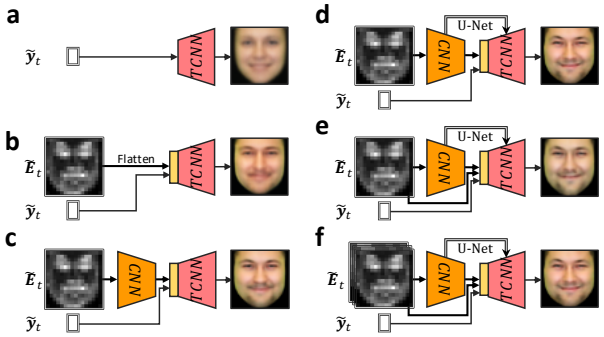
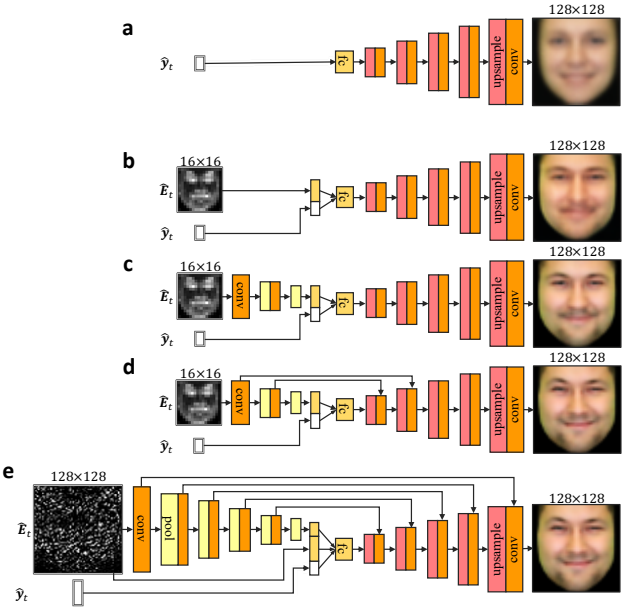
A. Supplementary method

A.1. XAI-Aware Inversion Attack Models

We describe the architectures of our proposed models. Supplementary Figure 1 illustrates the meta-architectures for all XAI Input Methods for inversion attacks. Supplementary Figure 2 illustrates the detailed neural network layer architectures of key inversion models. Supplementary Tables 2-8 describe details of the layer settings for various target and attack models, describing convolutional layers (conv), max pooling layers (pool), fully connected layers (fc), and transposed convolutional layers with large strides (upsample).



Supplementary Figure 1. Architectures of XAI input inversion attack models: a) Baseline threat model, b) Flattened $\tilde{\boldsymbol{E}}_t$ concatenated with $\tilde{\boldsymbol{y}}_t$, c) CNN for dimensionality reduction, d) U-Net for dimensionality reduction and spatial knowledge, e) Combined Flatten+U-Net, f) Combined Flatten+U-Net on multiple explanations as a 3D tensor.



Supplementary Figure 2. Detailed architectures of inversion attack models for different XAI Input Methods. a) Prediction only TCNN with alternating transposed conv (upsample) and conv layers from $\tilde{\boldsymbol{y}}_t$ only. b) Flatten(CAM) method first flattens the explanation $\tilde{\boldsymbol{E}}_t$ pixels as a 1D vector and concatenates with $\tilde{\boldsymbol{y}}_t$, then upsamples with the same TCNN layers as (a). c) CNN(CAM) method reduces the dimension in $\tilde{\boldsymbol{E}}_t$ with several conv and pool layers to produce a 1D embedding vector that is concatenated with $\tilde{\boldsymbol{y}}_t$. d) U-Net(CAM) method adds to CNN method with bypass connectors from each conv layer in the CNN to corresponding conv layer in the TCNN. For CAM explanations which are smaller than images , there are no bypass connectors for later larger TCNN layers. e) Flatten+U-Net(Gradient) method with more conv and pool layers due to the larger pixel size of the gradient explanation, with 1D embedding concatenated with Flatten(Gradient) and $\tilde{\boldsymbol{y}}_t$. Since the gradient explanation $\tilde{\boldsymbol{E}}_t$ and reconstructed image output are in the same size , the U-Net has all bypass connectors.

