Friedrich-Alexander-Universität
Technische Fakultät

Shadows of Deception: Unveiling AI-generated Images Through Inconsistencies in Scene Lighting

Muhammad Abdullah

Supervisors: PD. Dr. Christian Riess, Prof. Dr. Bernhard Egger

Light in real scenes follows the laws of physics. Can data-driven AI-generated images learn these laws?

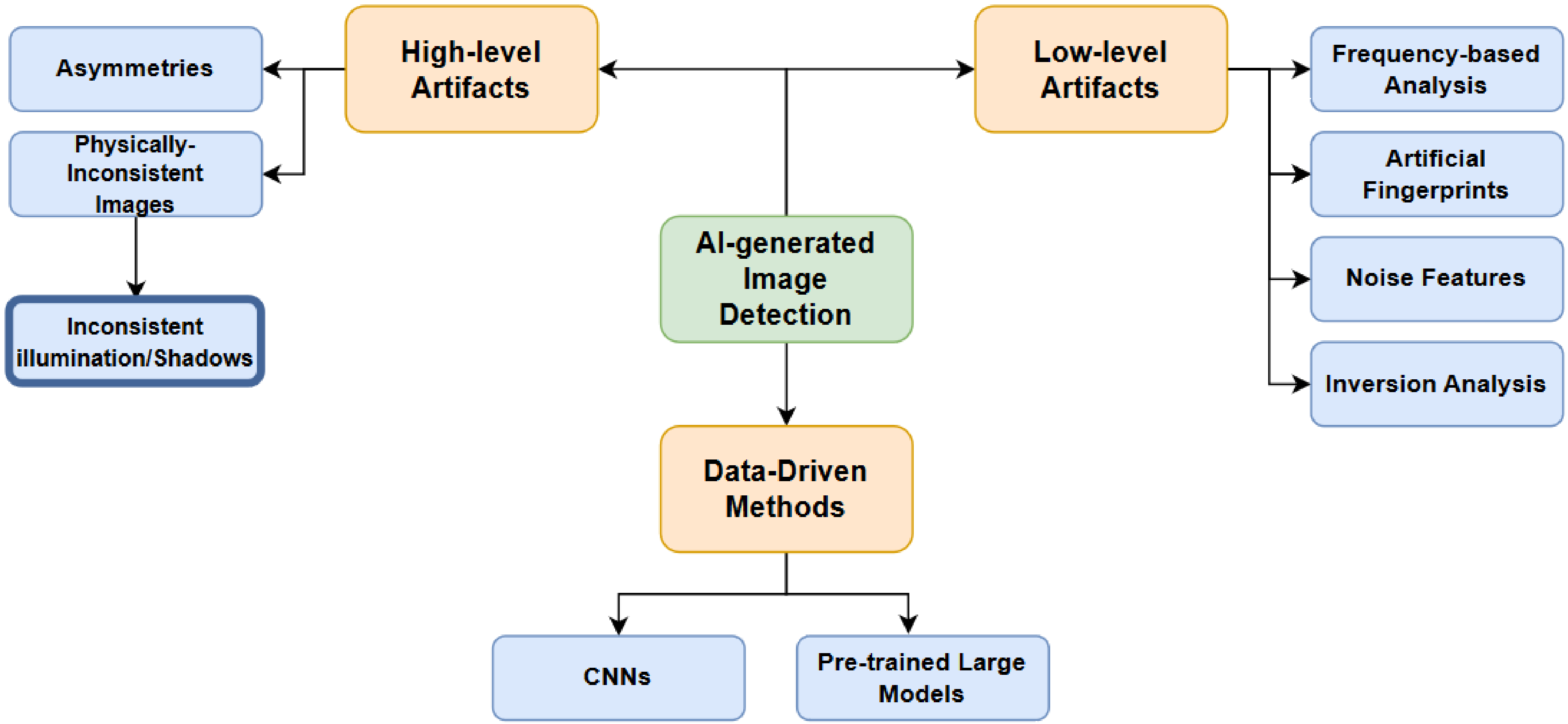
AI-generated image

AI-generated image

Outline

Related Work: Detecting AI-generated Images

Classification of different approaches used for detecting AI-generated images

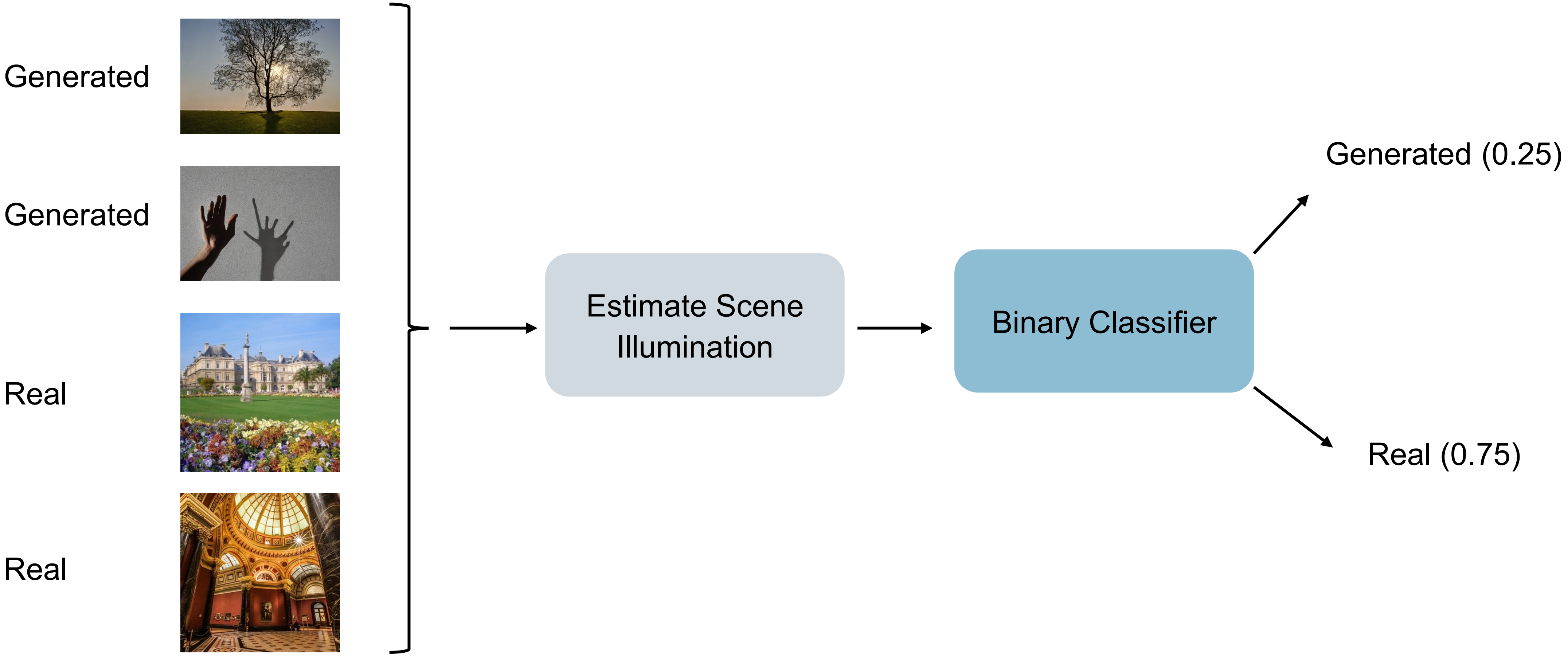


Classification of AI-generated images detection methods[1]

 1. Tariang et al. "Synthetic Image Verification in the Era of Generative AI"

Scene Illumination for Detecting AI-generated Images

Estimated scene illumination for real and generated images is passed to a binary classifier to classify real and generated images



