CNN-ViT Showdown for Racial Equality in Upscaling and Racial Recognition

Shou-Jen Chen [*1] **Justin Tsai** [*1] **Derek Xu** [*1]

Abstract

Facial image data plays a central role in many modern computational systems, forming the basis for tasks such as upscaling and recognition. However, despite significant progress in machine learning, many current models still have racial performance disparities, processing faces of different groups with unequal accuracy. In this study, we compare the performance of models with ResNet and FastViT backbones on both image upscaling and racial classification tasks across diverse racial groups to better understand how architectural choices influence fairness. We find that while superresolution performance remains largely race-agnostic even under extremely imbalance data training, classification fairness varies by architecture, with ResNet50 maintaining more stable accuracy across minority groups than FastViT. The code is available at https://github.com/Maltomatic/RaceComp.

1. Problem Statement

Modern facial analysis systems rely heavily on deep neural architectures trained on large image datasets, yet, racial imbalance continues to cause performance disparities across demographic groups. Athough architectural advancements, such as ResNet (He et al., 2015) and vision transformers (Dosovitskiy et al., 2021), it remains unclear which model classes are intrinsically more resilient to biased training distributions when applied to upscaling and race recognition tasks. In this work, we focus on racial classification rather than identity recognition, allowing us to isolate fairness effects tied to demographic labels.

The goal of this work is to determine whether convolutional or transformer-based backbones are more stable under racial imbalance. We focus on U-Net models equipped with ResNet50 or FastViT encoders and evaluate them on racially diverse datasets.

We will investigate the following research questions:

- **Architecture Fairness Robustness:** How differently do CNN-based and ViT-based U-Net architectures degrade across racial groups when trained with intentionally imbalanced data?
- **Task Sensitivity:** Are the fairness gaps larger for pixel-level tasks(upscaling) or semantic tasks (classification)?
- **Effect of Imbalance:** How do performance change when minority groups are undersampled at a different rate?
- **Backbone Stability**: Between ResNet50 and FastViT, which backbone has a more consistent per-race performance?

2. Related Work

2.1. Reducing Racial Bias in Face Recognition

There have also been other works that aimed to reduce racial bias in facial recognition in various ways. For instance, one study applied reinforcement learning to reduce bias in facial recognition by reducing the gap between the model's performance on different races (Wang & Deng, 2019a). Furthermore, another study focused on creating a less racially biased dataset that is more balanced between races, which can be used to train fairer models (Jain et al., 2023). However, racial bias has only been studied in very few types of models, which is why we plan to investigate the extent of racial bias in facial data for the ResNet50 and FastViT models. Recent work has also explored fairness properties of transformer-based vision models, though conclusions remain mixed depending on task and dataset.

2.2. Upsampling

While we are using CNNs and ViTs to upscale images, many simpler methods of image upscaling exist as well. For instance, nearest-neighbor upsampling simply involves

[1] Department of EECS, University of California, Berkeley, U.S.. Correspondence to: Shou-Jen Chen <lawrancechen@berkeley.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

copying pixels to fill a larger version of an image. For instance, each pixel becomes a 2x2 arrangement of itself if both the length and width are doubled this way. However, this method is not completely reliable since it fails to correspondingly increase the image resolution. Another method that can increase both image dimensions and resolution is bilinear interpolation, which makes each new pixel's values a weighted average of the values in surrounding pixels. However, the averaging mechanism is prone to blurring, which makes it poorly replicate fine details in the image (Ups).

In addition to non-learnable ways of upscaling images, more basic methods of upscaling using machine learning can be implemented with CNNs. CNNs are used to change the size of feature maps for tasks such as image resizing. For instance, one of the methods described previously can be used to initially increase the image's dimension, and then the resultant image can be passed through a CNN (Ups). Also, the initial upscaling stage can be skipped and the original image can be passed into the CNN as well (Ups; CSD, 2025).