Type	Kernel	Stride	Padding	Feature Map	Outputs
input				128×128	1
conv	3×3	1×1	1	128×128	128
pool	2×2	2×2	0	64×64	128
conv	3×3	1×1	1	64×64	256
pool	2×2	2×2	0	32×32	256
conv	3×3	1×1	1	32×32	512
pool	2×2	2×2	0	16×16	512
fc					512
fc					$\|C\|$

Supplementary Table 1. Network layers for iCV-MEFED target models. $|C|$ is the number of classes.

Type	Kernel	Stride	Padding	Feature Map	Outputs
input				256×256	1
conv	3×3	1×1	1	256×256	128
pool	2×2	2×2	0	128×128	128
conv	3×3	1×1	1	128×128	256
pool	2×2	2×2	0	64×64	256
conv	3×3	1×1	1	64×64	512
pool	2×2	2×2	0	32×32	512
conv	3×3	1×1	1	32×32	1024
pool	2×2	2×2	0	16×16	1024
fc					1024
fc					$\|C\|$

Supplementary Table 2. Network layers for CelebA target models. $|C|$ is the number of classes.

Type	Kernel	Stride	Padding	Feature Map	Outputs
input				32×32	1
conv	3×3	1×1	1	32×32	128
pool	2×2	2×2	0	16×16	128
conv	3×3	1×1	1	16×16	256
pool	2×2	2×2	0	8×8	256
fc					512
fc					$\|C\|$

Supplementary Table 3. Network layers for MNIST target model.

Type	Kernel	Stride	Padding	Feature Map	Outputs
input					$\|C\|$
fc					$\|C\|$
upsample	4×4	1×1	0	4×4	1024
conv	3×3	1×1	1	4×4	1024
upsample	4×4	2×2	1	8×8	512
conv	3×3	1×1	1	8×8	512
upsample	4×4	2×2	1	16×16	256
conv	3×3	1×1	1	16×16	256
upsample	4×4	2×2	1	32×32	128
conv	3×3	1×1	1	32×32	128
upsample	4×4	2×2	1	64×64	64
conv	3×3	1×1	1	64×64	64
upsample	4×4	2×2	1	128×128	1
conv	3×3	1×1	1	128×128	1

Supplementary Table 4. Network layers for Prediction only inversion attack model on iCV-MEFED(Supplementary Figure 2a).

Type	Kernel	Stride	Padding	Feature Map	Outputs
input($\tilde{\boldsymbol{y}}_t$)					$\|C\|$
input($\tilde{\boldsymbol{E}}_t$)				16×16	1
fc($\tilde{\boldsymbol{y}}_t$,$\tilde{\boldsymbol{E}}_t$)				1×1	$\|C\| + 16^2$
upsample	4×4	1×1	0	4×4	1024
conv	3×3	1×1	1	4×4	1024
upsample	4×4	2×2	1	8×8	512
conv	3×3	1×1	1	8×8	512
upsample	4×4	2×2	1	16×16	256
conv	3×3	1×1	1	16×16	256
upsample	4×4	2×2	1	32×32	128
conv	3×3	1×1	1	32×32	128
upsample	4×4	2×2	1	64×64	64
conv	3×3	1×1	1	64×64	64
upsample	4×4	2×2	1	128×128	1
conv	3×3	1×1	1	128×128	1

Supplementary Table 5. Network layers for Flatten(CAM) inversion attack model on iCV-MEFED(Supplementary Figure 2b).