Modeling Scene Illumination

Modeling Scene Illumination

Spherical harmonic lighting

The Rendering Equation[1]

$$L(\mathbf{x},\vec{\omega}_o) = L_e(\mathbf{x},\vec{\omega}_o) + \int_S f_r(\mathbf{x},\vec{\omega}_i \to \vec{\omega}_o) L(\mathbf{x}',\vec{\omega}_i) G(\mathbf{x},\mathbf{x}') V(\mathbf{x},\mathbf{x}') d\omega_i$$

where

$L(\mathbf{x},\vec{\omega}_o)$ = the intensity reflected from position $\mathbf{x}$ in direction ω_o

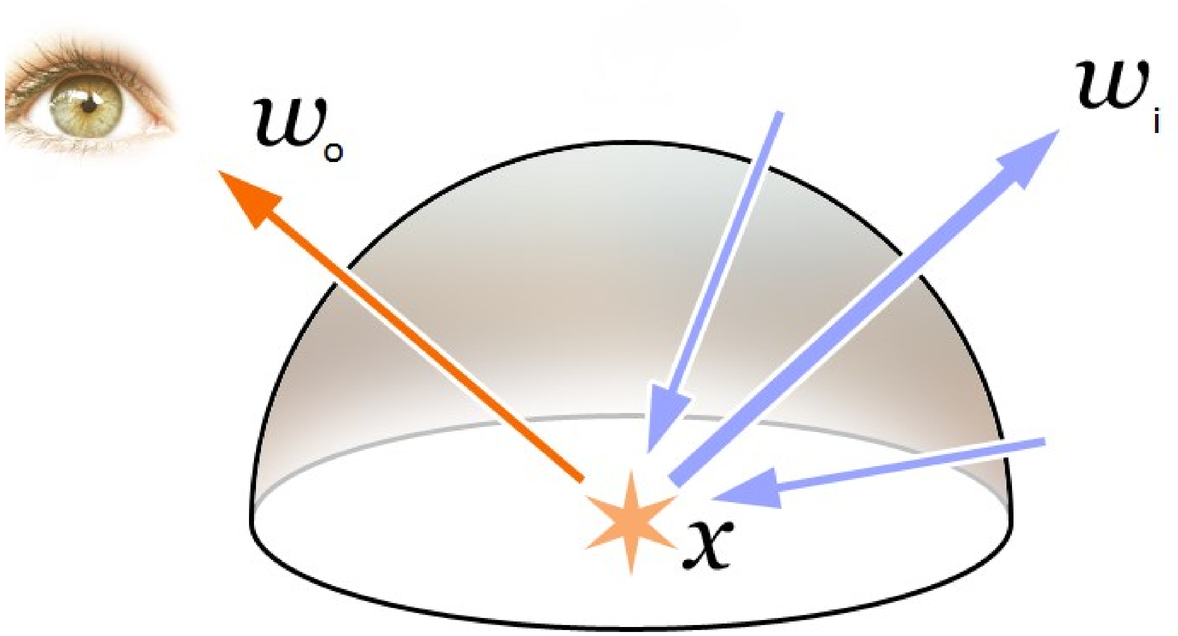
$L_e(\mathbf{x},\vec{\omega}_o)$ = the light emitted from $\mathbf{x}$ by this object itself

$f_r(\mathbf{x},\vec{\omega}_i \to \vec{\omega}_o)$ = the BRDF of the surface at point $\mathbf{x}$, transforming incoming light ω_i to reflected light ω_o

$L(\mathbf{x}',\vec{\omega}_i)$ = light from $\mathbf{x}'$ on another object arriving along ω_i

$G(\mathbf{x},\mathbf{x}')$ = the geometric relationship between $\mathbf{x}$ and $\mathbf{x}'$

