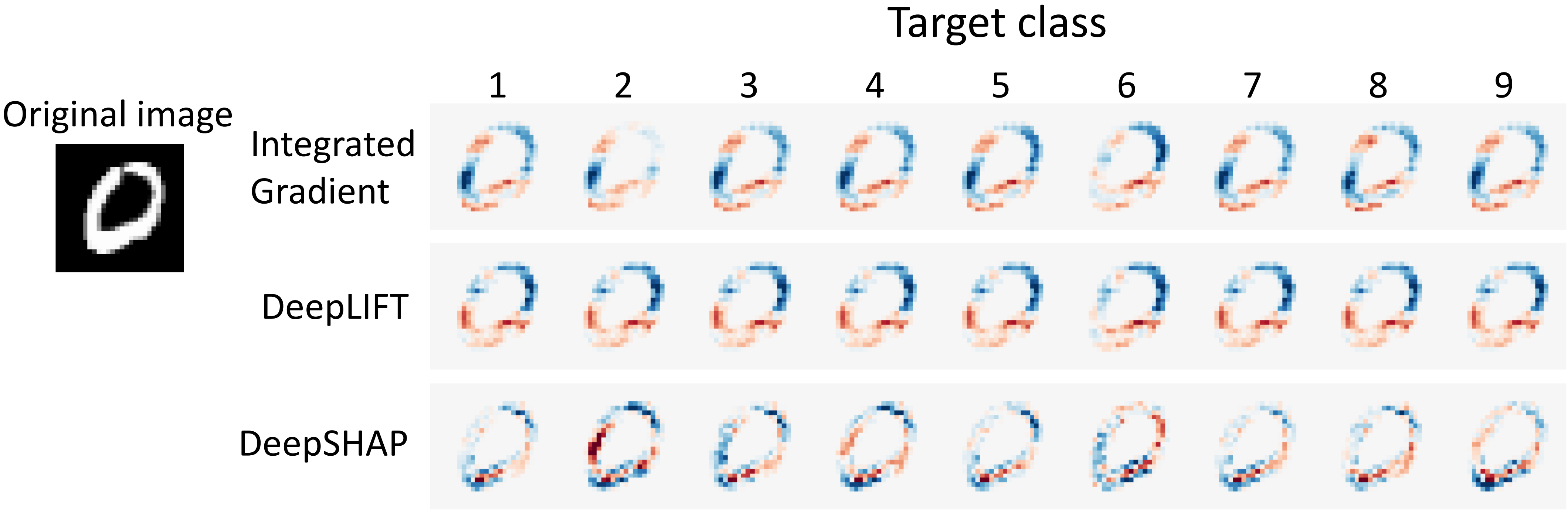
GANMEX: One-vs-One Attributions Guided by GAN-based Counterfactual Baselines

Sheng-Min Shih, Pin-Ju Tien, Zohar Karnin

Amazon Web Service

One-vs-One Explanations for Multi-class Classifiers

- One-vs-all: Why does the instance belong to class A?
- One-vs-one: Why does the instance belong to class A but not class B?



GANMEX: One-vs-One Attributions Guided by GAN-based Counterfactual Baselines

Feature Attribution Methods and Baselines

- **Integrated Gradient (IG)**:

$$\mathcal{IG}_i = (x_i - \boxed{\tilde{x}_i}) \int_{\alpha=0}^{1} \partial_{x_i} S(\tilde{x} + \alpha(x - \tilde{x}))d\alpha.$$

- **DeepLIFT**:

$$r_i^{(L)} = \begin{cases} S_i(x) - S_i\boxed{(\bar{x})} & \text{if unit } i \text{ is the target unit of interest} \\ 0 & \text{otherwise} \end{cases}$$

$$r_i^{(l)} = \sum_j \frac{z_{ji} - \bar{z}_{ji}}{\sum_{i'} z_{ji} - \sum_{i'} \bar{z}_{ji}} r_j^{(l+1)}$$

- **Occlusion**: full-feature perturbations by removing each feature and calculating the impacts on the DNN output.