Type	Kernel	Stride	Padding	Feature Map	Outputs
input($\tilde{\boldsymbol{y}}_t$)					$\|C\|$
input($\tilde{\boldsymbol{E}}_t$)				16×16	1
conv($\tilde{\boldsymbol{E}}_t$)	3×3	1×1	1	16×16	256
pool	2×2	2×2	0	8×8	256
conv	3×3	1×1	1	8×8	512
pool	2×2	2×2	0	4×4	512
conv	3×3	1×1	1	4×4	1024
pool	2×2	2×2	0	2×2	1024
fc					64
fc($\tilde{\boldsymbol{y}}_t$,conv)					$\|C\| + 64$
upsample	4×4	1×1	0	4×4	1024
conv	3×3	1×1	1	4×4	1024
upsample	4×4	2×2	1	8×8	512
conv	3×3	1×1	1	8×8	512
upsample	4×4	2×2	1	16×16	256
conv	3×3	1×1	1	16×16	256
upsample	4×4	2×2	1	32×32	128
conv	3×3	1×1	1	32×32	128
upsample	4×4	2×2	1	64×64	64
conv	3×3	1×1	1	64×64	64
upsample	4×4	2×2	1	128×128	1
conv	3×3	1×1	1	128×128	1

Supplementary Table 6. Network layers for CNN(CAM) inversion attack model on iCV-MEFED(Supplementary Figure 2c).

Type	Kernel	Stride	Padding	Feature Map	Outputs
input($\tilde{\boldsymbol{y}}_t$)					$\|C\|$
input($\tilde{\boldsymbol{E}}_t$)				16×16	1
conv($\tilde{\boldsymbol{E}}_t$)	3×3	1×1	1	16×16	256
pool	2×2	2×2	0	8×8	256
conv	3×3	1×1	1	8×8	512
pool	2×2	2×2	0	4×4	512
conv	3×3	1×1	1	4×4	1024
pool	2×2	2×2	0	2×2	1024
fc					64
fc($\tilde{\boldsymbol{y}}_t$,conv)					$\|C\| + 64$
upsample	4×4	1×1	0	4×4	1024
conv*	3×3	1×1	1	4×4	1024
upsample	4×4	2×2	1	8×8	512
conv*	3×3	1×1	1	8×8	512
upsample	4×4	2×2	1	16×16	256
conv*	3×3	1×1	1	16×16	256
upsample	4×4	2×2	1	32×32	128
conv	3×3	1×1	1	32×32	128
upsample	4×4	2×2	1	64×64	64
conv	3×3	1×1	1	64×64	64
upsample	4×4	2×2	1	128×128	1
conv	3×3	1×1	1	128×128	1

Supplementary Table 7. Network layers for U-Net(CAM) inversion attack model. conv* indicates connected via bypass connector from CNN conv of same feature map size. Refer to Supplementary Figure 2d for details of bypass connectors.

Type	Kernel	Stride	Padding	Feature Map	Outputs
input($\tilde{\boldsymbol{y}}_t$)					$\|C\|$
input($\tilde{\boldsymbol{E}}_t$)				128×128	1
conv($\tilde{\boldsymbol{E}}_t$)	3×3	1×1	1	128×128	1
pool	2×2	2×2	0	64×64	1
conv	3×3	1×1	1	64×64	64
pool	2×2	2×2	0	32×32	64
conv	3×3	1×1	1	32×32	128
pool	2×2	2×2	0	16×16	128
conv	3×3	1×1	1	16×16	256
pool	2×2	2×2	0	8×8	256
conv	3×3	1×1	1	8×8	512
pool	2×2	2×2	0	4×4	512
conv	3×3	1×1	1	4×4	1024
pool	2×2	2×2	0	2×2	1024
fc					64
fc($\tilde{\boldsymbol{y}}_t$,conv, $\tilde{\boldsymbol{E}}_t$)					$\|C\| + 64 + 128^2$
upsample	4×4	1×1	0	4×4	1024
conv*	3×3	1×1	1	4×4	1024
upsample	4×4	2×2	1	8×8	512
conv*	3×3	1×1	1	8×8	512
upsample	4×4	2×2	1	16×16	256
conv*	3×3	1×1	1	16×16	256
upsample	4×4	2×2	1	32×32	128
conv*	3×3	1×1	1	32×32	128
upsample	4×4	2×2	1	64×64	64
conv*	3×3	1×1	1	64×64	64
upsample	4×4	2×2	1	128×128	1
conv*	3×3	1×1	1	128×128	1

Supplementary Table 8. Network layers Flatten+U-Net(Gradient) inversion attack model. conv* indicates connected via bypass connector from CNN conv of same feature map size. Refer to Supplementary Figure 2e for details of bypass connectors.