$V(\mathbf{x},\mathbf{x}')$ = a visibility test, returns 1 if $\mathbf{x}$ can see $\mathbf{x}'$, 0 otherwise



https://en.wikipedia.org/wiki/Rendering_equation

Modeling Scene Illumination

Lighting as an irradiance environment map

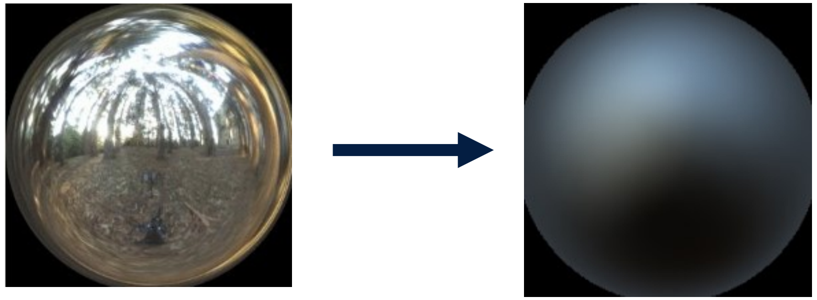
We make some assumptions: A diffused surface is illuminated by a distant light source L. We can rewrite the rendering equation:

$$L = f_r(x) \int_S L(w_i)(n \cdot w_i)\, dw_i$$

$$\text{Irradiance} = E(n) = \int_S L(w_i)(n \cdot w_i)\, dw_i$$

$$L = f_r(x)\, E(n)$$

Irradiance is parameterized by the surface normal only. It can be considered as a function defined over a sphere.



Environment map

Irradiance environment map

Modeling Scene Illumination

Irradiance as spherical harmonics

Each function defined on the surface of sphere can be written as a sum of spherical harmonics that form orthonormal basis.

Assuming Lambartian surface, we can write the irradiance as

$\boldsymbol{E(n) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} c_{lm} Y_{lm}(n(x)) = Bi}$

$\boldsymbol{L = f_r \odot Bi}$

$\boldsymbol{L = Mi} \longrightarrow$ Solve using least squares method

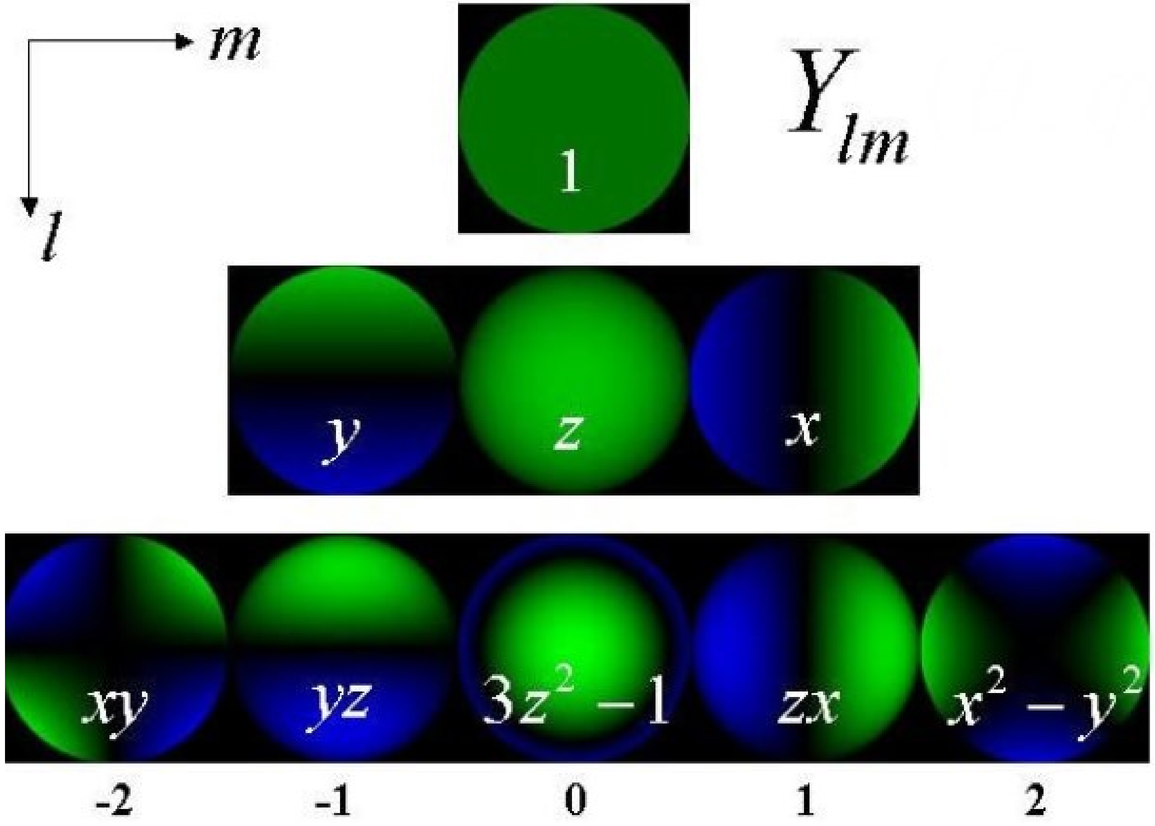
$\boldsymbol{c_{lm}}$ = Spherical harmonic coefficients

$\boldsymbol{Y_{lm}}$ = Spherical harmonic basis functions

n(*x*) = Surface normal at point x

i = Vector containing spherical harmonic coefficients

M = Matrix containing spherical harmonic basis



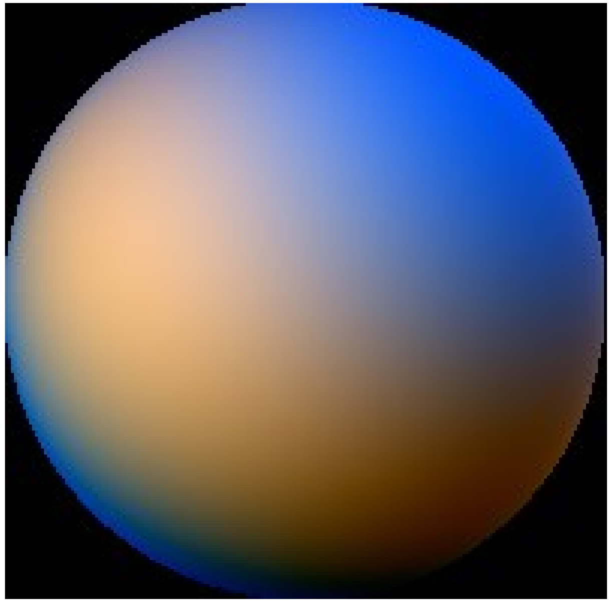
The first 3 orders of spherical harmonics[1]

Modeling Scene Illumination

Spherical harmonic lighting

Irradiance for a lambartian object can be well approximated by up to 2 orders of spherical harmonics[1].