A more sophisticated solution for classification involves using the ResNet architecture, which is a deep convolutional neural network that reduces the impact of issues like vanishing gradients. This can be incorporated into the U-Net architecture, which can learn features from images and use them for upscaling. Specifically, the U-Net architecture consists of two parts: an encoder for learning features from the images, and a decoder which produces the upscaled images, which could use ResNet as a backbone. The U-Net with ResNet architecture is widely used for image analysis in many areas (Tomar, 2024; cod, 2025; CSD, 2022).

Furthermore, other methods have also been used to improve the quality of image upscaling. For instance, the Adaptive Deviation Modulator (AdaDM) has been developed in place of normalization for neural network-based solutions for upscaling, such as ResNet, to improve generalization (Liu et al., 2021). Furthermore, there has also been research about improving the performance of image upscaling with CNNs and subpixel convolution (Shi et al., 2016).

Another related work that involved creating textures through upscaling examined the Super-Resolution Generative Adversarial Network (SRGAN). This work proposed Enhanced SRGAN (Wang et al., 2018), which makes substantial use of Residual-in-Residual Dense Blocks (RRDB) in order to mitigate the effects of hallucination in SRGAN (Wang et al., 2018).

2.3. Classification

One of the many approaches of image classification using CNNs is ConvMixer. Research has shown that the efficiency of this approach can be improved through pixel shuffling downsampling, which improves classification performance compared to previous methods of image patching (Ibrahem et al., 2025).

3. Background

Many visual-related task uses images of people from diverse racial backgrounds. Tasks such as classification, enhancement, and quality evaluation often rely on models trained with uneven demographic representation, which can lead to unequal performance across racial groups (Puyol-Antón et al., 2022; Cross et al., 2024). For example, an upscaling model may consistently produce higher-quality results for one race while degrading fidelity for others. This highlights the need to consider equity when designing algorithms that process human-centered visual data.

One particular system where racial biases can have catastrophic impacts are self-driving systems. These systems must reliably detect pedestrians to make correct stopping decisions. If they have different abilities to recognize different races, some races may be more susceptible to crashes in self-driving vehicles. As a result, mitigating racial disparities in perception systems is essential to ensuring safety for all users.

Racial recognition systems in search engines also show how harmful racial biase can be. When these systems have different abilities of recognizing members of different races, they could overrepresent certain groups while underrepresenting others. This imbalance creates inequitable visibility within the results of search engines, which can cause some individuals to receive less attention because of their race.

3.1. ResNet50

ResNet50 (He et al., 2015) is a 50-layer convolutional neural network that introduces residual connections to stabilize the training of deep CNNs and extract hierarchical local features. Its strong performance, widespread adoption, and moderate computational cost make it a standard backbone for vision tasks. We use ResNet50 as the representative CNN encoder to evaluate how convolution-based architectures behave under racially imbalanced training data.

3.2. FastViT

FastVit (Vasu et al., 2023) a hybrid Vision Transformer architecture developed by Apple Inc. designed specifically for low memory usage and fast computational speed. It combines convolutional token mixing with global self-attention to achieve higher accuracy with lower memory and latency. It retains the main structure of classic ViT models, and was therefore chosen as a token representation of ViT models in this comparison.

3.3. Metrics

To evaluate the quality of image upscaling, prior work in super-resolution and image reconstruction has established several standard quantitative metrics (Wang et al., 2018; Ledig et al., 2017). PSNR and SSIM capture different aspects of image fidelity and are most commonly used to benchmark both classical and deep learning–based models.

Peak Signal-to-Noise Ratio (PSNR). PSNR is one of the most widely used metrics for evaluating reconstruction quality. It measures the ratio between the maximum possible pixel intensity and the mean squared error (MSE) between the reconstructed image and the ground truth. Higher PSNR values indicate lower distortion and therefore better fidelity. In super-resolution work (Wang et al., 2018; Ali et al., 2023), differences of 0.1–0.5 dB are generally considered meaningful, while differences below 0.1 dB are typically imperceptible and often fall within normal training variance.

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right), \tag{1}$$

In our context, if PSNR varies across racial groups, this may indicate that the model is reconstructing certain groups with consistently higher error, which would be a potential fairness concern.