B. Supplementary results

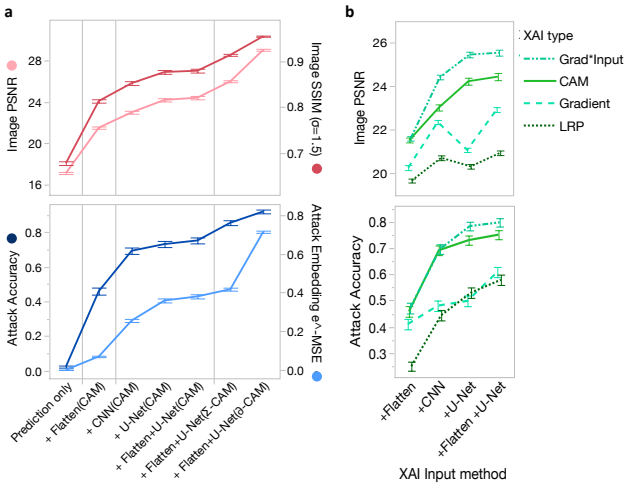
This section adds results with other metrics, from the Fredrikson inversion model [11] baseline and more demonstration instances for each dataset.

B.1. Attacking with different XAI Input Methods

We calculated the peak signal-to-noise ratio (PSNR) as another popular image similarity metric, where PSNR $= log_{10}(\text{MAX}^2/\text{MSE})$ and MAX refers to the dynamic range of the image pixel (255 for 8-bit greyscale images). Supplementary Figure 3a shows the inversion performance with PSNR replacing $1 - \text{MSE}$ in Figure 4a in the main paper.

In addition to Pixelwise Similarity $1 - \text{MSE}_x$ and Attack Embedding Similarity $e^{-\text{MSE}_s}$ reported in Figure 4b of the main paper, we report Image PSNR and Attack Accuracy as alternative metrics for inversion attack performance (Supplementary Figure 3b).

Supplementary Figure 6 demonstrates more instances with explanations and image reconstructions across all XAI Input Methods.

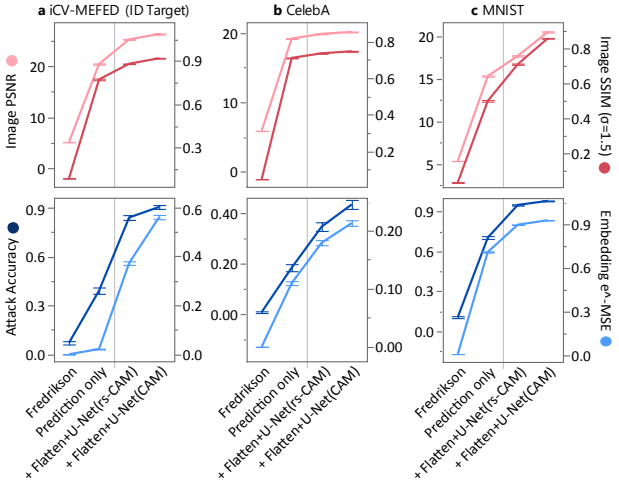


Supplementary Figure 3. Inversion attack performance for different XAI input methods and XAI types of iCV-MEFED [28] (Emotion target task). PSNR replacing $1 - \text{MSE}$. Error bars indicate 90% confidence interval.

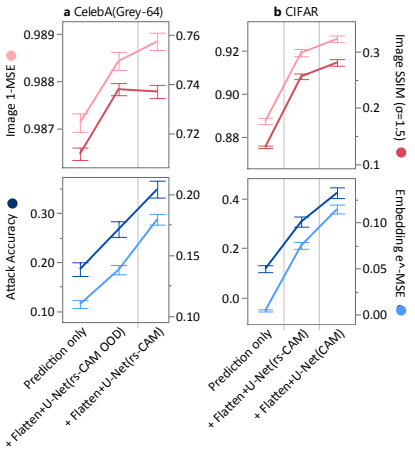
B.2. Attacking non-explainable target models