- Low frequency estimation but fewer parameters
- Total 9 spherical harmonics for 1-channel images
- 27 for 3-channel images (9 for each channel)
- Spherical harmonics coefficients can be used as a proxy for scene lighting
- On Right, a lambartian sphere is lightened using spherical harmonic coefficients



An exemplary SH illumination

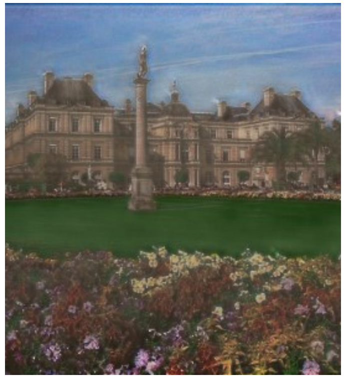
Inverse Rendering

Inverse Rendering

Decompose an image into albedo, surface normal, scene illumination, and shadows

Original Image

Inverse Rendering



Albedo map

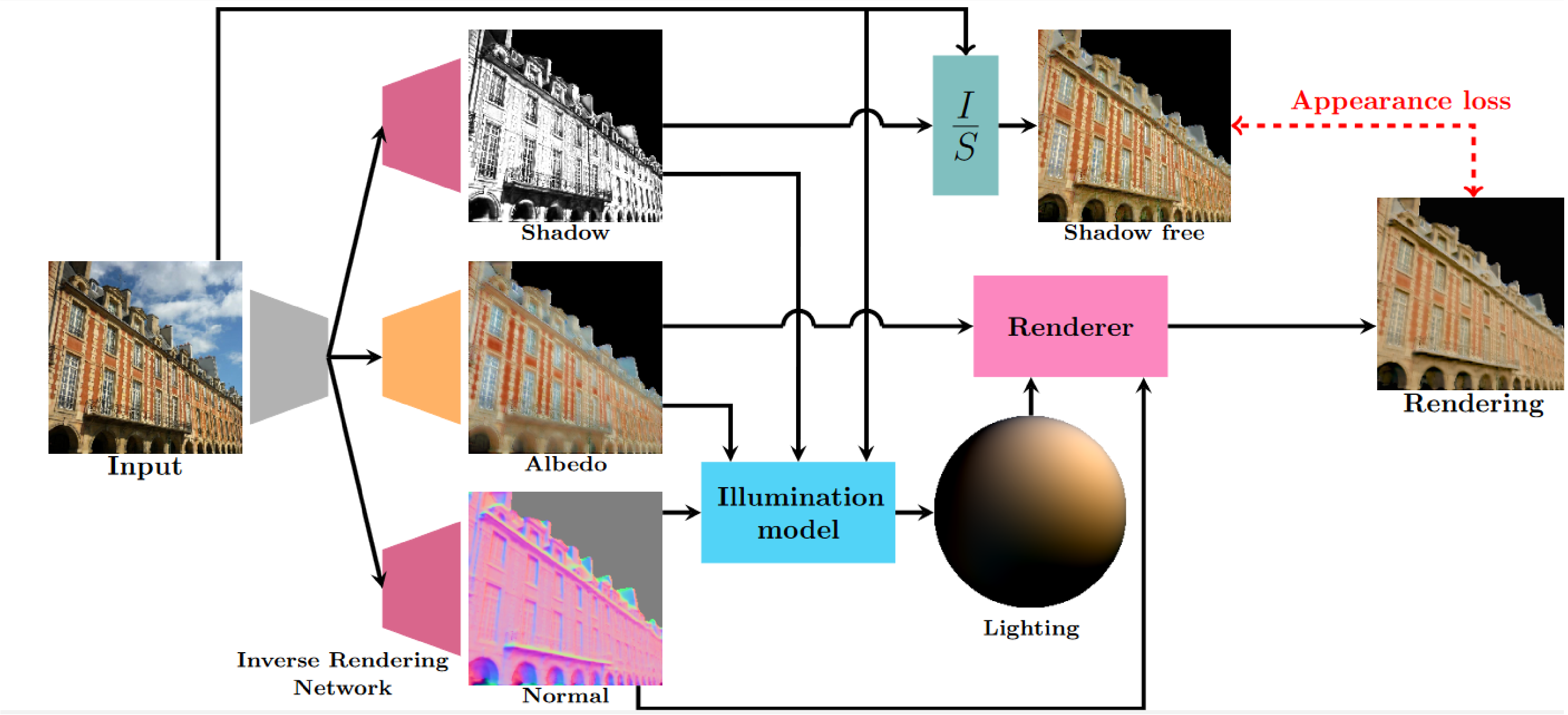
Scene illumination

Surface normal map

Shadow map

Inverse Rendering Network (IRN)

Directly regress albedo, surface normal and shadow maps from image using autoencoder