Structural Similarity Index (SSIM). SSIM complements measures perceptual similarity. It compares local statistics of luminance, contrast, and structural patterns between two images. SSIM values range from –1 to 1, where 1 indicates identical structural content. This is particularly relevant in fairness analysis: even if two racial groups have similar PSNR, different SSIM values may indicate subtle structural degradation disproportionately affecting certain groups (Nilsson & Akenine-Möller, 2020).

$$\text{SSIM}(x, \hat{x}) = \frac{(2\mu_x\mu_{\hat{x}} + C_1)(2\sigma_{x\hat{x}} + C_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + C_2)}, \tag{2}$$

Other Perceptual Metrics. Several more advanced perceptual metrics exist, such as Learned Perceptual Image Patch Similarity(LPIPS) (Zhang et al., 2018), Fréchet Inception Distance(FID) (Heusel et al., 2018), and Natural Image Quality Evaluator(NIQE) (Mittal et al., 2013). These metrics often better capture high-level texture realism or perceptual artifacts. However, they also introduce model bias from the feature extractors they rely on, which can complicate fairness analysis. LPIPS and FID depend on deep feature spaces learned from datasets with uneven demographic representation, so they may inadvertently encode demographic biases and thereby distort measurement. For these reasons, and to maintain interpretability, we restrict our evaluation to metrics with well-understood statistical behavior.

4. Methods

In this project, we aim to apply transfer learning to ResNet-50 and FastViT (both ImageNet-1k-pretrained). We use the racially-balanced FairFace and Racial Faces in-the-Wild (RFW) datasets as a starting point, then intentionally reduce the representation of races to compare how they generalize in the face of imbalanced racial data. We judge the models' "fairness" in upscaling and racial recognition across races when training data is skewed (1). Since the model architectures themselves may influence how bias occurs, we define fairness in terms of how consistently each architecture performs across racial groups, rather than in absolute image quality. Throughout this work, we define fairness as performance stability across racial groups under distribution shift, rather than absolute parity or equality of outcomes. Specifically, we will measure the change in per-group performance when trained on balanced versus imbalanced datasets to capture how resilient each model is to biased data. In short, our goal is to determine what model architecture might be better suited for maintaining fairness in real-world racial recognition. We intentionally refrain from using LLM-based methods like VLMs because beyond computational restrictions, their massive pretraining and complex encodings could obscure the resulting differences. ResNet50 and FastViT were selected as token representations for CNNs and ViTs to partake in the comparison for their simplicity, similarity to the core network structure, lighter memory footprint, and lower computational resources needed.

4.1. Datasets

We evaluate and train our models using two racially diverse datasets:

FairFace (for Upscaling) A 108k-image dataset with labels across seven demographic groups. After filtering and splitting:

Race	Count
White	16527
Latino/Hispanic	13367
Indian	12319
East Asian	12287
Black	12233
Southeast Asian	10795
Middle Eastern	9216

Table 1. **Class distribution of training data in the dataset by race.**

Images are center-cropped, resized to 56×56, and converted to RGB float tensors.

Minority Setting	**Black**		**White**		**East Asian**		**SE Asian**		**Indian**		**Middle Eastern**		**Latino/Hispanic**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
All	22.77	0.932	22.75	0.932	22.75	0.932	22.75	0.932	22.77	0.932	22.75	0.932	22.75	0.932
East Asian	23.14	0.937	23.12	0.937	23.12	0.938	23.13	0.937	23.16	0.938	23.13	0.937	23.13	0.938
Indian	24.92	0.953	24.87	0.953	24.90	0.953	24.88	0.953	24.90	0.953	24.89	0.953	24.88	0.953
Black	25.87	0.958	25.87	0.958	25.86	0.958	25.86	0.958	25.88	0.958	25.87	0.958	25.86	0.958
White	23.77	0.937	23.75	0.937	23.76	0.937	23.75	0.937	23.76	0.937	23.76	0.937	23.76	0.937
Middle Eastern	23.18	0.935	23.16	0.935	23.17	0.935	23.17	0.935	23.20	0.935	23.17	0.935	23.18	0.935
Latino Hispanic	24.47	0.946	24.44	0.946	24.45	0.946	24.44	0.946	24.47	0.946	24.47	0.946	24.45	0.946
Southeast Asian	23.06	0.936	23.03	0.936	23.04	0.936	23.04	0.936	23.04	0.936	23.04	0.936	23.03	0.936

Table 2. **Per-race PSNR (dB) and SSIM for FastViT upscaling across different minority-downsampling settings.** Rows correspond to the race that was downsampled during training, while columns show the resulting model performance on each race's validation set.