Inversion performance improved in the order: Fredrikson < Prediction only < rs-CAM < CAM (Supplementary Figure 4). Fredrikson has poorest performance because it performs class inversions (i.e., only one reconstructed image per class), while the other approaches are instance inversions (i.e., reconstruct differently per instance). For ecological validity regarding privacy, we evaluated the inversion attacks on out-of-distribution (OOD) data (FaceScrub citeng2014data) and a real-world photo dataset (CIFAR-10 [22]). Supplementary Fig. 5a shows that attack models trained with OOD data still perform improved inversion attack compared to prediction only, albeit weaker than models trained with independent and identically distributed (IID) data [27]. Supplementary Fig. 5b shows similar inversion attack performance with CIFAR compared to iCV-MEFED, CelebA, and MNIST.

Supplementary Figures 7-11 demonstrate explanations and image reconstructions more demonstration instances for baseline and XAI-aware, and XAI Input methods for non-explainable and explainable target models.

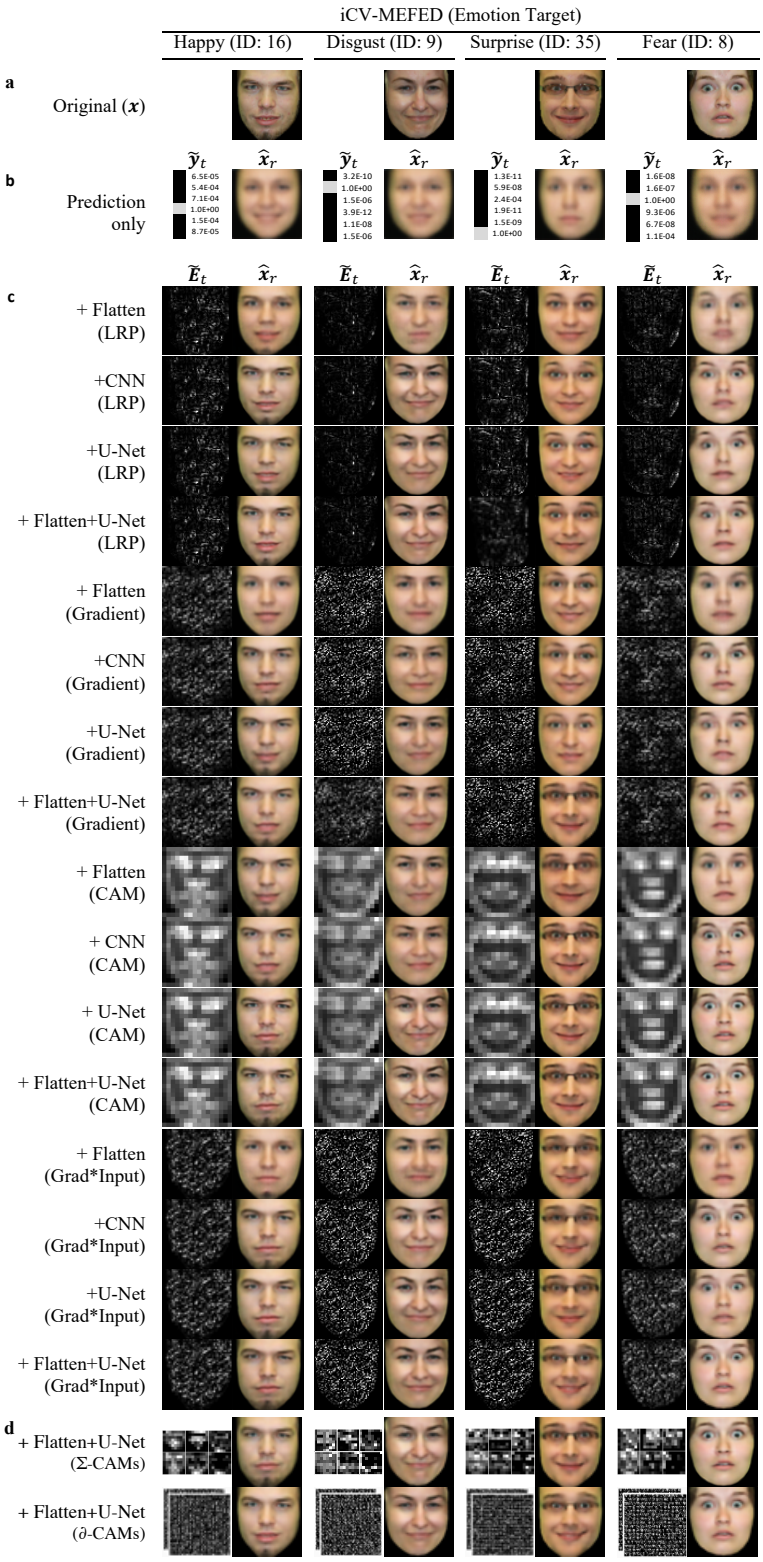


Supplementary Figure 4. Inversion attack performance across different datasets showing increased privacy risk when exploiting target explanations (CAM) and with attention transfer. Two non-XAI-aware baselines Fredrikson [11] and Prediction only [50] are significantly poorer than XAI-aware inversions. Error bars indicate 90% confidence interval.

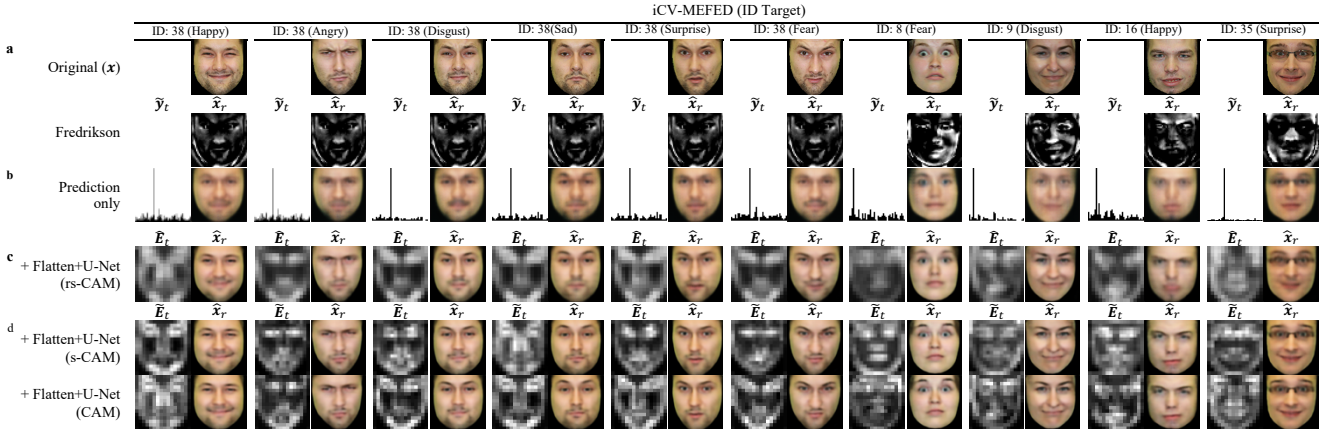


Supplementary Figure 5. Inversion attack performance for attack model trained on OOD data (a) and for CIFAR data (b) showing increased privacy risk when exploiting target explanations (CAM) and with attention transfer. Error bars indicate 90% confidence interval.