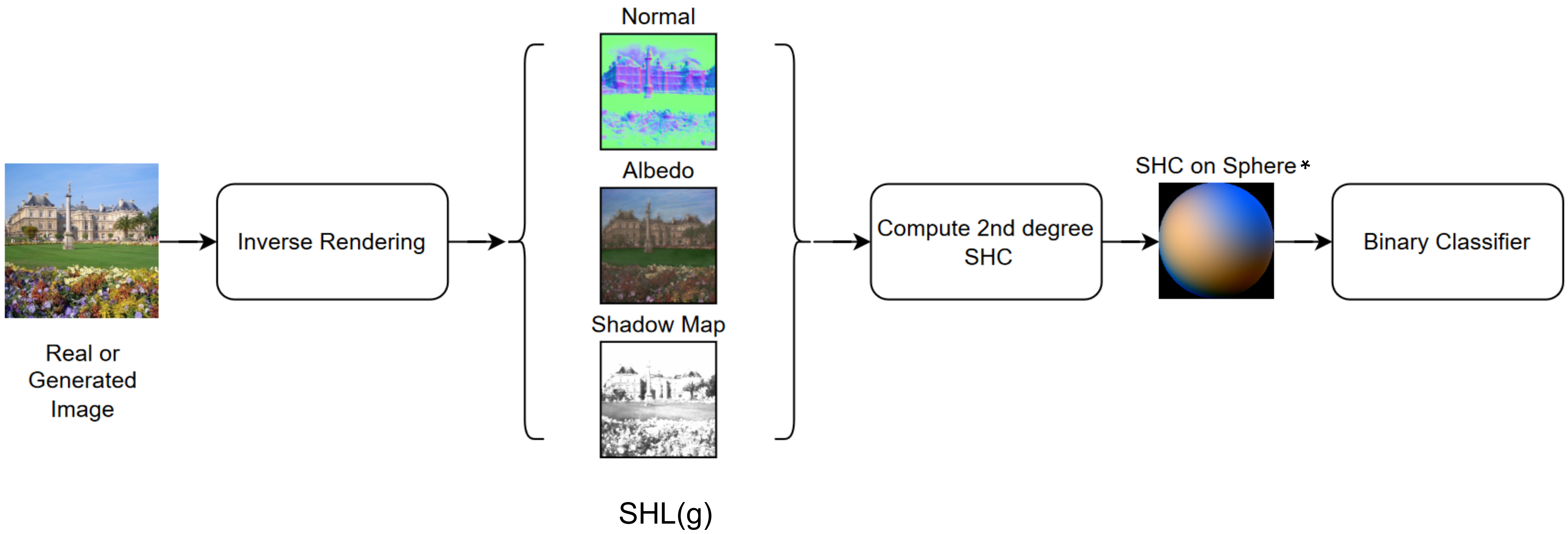


Inverse rendering network architecture[1]

Methodology

Proposed Pipeline

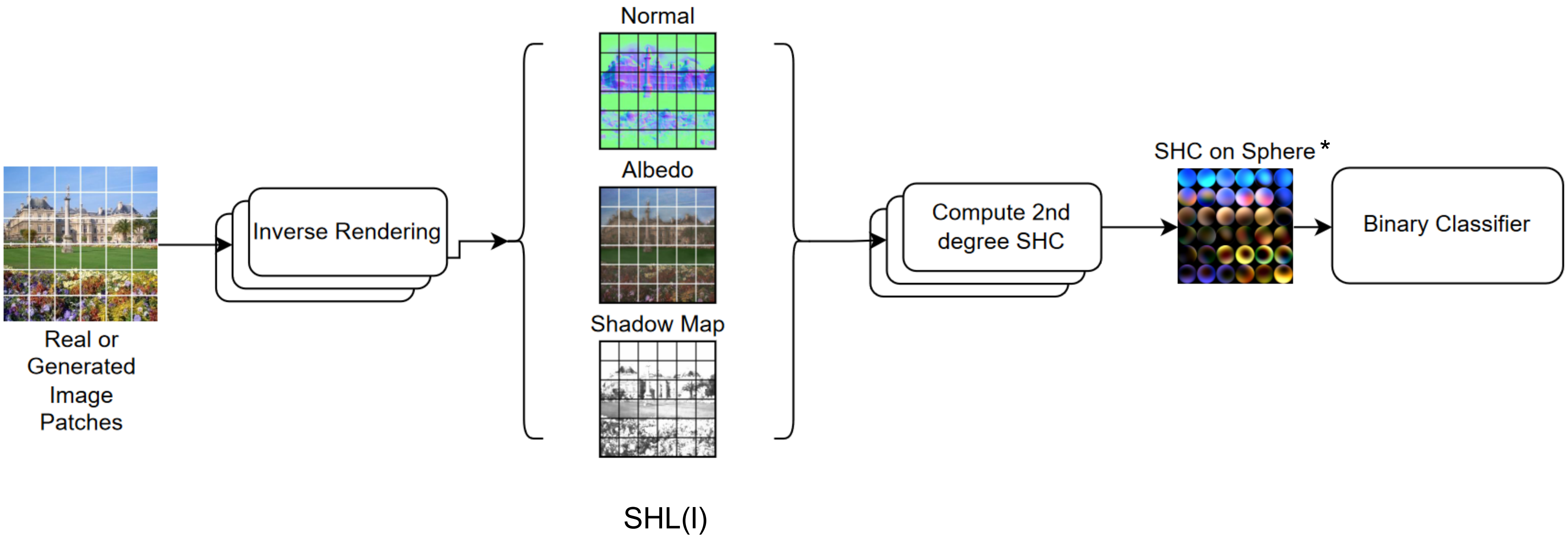
Compute 27 spherical harmonics coefficients for the whole image and pass them to different binary classifers



* This image is for visualisation only. We directly pass 27 spherical harmonics coefficients to the binary classifier.

Proposed Pipeline

Divide image into patches and compute spherical harmonics coefficients for each patch and pass them to the binary classifier



* This image is for visualisation only. We accumulate 27 spherical harmonics coefficients for each patch and pass them to the binary classifier.