RFW (Racial Faces in the Wild) (for Classification) We use four verification subsets: (1) African (2) Asian (3) Caucasian (4) Indian

For racial classification, we ignore their individual identities and separate the dataset into 4 races, matching prior preprocessing from FairFace-based pipelines. We use RFW because our initial plan was to evaluate facial recognition, and RFW provides identity labels needed for that task. However, both ResNet and ViT did not exhibit sufficient performance variation to meaningfully analyze fairness effects. So, we shifted our focus to racial recognition while still using RFW's race-labeled subsets.

Imbalance Simulation: For each imbalance training, minority races are sampled with 5% representation of majority races. This results in effective training distribution skewed 1:20. To avoid accidentally sampling new, unseen images in different epochs and breaking the underrepresentation simulation, at the start of the training run for each minority, we randomly eliminate images of minority races in the whole dataset to just 5% of its initial data size compared to 100% representation in normal races.

Train/Validation Splits: We use a 80% training and 20% validation split.

4.2. Upscaling

We developed upscaling algorithms using U-Net with the ResNet50 and FastViT backbones. For upscaling, we trained our models on the FairFace dataset (Karkkainen & Joo, 2021). For each model, we will utilize transfer learning by unfreezing layers from the back in stages.

4.3. Classification

Similar to upscaling, we also use the ResNet50 and FastVIT structures for racial classification, with the upsampling decoder head removed. Instead of FairFace, we used the RFW dataset for examining upscaling (Wang et al., 2019; 2021; Wang & Deng, 2019b; 2021).

4.4. Training

For training both the upscaling and classification models, we used most of the dataset for model training. Each image in the training set will be represented multiple times in the actual training set with different transformations applied to it, as we aim to build robust models that can adapt to faces presented in various ways. During training, we minimized cross-entropy loss for classification and used VGG19-based perception loss for image upscaling. We fine-tuned only the later stages of each backbone. For ResNet50 in the upscaling task, we first unfroze layer4 and trained for 2 epochs, then unfroze layer3 for another 2 epochs. For FastViT, we unfroze layers 5–7 for 2 epochs, followed by layers 3–4 for an additional 2 epochs. For the classification task, we added a third stage: ResNet50's layer2 and FastViT's layers 3-4 were also unfrozen and trained for one more epoch.

We selected these stages based on empirical stability: deeper layers adapted well without causing overfitting or destroying pretrained representations, while unfreezing too many layers led to unstable training. For classification, we added a third stage by unfreezing ResNet50's layer2 and FastViT's layers 3–4 for one more epoch to allow limited mid-level features to adapt without fully retraining the networks.

5. Results

We used PSNR and SSIM to evaluate our upscaling task and accuracy on our classification task, measuring the proportion of correctly identified race labels to assess overall and per-group performance.

5.1. Analysis

(1) Upscaling

We first examine the upscaling results for both architectures. Across all minority-downsampling settings, both FastViT and ResNet50 show stable performance in PSNR and SSIM, with only minimal variation across races (Table 2, Table 3).

Minority Setting	Black		White		East Asian		SE Asian		Indian		Middle Eastern		Latino/Hispanic	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
All	29.92	0.983	29.90	0.983	29.93	0.983	29.91	0.983	29.92	0.983	29.92	0.983	29.90	0.983
East Asian	29.21	0.979	29.19	0.979	29.19	0.979	29.20	0.979	29.22	0.979	29.20	0.979	29.19	0.979
Indian	29.15	0.979	29.13	0.979	29.14	0.979	29.14	0.979	29.15	0.979	29.14	0.979	29.14	0.979
Black	29.86	0.982	29.81	0.982	29.83	0.982	29.79	0.982	29.87	0.982	29.85	0.982	29.77	0.982
White	29.91	0.983	29.84	0.983	29.87	0.983	29.83	0.983	29.92	0.983	29.88	0.983	29.80	0.983
Middle Eastern	29.93	0.983	29.87	0.983	29.91	0.983	29.86	0.983	29.95	0.983	29.91	0.983	29.84	0.983
Latino Hispanic	29.77	0.982	29.72	0.982	29.75	0.983	29.69	0.982	29.79	0.983	29.75	0.983	29.68	0.983
Southeast Asian	29.86	0.983	29.82	0.983	29.85	0.983	29.80	0.983	29.89	0.983	29.84	0.983	29.78	0.983