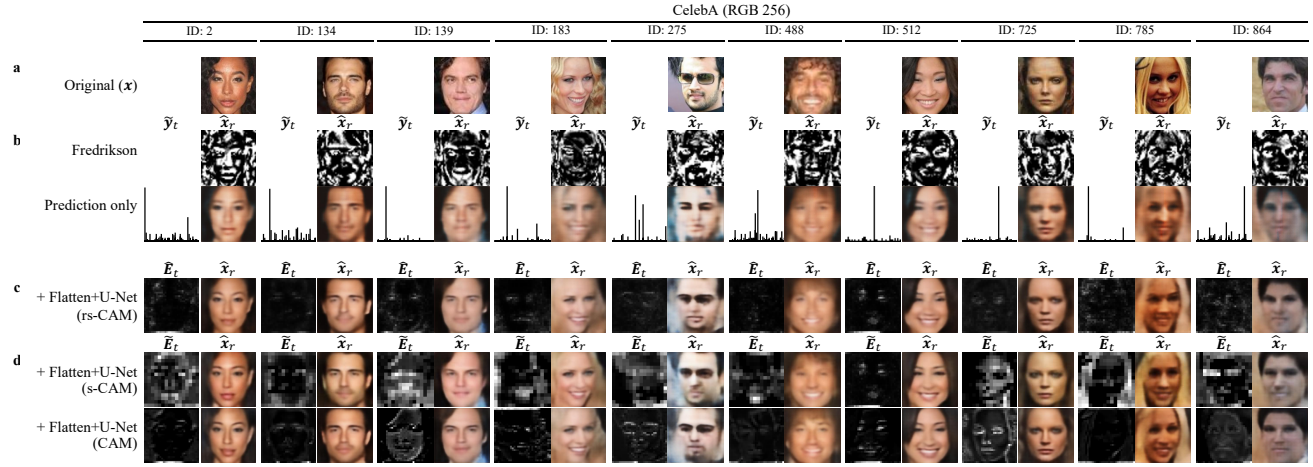
Supplementary Figure 6. Demonstration of image reconstruction from XAI-aware inversion attack with emotion prediction as the target task, and face reconstruction as the attack task. Six emotions of a single identity come from the iCV-MEFED dataset [28]. Reconstructed images are shown with corresponding information (ie. target prediction $\tilde{\boldsymbol{y}}_t$, explanations $\tilde{\boldsymbol{E}}_t$ as Gradients [43], Grad-CAM [39] or Gradient $\odot$ Input [42] saliency maps). Towards original images (a), reconstructions from Prediction only (b) are poor and similar across different faces, and are significantly improved when exploiting single (c) and multiple (d) explanations. Fredrikson baseline not used, since it is a class-based inversion and will only have the same reconstructed image for each emotion class, regardless of face identity or instance.



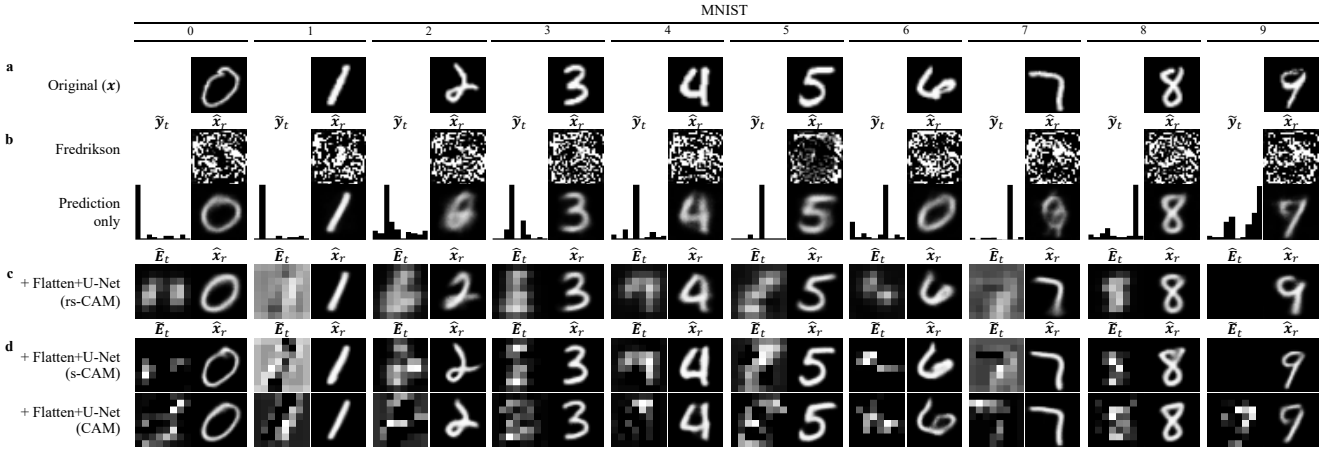
Supplementary Figure 6 (Cont.). Demonstration of image reconstruction from XAI-aware inversion attack with emotion prediction as the target task, and face reconstruction as the attack task. Four emotions of different identities come from the iCV-MEFED dataset [28].