Binary Classifiers

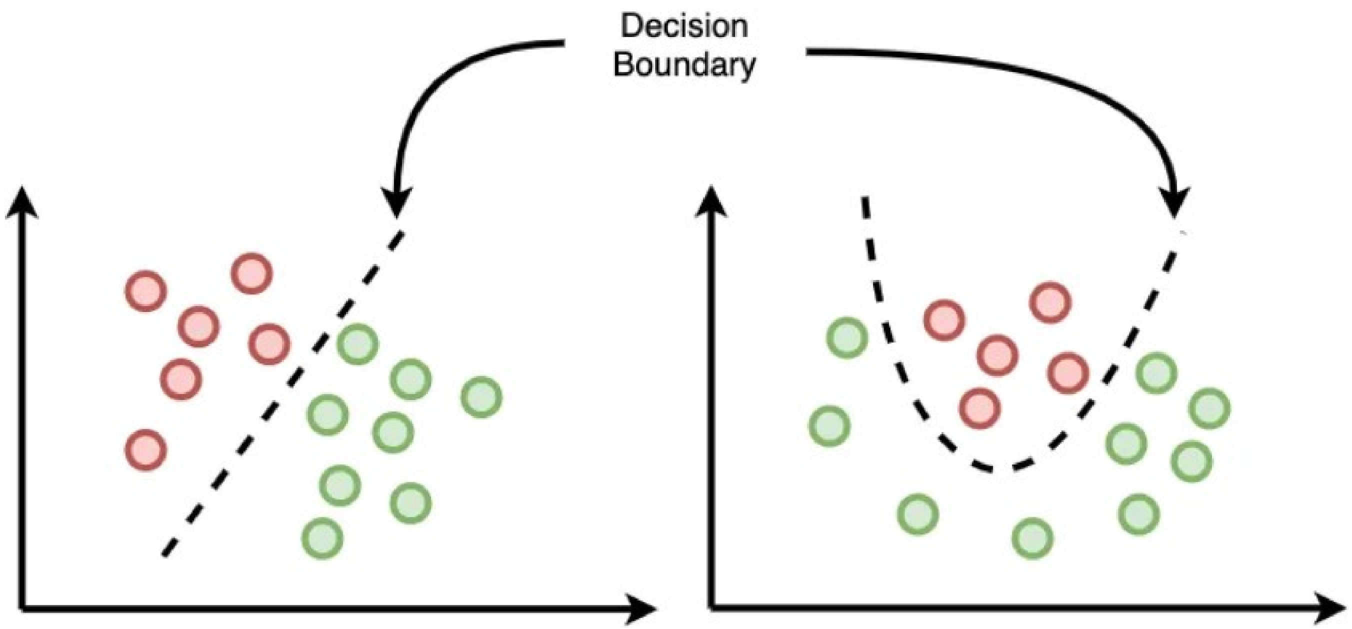
Different linear and non-linear binary classifiers are trained for classification

Traditional classifiers:

- Random Forest (RF)
- Support Vector Machine (SVM)
- Logistic Regression (LR)

Neural network-based classifiers:

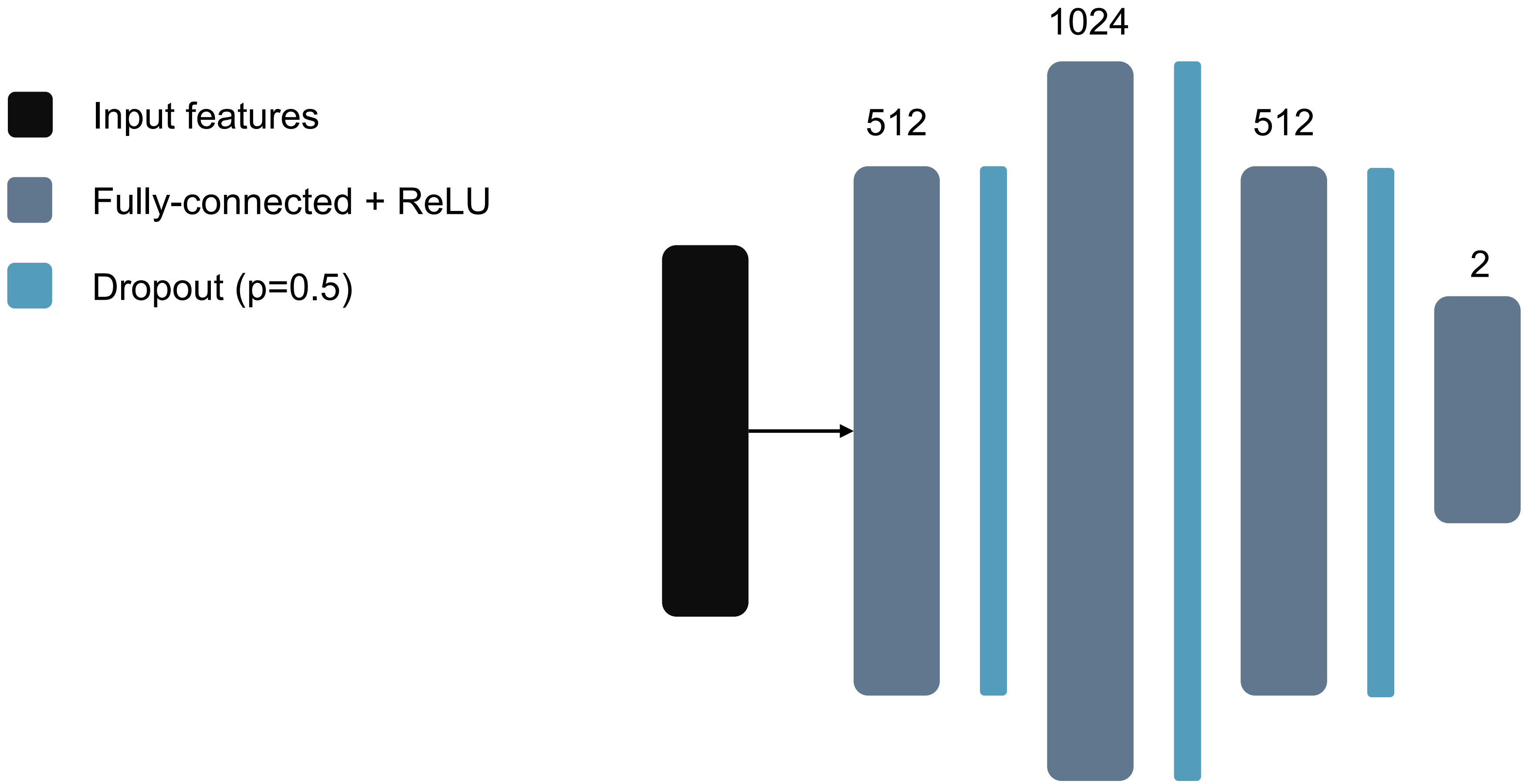
- MLP
- Modified Vision Transformer (ViT)