Table 3. **Per-race PSNR (dB) and SSIM for ResNet50 upscaling across different minority-downsampling settings.** Rows correspond to the race that was downsampled during training, while columns show the resulting model performance on each race's validation set.

For FastViT upscaling (Table 2), PSNR differences for each minority setting is 0.02–0.05 dB, and SSIM are tightly grouped between 0.93 and 0.96 for all races and training conditions. Similarly, ResNet50 upscaling (Table 3) is even more stable, with PSNR consistently around 29 dB and SSIM of around 0.98 for every race. These results indicate that, for the upscaling task, neither model demonstrates meaningful racial sensitivity, and minority downsampling has negligible impact on reconstruction quality.

In absolute terms, ResNet50 outperforms FastViT for upscaling, achieving 5.79 dB higher PSNR and consistently higher SSIM. This suggests that CNN-based encoders preserve low-level image structure more effectively than transformer-based ones.

(2) Classification

The classification results present a different pattern. Classification accuracy shows clear racial sensitivity when a particular race is downsampled during training. For FastViT classification (Table 4), one group stood out prominently: African faces maintain high accuracy across all settings, even when African is set to minority. When Asian, Caucasian, and Indian are each set to minority, there's a substantial accuracy drops. ResNet50 follows these same patterns: Caucasian, Asian, and Indian accuracies decline to 0.75 when its respected images are downsampled (Table 5). Notably, African faces remain consistently high-performing even when downsampled.

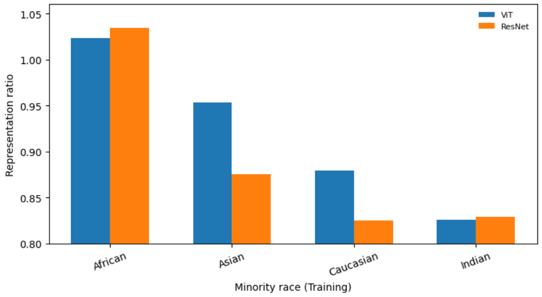

Figure 1. Representation Ratio on Racial Classification Task

(3) ResNet50 v.s. FastViT

Comparing architectures, both models exhibit nearly identical fairness trends in classification. The races that is most affected by downsampling are Indian and Caucasian, with a drop of performance of 18.08% and 16.03% respectively across two architectures. Asian faces has moderate sensitivity with a drop 11.545%. African faces are the most stable with a performance drop of 3.08%.

Minority Setting	African	Asian	Caucasian	Indian
African	0.953	0.940	0.919	**0.914**
Asian	0.976	**0.878**	0.910	0.919
Caucasian	0.984	0.949	**0.803**	0.916
Indian	0.981	0.943	0.940	**0.745**

Table 4. **Per-race classification accuracy for FastViT on racial recognition task under different minority-downsampling settings.** Rows correspond to the race used as the minority (downsampled) during training, and columns indicate the resulting classification accuracy on each race's validation set.

On average, FastViT has a drop in performance of 10.37% and ResNet50 exhibits a drop of 14.13%. This suggests that FastViT is less susceptible to downsampling and has lower racial sensitivity.