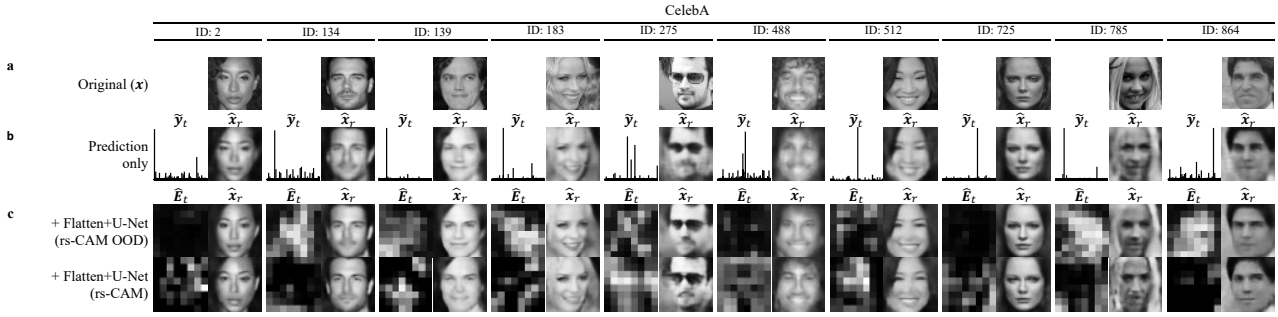
Supplementary Figure 7. Demonstration of image reconstruction with baseline and XAI-aware inversion attack models for iCV-MEFED [28] with identification as target and attack tasks.



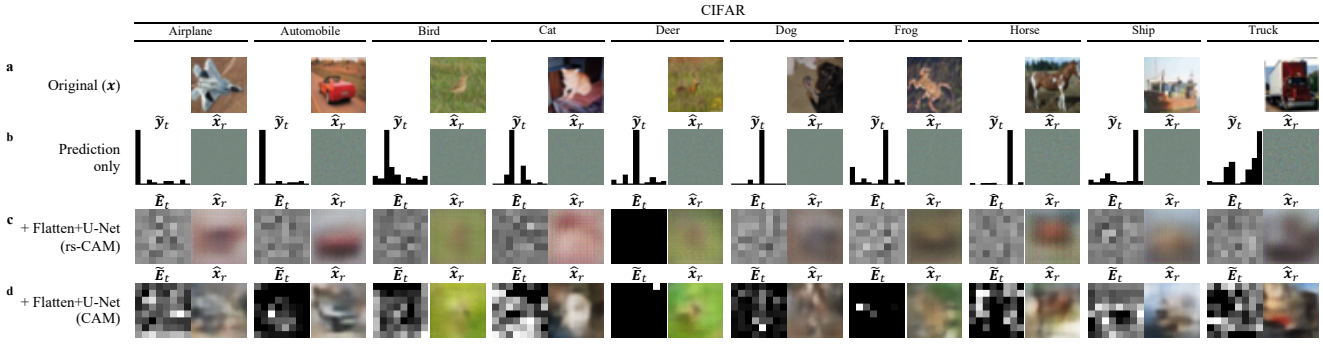
Supplementary Figure 8. Demonstration of image reconstruction with baseline and XAI-aware inversion attack models for CelebA [27] with identification as target and attack tasks.



Supplementary Figure 9. Demonstration of image reconstruction with baseline and XAI-aware inversion attack models for MNIST [24] with handwriting digit recognition as target and attack tasks.



Supplementary Figure 10. Demonstration of image reconstruction with baseline and XAI-inversion attack models trained on OOD data (FaceScrub [31]) dataset to attack a target model (trained on (CelebA [27]) .



Supplementary Figure 11. Demonstration of image reconstruction with baseline and XAI-aware inversion attack models for CIFAR-10 with recognition as target and attack tasks.