13
w/o Grad	6.69 [5.96, 7.42]	8.02
w/o OCR	7.07 [6.34, 7.80]	8.39
GAN .01	6.02 [5.29, 6.75]	7.75
GAN .025	5.54 [4.82, 6.27]	7.67
GAN .05	1.01 [0.47, 1.57]	1.47
Larger	−1.05 [−1.61, −0.49]	−0.23

Paired test-sample uncertainty. For the CTR TransOCR comparisons, we pair correctness indicators from Full and each comparison variant by sample ID. Each image-level bootstrap replicate draws N paired indices with replacement and recomputes the mean accuracy difference. We use 100,000 replicates and take the 2.5th and 97.5th percentiles as the endpoints of the 95% interval. These intervals are conditional on the trained checkpoints and are not adjusted for multiple comparisons. They exclude variation across training runs and uncertainty associated with hyperparameter selection. Source-image groups are unavailable for CTR, so independence among crops is an approximation. Every difference in tables 13 and 14 is defined as Full minus the named variant; negative values favor the variant. Reported aggregate scores for CTR PaddleOCR and RealCE support point differences but do not determine paired confidence intervals, discordant-pair counts, or McNemar test results. We therefore omit these quantities for these comparisons.

For GAN $\lambda = 0.05$, TransOCR recognizes 300 samples correctly under Full but incorrectly under GAN, and 218 samples incorrectly under Full but correctly under GAN. GAN therefore yields 82 fewer correct predictions overall. The 19M student yields 85 more correct TransOCR predictions than Full, with the paired interval favoring the larger model. The reported aggregate PaddleOCR scores place Full 1.47 percentage points above GAN and 0.23 percentage points below the 19M student. These aggregate scores do not determine how many individual predictions differ in each direction.

The 200 RealCE crops come from 73 source images identified in the manifest. The point estimates in table 14 weight each crop equally, consistent with the benchmark scores. Paired resampling should account for dependence among crops from the same source image. Because the reported aggregate scores do not determine paired outcomes within these groups, we report neither image-level nor source-cluster confidence intervals for these RealCE comparisons.

Table 14: RealCE-200 aggregate accuracy differences between Full and each variant. Positive values favor Full. Differences are computed from the reported main-table results and the comparison results in table 11. Paired confidence intervals are unavailable for these comparisons.

Full minus variant	ΔT-ACC (pp)	ΔP-ACC (pp)
Sup-only	3.00	4.50
GT-target	2.50	6.00
w/o Grad	2.00	7.50
w/o OCR	3.50	5.00
GAN .01	2.50	3.00
GAN .025	1.50	5.00

H EFFICIENCY BOUNDARIES AND MEMORY

Table 15: Additional student efficiency measurements using the same forward-pass measurement scope as table 3. Peak values report the maximum allocated GPU memory during the measured forward pass.

Student	MACs (G)	Forward (ms)	Peak (MiB)
PixelSTR	4.9517	8.36	88.19
Larger	9.4953	11.08	144.42

Table 3 uses the same forward-pass measurement protocol for all methods: a batch size of one, FP32 execution on the same idle RTX PRO 6000, the same fixed list of 100 CTR images, inputs in $[0, 1]$, warm-up, three measurement rounds, and synchronized CUDA timing. Inputs are resized and placed on the GPU before measurement. Timing excludes file reading, external bicubic resizing, host-to-device transfer, output saving, and model loading. Student parameter and MAC counts cover the complete NAFNet. Teacher and FFN-pruned counts cover the restoration modules and exclude the VAE and auxiliary text modules, while latency measures the executed single-step forward path.