To better quantify fairness differences across the models, we compute a representation ratio for each race under each minority-downsampling condition. This metric compares the accuracy of the minority race to the model's average accuracy across all races in that same training setting.

$$\text{RepresentationRatio}_r = \frac{\text{Acc}(r)}{\text{AvgAcc}}$$

Image 1 shows that African faces remain the most robust group for both ViT (1.023) and ResNet (1.034), even slightly exceeding the models' average accuracy, indicating that reducing African training samples does not meaningfully harm their performance. Asian faces show moderate underrepresentation, dropping to 0.954 in ViT and a more pronounced

0.875 in ResNet. In contrast, Caucasian and Indian faces exhibit the strongest declines across both models: ViT falls to 0.880 for Caucasians and 0.826 for Indians, while ResNet shows similarly low ratios of 0.825 and 0.829. These results show a consistent fairness issue where African faces are resilient to data reduction, Asian faces are moderately affected, and Caucasian and Indian faces are the most sensitive to underrepresentation across both architectures.

Minority Setting	**African**	**Asian**	**Caucasian**	**Indian**
African	0.943	0.920	**0.887**	0.897
Asian	0.964	**0.778**	0.937	0.876
Caucasian	0.978	0.942	**0.740**	0.928
Indian	0.986	0.920	0.937	**0.743**

Table 5. **Per-race classification accuracy for ResNet50 on racial recognition task under different minority-downsampling settings.** Rows correspond to the race used as the minority (downsampled) during training, and columns indicate the resulting classification accuracy on each race's validation set.

Overall, these results demonstrate that upscaling is possibly a race-agnostic task for both architectures, while classification displays race-dependent performance variation driven mostly by training imbalance and marginally on architectural structure. Caucasian and Indian groups show the strongest dependence on adequate representation, whereas African faces generalize well even under heavy downsampling. The shared behavior across FastViT and ResNet50 indicates that fairness challenges mostly originate from the underlying data distribution rather than the choice of model architecture. While ResNet50 exhibits more uniform per-race behavior, FastViT experiences smaller average relative drops under downsampling, leading to different interpretations of ”fairness” depending on the metric.

6. Conclusion

6.1. Superresolution

The results paint a picture that while may be upsetting for computer scientists who wish for observably different behavior based on the training data for image upscaling (superresolution) and generation, is a statement made from a broader societal perspective: we are not as unalike as it seems. For the upscaling task, the tested models are overwhelmingly racially agnostic, performing well across all races despite massive minority/majority representation differences. This is a significant and optimistic result in an age where image generation and artistic creations from ML models are increasingly common.

6.2. Classification

We note that in general, FastViT performs better than ResNet50 in regards to fairness. Aside from higher accuracy overall, its accuracy of minority racial representations is not compromised as much relatively. We hypothesize that the relative weakness of CNNs is due to their inherent structure: their filters are capable of generalizing to broad features effectively, but may fail to capture finer details and their relative relation to the rest of the image. Vision Transformers and their strength for context may help them learn to recognize subtle differences in image context slightly better. Considering the oft-cited downside of Vision Transformers, which is that they typically require more training data than CNNs, this may imply a ViT may work as a robust and fair model for facial or racial recognition tasks provided that even the underrepresented minority or minorities have sufficient samples. When a race does not have outstanding features that are easily learned and represented or are similar to another race with more prominent features (eg. in RFW some Indian image samples are similar to African image samples, but the African class has more prominent features overall), there is a noticeable decrease in racial recognition accuracy. More training data, understandably, helps a model understand discriminative cues and therefore racially profile an individual. A lack of that data, meanwhile, significantly reduces the accuracy, implying that the model can only fall back to broader, more features that may be shared with other races. Again, this hints at more external similarities between races than we have been led to think, but at the same time, this may mean that the performance of models that rely on visible facial features—device facial unlocking, for a common task—may be compromised for certain users if a carelessly imbalanced dataset is employed.

6.3. Future Work

In the future, we can add more variables to capture a more in-depth picture of fairness, such as increasing the downsampling races to gain a pairwise comparison of multiple downsampled data. We can evaluate how other models perform on tasks involving image data with different races, such as more sophisticated and cutting-edge models. Furthermore, we can also evaluate our U-Net based models on other tasks involving images with people of more finely-differentiated races as well, which could potentially be more complex. We might also investigate how the architectures we examined can be improved to be less racially biased and perform similarly on various visual tasks involving different races.