We use linear/non-linear binary classifiers to explore different decision boundaries.
https://towardsdatascience.com/logistic-regression-and-decision-boundary-eab6e00c1e8

Binary Classification with MLP

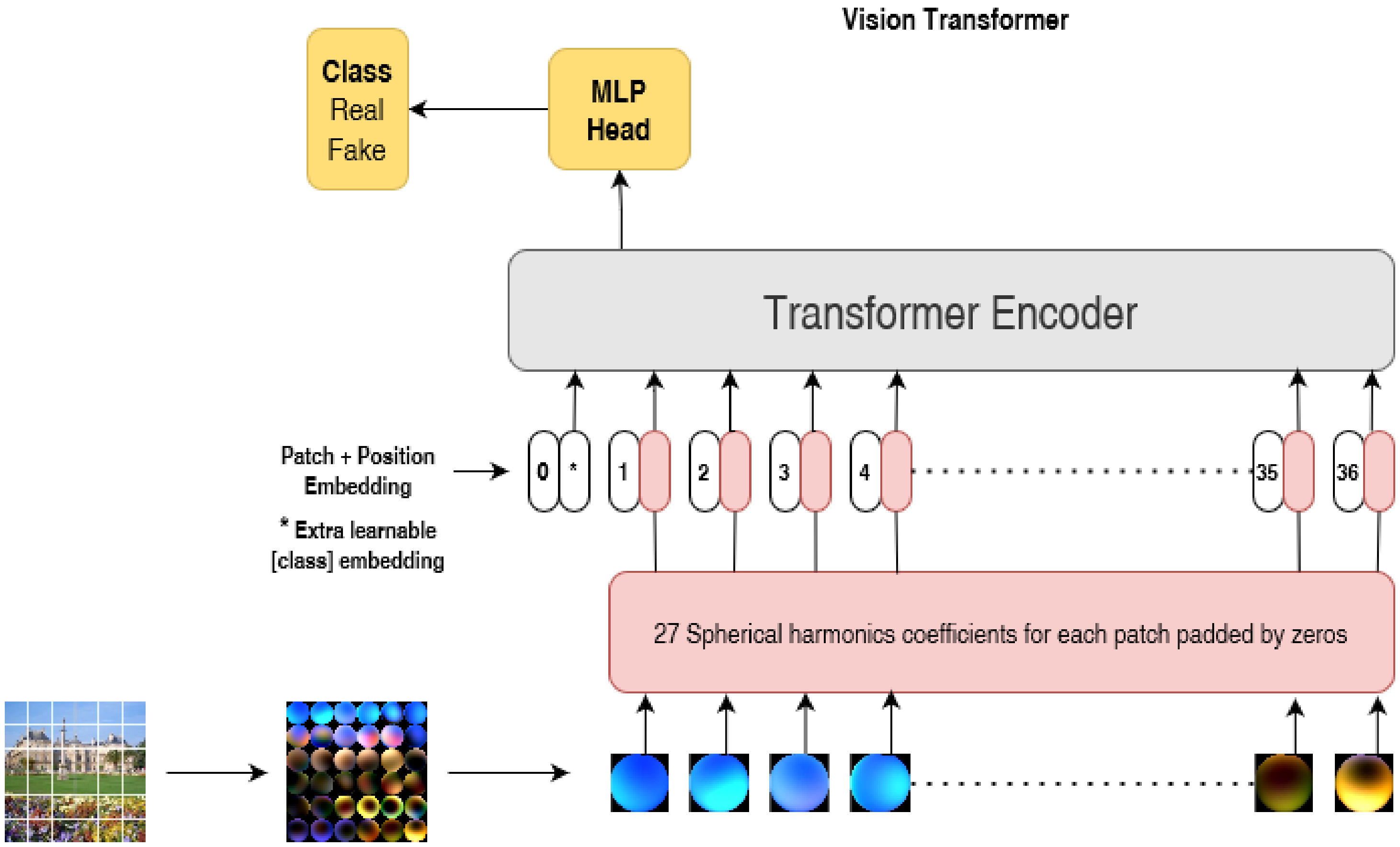
4-layer MLP



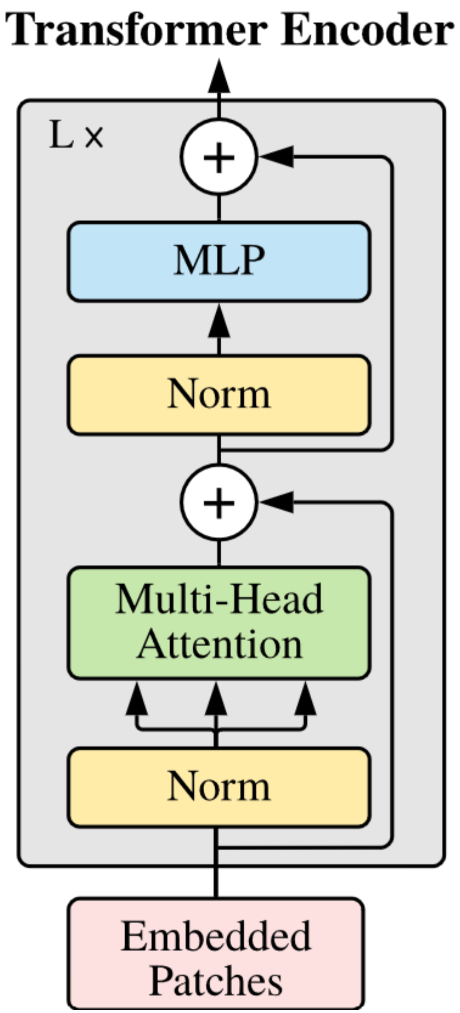
MLP architecture

Binary Classification with ViT

Modified ViT: Instead of image patched, pass spherical harmonics for each patch



SHL(ViT)