All require baselines

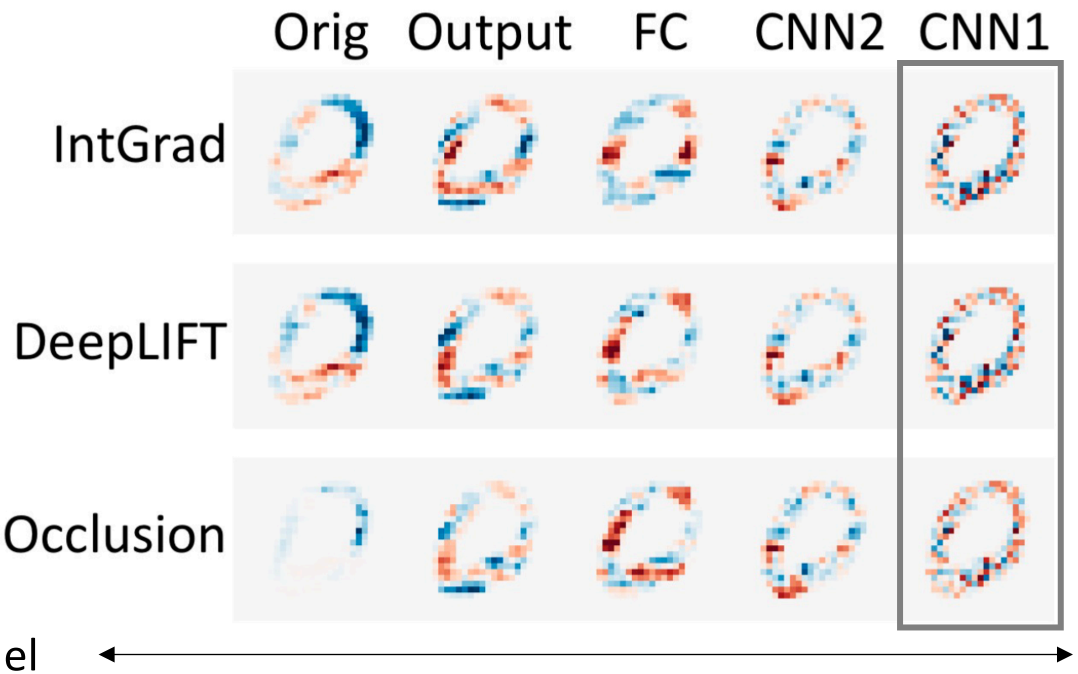
- **Expected Gradient**

$$\mathcal{EG}_i = \mathbb{E}_{\boxed{\tilde{x}} \sim X_T \alpha \sim U(0,1)} (x - \tilde{x})_i \partial_{x_i} S(\tilde{x} + \alpha(x - \tilde{x}))$$

GANMEX: One-vs-One Attributions Guided by GAN-based Counterfactual Baselines

Baseline Problems and the Failed Sanity Check

- Attributions methods are blind to the color chosen as a baseline (Sundararajan & Taly 2018; Adebayo et al. 2018; Kindermans et al. 2017; Sturmfels et al. 2020)
- Common used baselines includes zero values, max values, blurred images…
- Sanity Check
  (Adebayo et al. 2018)



Original model

Completely randomized model

Failed

Saliency maps should be completely randomized

One-vs-One Baseline Requirements

Class-targeted baselines are required for one-vs-one attribution.

1. The baseline belongs to the target class (with respect to the classifier).
2. The baseline is close to the original input.
3. The baseline is a realistic image.

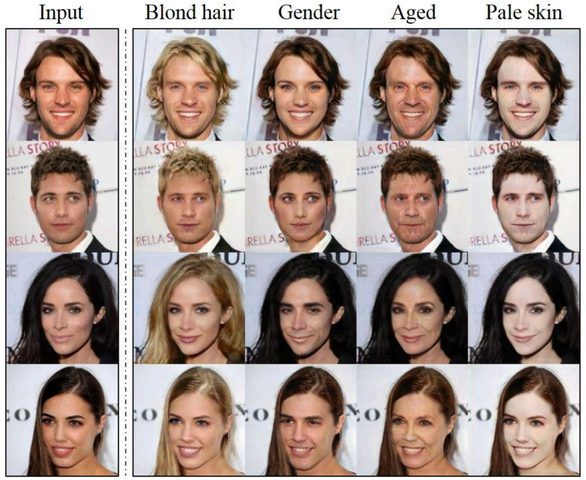
=> Counterfactual Explanation!