References

Upsampling layers in cnn decoders. URL https://apxml.com/courses/applied-autoencoders-feature-extraction/chapter-5-convolutional-autoencoders-image-data/upsampling-techniques-decoders.

Jul 2022. URL https://blog.csdn.net/weixin_54255111/article/details/125625991.

Mar 2025. URL https://blog.csdn.net/shizheng_Li/article/details/146232927.

Nov 2025. URL https://www.codegenes.net/blog/unet-with-resnet-backbone-pytorch/.

Ali, S., Jamil, U., Jabbar, M., Sajid, A., and Jabbar, M. Evaluation of psnr value for image super-resolution using deep learning. *Lahore Garrison University Research Journal of Computer Science and Information Technology*, 12 2023. doi: 10.54692/lgurjcsit.2023.074457.

Cross, J. L., Choma, M. A., and Onofrey, J. A. Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, 3(11):e0000651, 2024. doi: 10.1371/journal.pdig.0000651. URL https://doi.org/10.1371/journal.pdig.0000651. Published November 7, 2024.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. URL https://arxiv.org/abs/1706.08500.

Ibrahem, H., Salem, A., and Kang, H.-S. Pixel shuffling is all you need: spatially aware convmixer for dense prediction tasks. *Pattern Recognition*, 158:111068, 2025. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2024.111068. URL https://www.sciencedirect.com/science/article/pii/S0031320324008197.

Jain, A., Dholakia, R., Memon, N., and Togelius, J. Zero-shot demographically unbiased image generation from an existing biased stylegan. December 2023. doi: 10.36227/techrxiv.24634239.v1. URL http://dx.doi.org/10.36227/techrxiv.24634239.v1.

Karkkainen, K. and Joo, J. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1548–1558, 2021.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. Photo-realistic single image super-resolution using a generative adversarial network, 2017. URL https://arxiv.org/abs/1609.04802.

Liu, J., Tang, J., and Wu, G. Adadm: Enabling normalization for image super-resolution, 2021. URL https://arxiv.org/abs/2111.13905.

Mittal, A., Soundararajan, R., and Bovik, A. C. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

Nilsson, J. and Akenine-Möller, T. Understanding ssim, 2020. URL https://arxiv.org/abs/2006.13846.

Puyol-Antón, E., Ruijsink, B., Mariscal Harana, J., Piechnik, S. K., Neubauer, S., Petersen, S. E., Razavi, R., Chowienczyk, P., and King, A. P. Fairness in cardiac magnetic resonance imaging: Assessing sex and racial bias in deep learning-based segmentation. *Frontiers in Cardiovascular Medicine*, Volume 9 - 2022, 2022. ISSN 2297-055X. doi: 10.3389/fcvm.2022.859310. URL https://www.frontiersin.org/journals/cardiovascular-medicine/articles/10.3389/fcvm.2022.859310.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, 2016. URL https://arxiv.org/abs/1609.05158.

Tomar, N. Step-by-step guide to resnet50 unet in tensorflow, Jan 2024. URL https://idiotdeveloper.com/step-by-step-guide-to-resnet50-unet-in-tensorflow

Vasu, P. K. A., Gabriel, J., Zhu, J., Tuzel, O., and Ranjan, A. Fastvit: A fast hybrid vision transformer using structural reparameterization, 2023. URL https://arxiv.org/abs/2303.14189.

Wang, M. and Deng, W. Mitigate bias in face recognition using skewness-aware reinforcement learning. *CoRR*, abs/1911.10692, 2019a. URL http://arxiv.org/abs/1911.10692.

Wang, M. and Deng, W. Mitigate bias in face recognition using skewness-aware reinforcement learning. *arXiv preprint arXiv:1911.10692*, 2019b.

Wang, M. and Deng, W. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.

Wang, M., Deng, W., Hu, J., Tao, X., and Huang, Y. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

Wang, M., Zhang, Y., and Deng, W. Meta balanced network for fair face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., and Tang, X. Esrgan: Enhanced super-resolution generative adversarial networks, 2018. URL https://arxiv.org/abs/1809.00219.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL https://arxiv.org/abs/1801.03924.