(Dhurandhar et al. 2018)

GANMEX (GAN-based Model EXplainability)

Baseline requirements

$$B_{c_t}(x) = \arg\min_{\tilde{x}\in\mathbb{R}^N} \left(\|x-\tilde{x}\| - \log R(\tilde{x}) - \log S_{c_t}(\tilde{x})\right)$$



StarGAN

$$\mathcal{L}_D = -\mathcal{L}_{\text{adv}} + \lambda_{\text{cls}}^{r}\mathcal{L}_{\text{cls}}^{r}$$

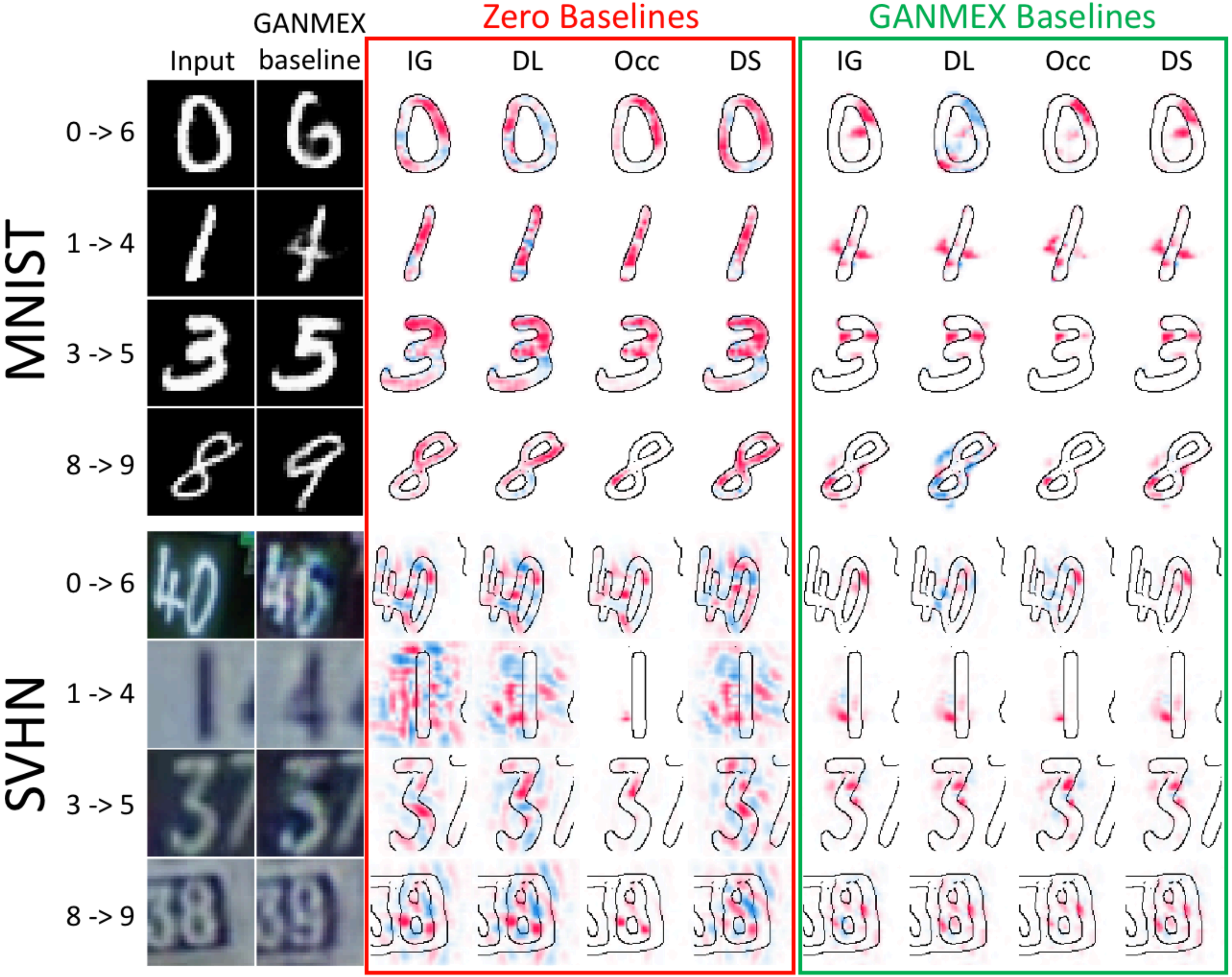
$$\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \lambda_{\text{cls}}^{f}\mathcal{L}_{\text{cls}}^{f} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}$$

GANMEX

$$\mathcal{L}_G = \log(1 - D_{\text{src}}(\tilde{x})) - \lambda_{\text{cls}}^{f}\log(\underset{\substack{\uparrow\\ \text{Trained classifier}}}{S_c}(\tilde{x})) + \lambda_{\text{rec}}\|x - G(\tilde{x}, c')\|_1 + \underset{\text{Similarity loss}}{\lambda_{\text{sim}}\|x - \tilde{x}\|_1}$$

GANMEX: One-vs-One Attributions Guided by GAN-based Counterfactual Baselines

GANMEX Results



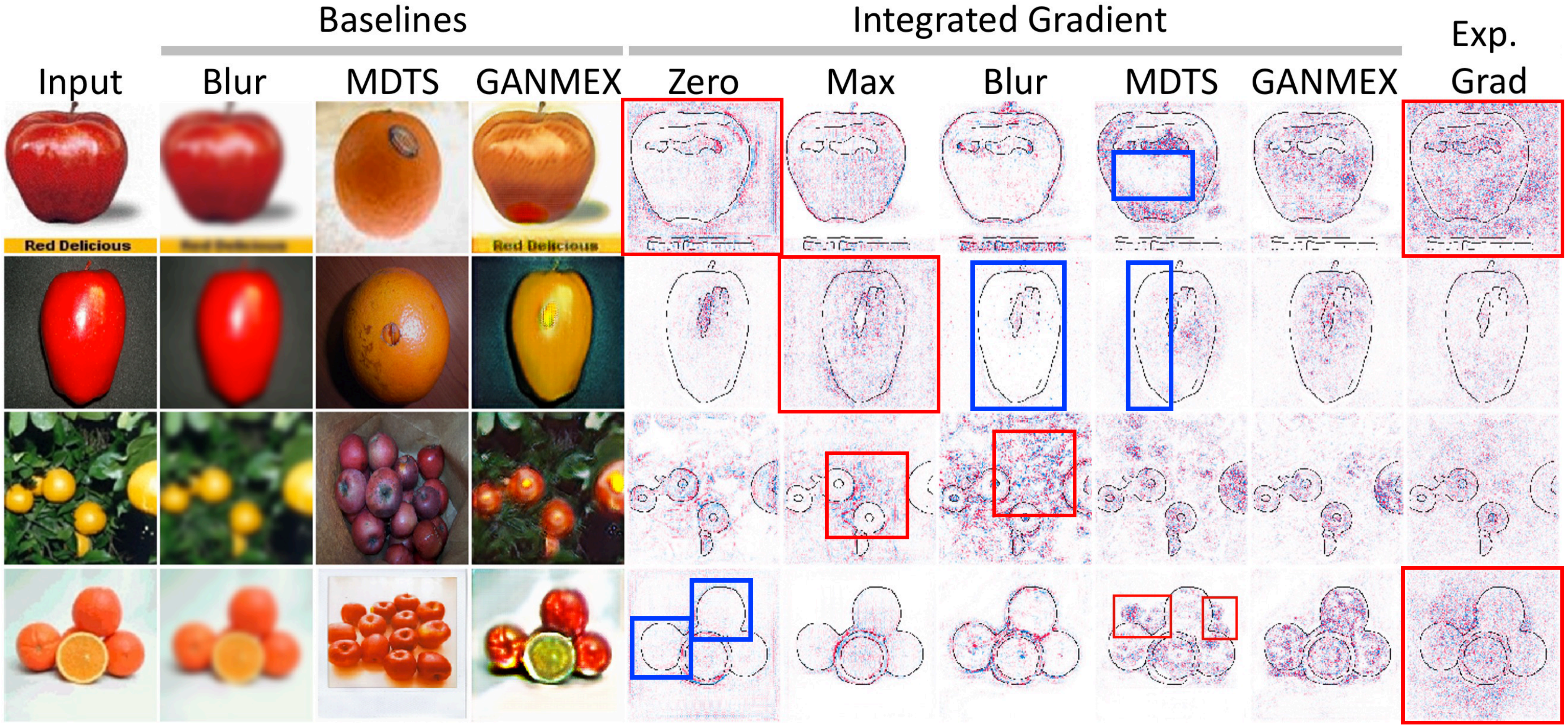
IG: Integrated Gradient

DL: DeepLIFT

Occ: Occlusion

DS: DeepSHAP

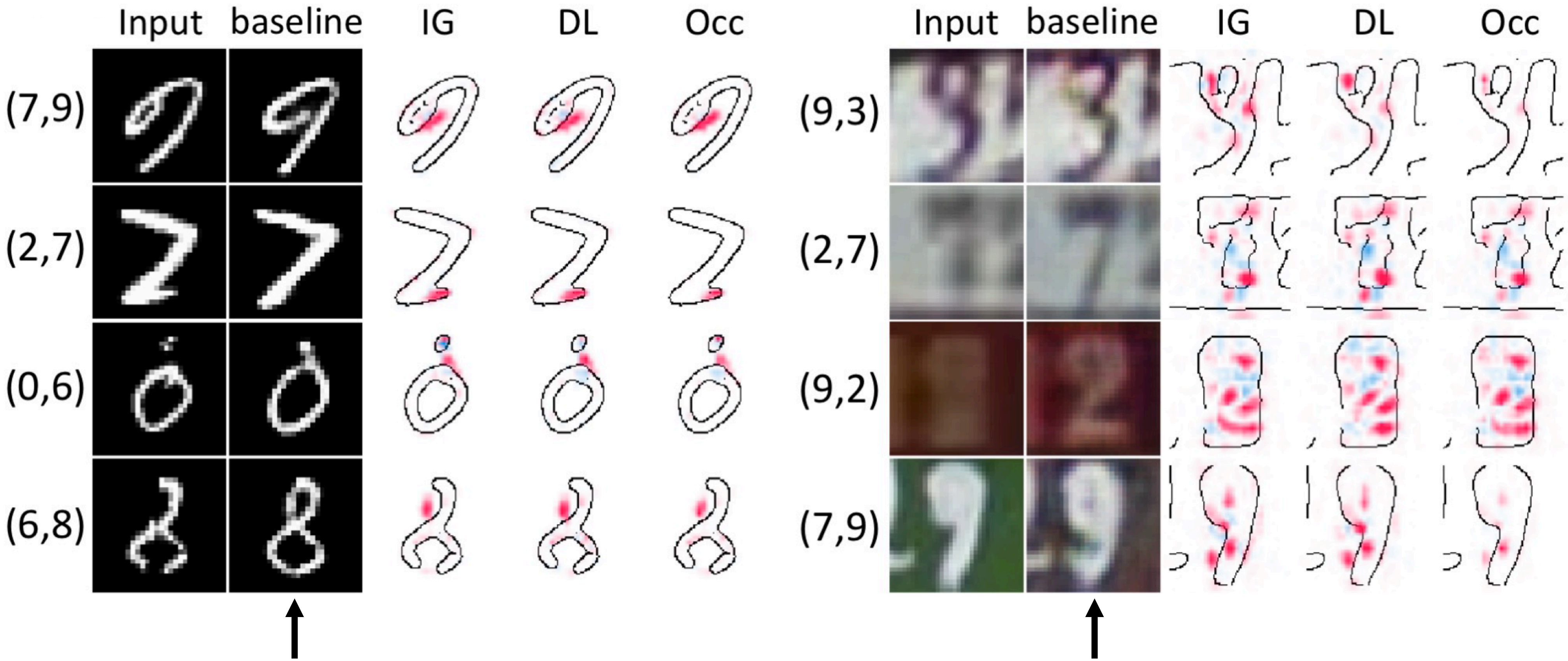
GANMEX vs. Other Baselines



□ Incorrected highlighted background □ Missed objects

Debugging for Mis-Classified Samples

Tuples are showing (*predicted label, correct label*)



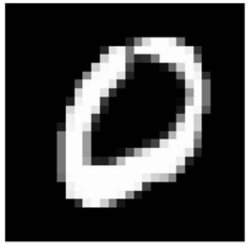
How correct images should look like according to the classifier

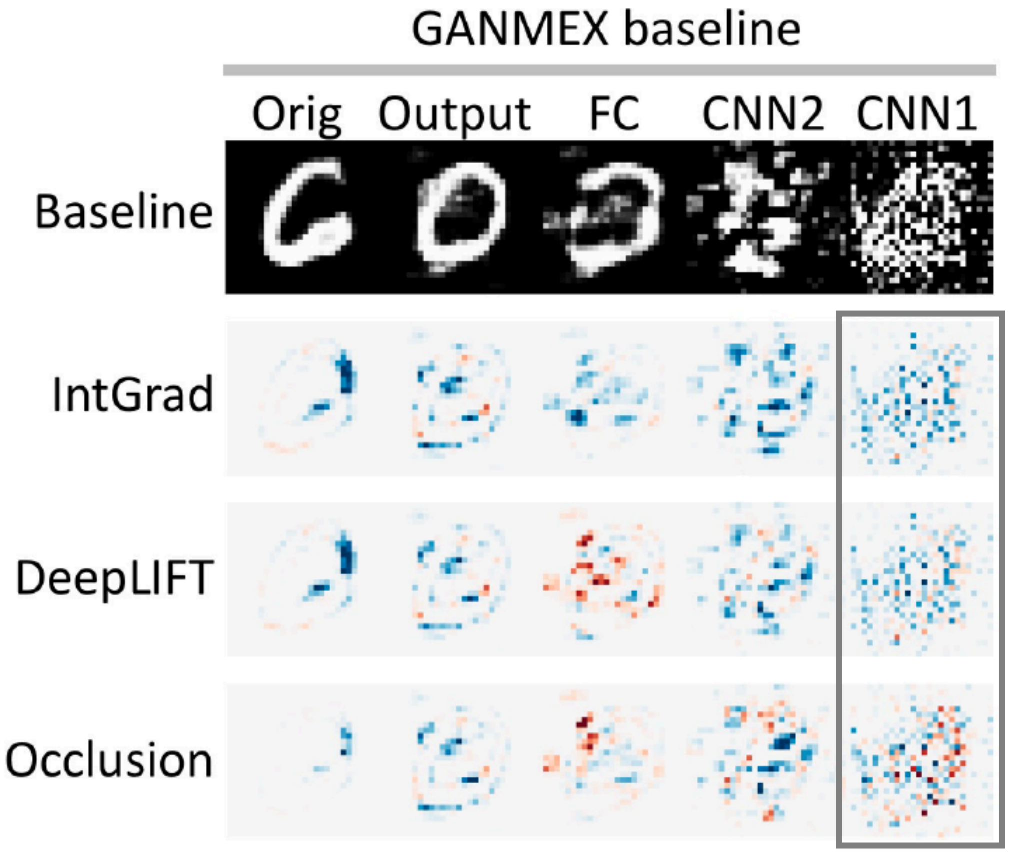
GANMEX: One-vs-One Attributions Guided by GAN-based Counterfactual Baselines

Cascading Randomization Sanity Checks

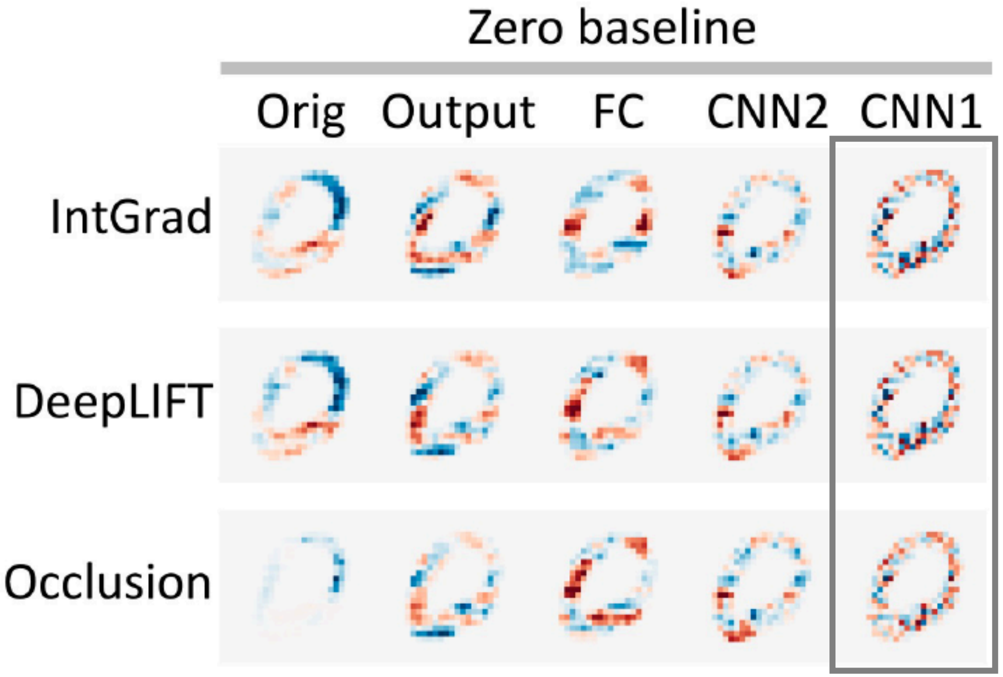
Original class: 0
Target class: 6

Input image





Pass



Failed

GANMEX: One-vs-One Attributions Guided by GAN-based Counterfactual Baselines

Conclusions

- GANMEX can be used with IG, DeepLIFT, Occlusion, DeepSHAP to improve the feature attribution.
- GANMEX addressed the failed sanity checks introduced by other baseline choices.
- Effective method for counterfactual explanations.