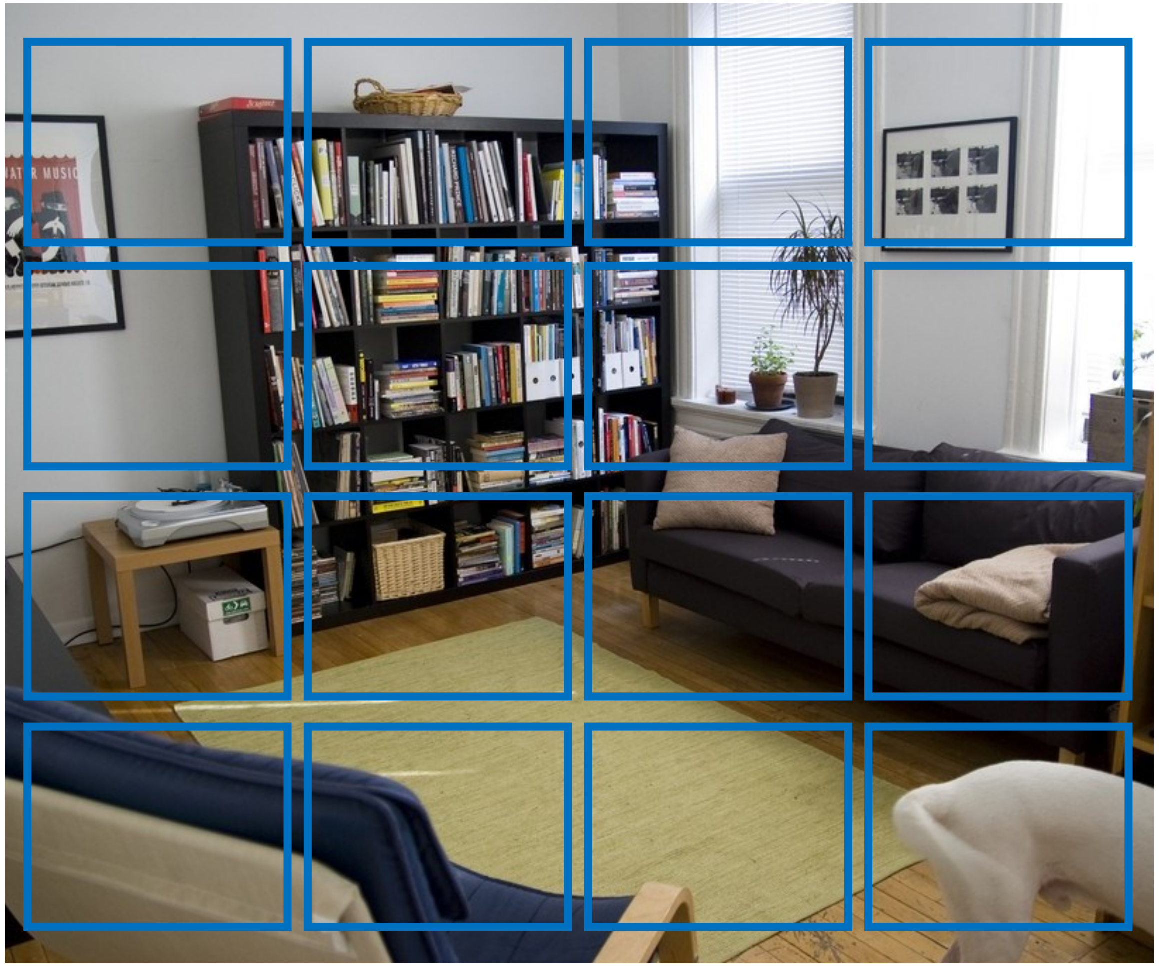
Baseline Methods

GIST Features as a Baseline

Learn scene-level representation instead of segmentation or object-level representations

Estimate structure of a scene by computing spectral signature using a few perceptual scene properties[1]

- Scene properties include openness, roughness, expansion, ruggedness
- Spectral templates for different values of scene properties that are learned from real images
- 32 spectral templates (8 for each scene property)
- Divide image into 4x4 patches
- Average filter response for each filter and patch
- Total 512 features for a single image



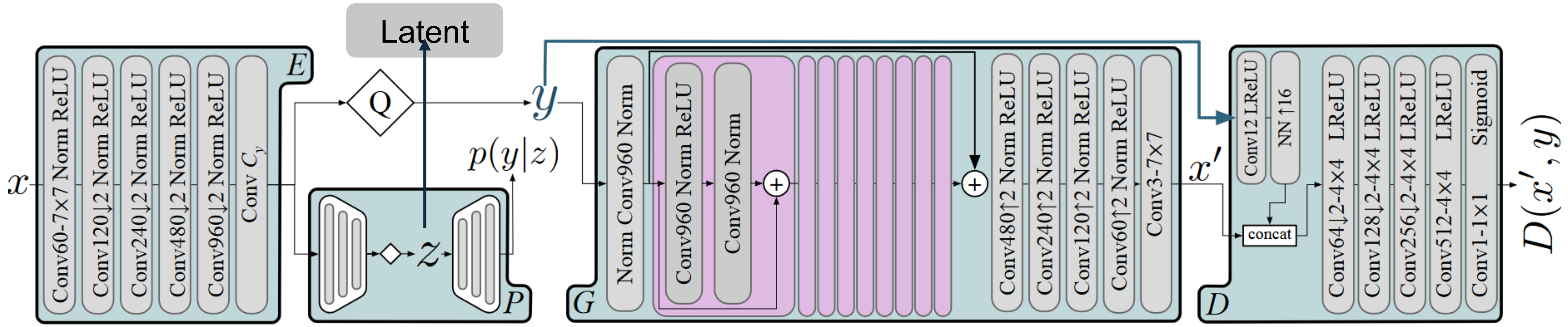
 1. Oliva et al. "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope"

Latent Space of a Compression Network

Real and generated images have different latent space representation

Compression network project input image into a very compact latent representation. HiFiC is one such compression network

- Reconstructs perceptually similar images at extremely low bit-rate
- An input image is projected into a latent representation using an encoder
- A conditional-GAN reconstruct the image from the latent representation
- We only use latent representation



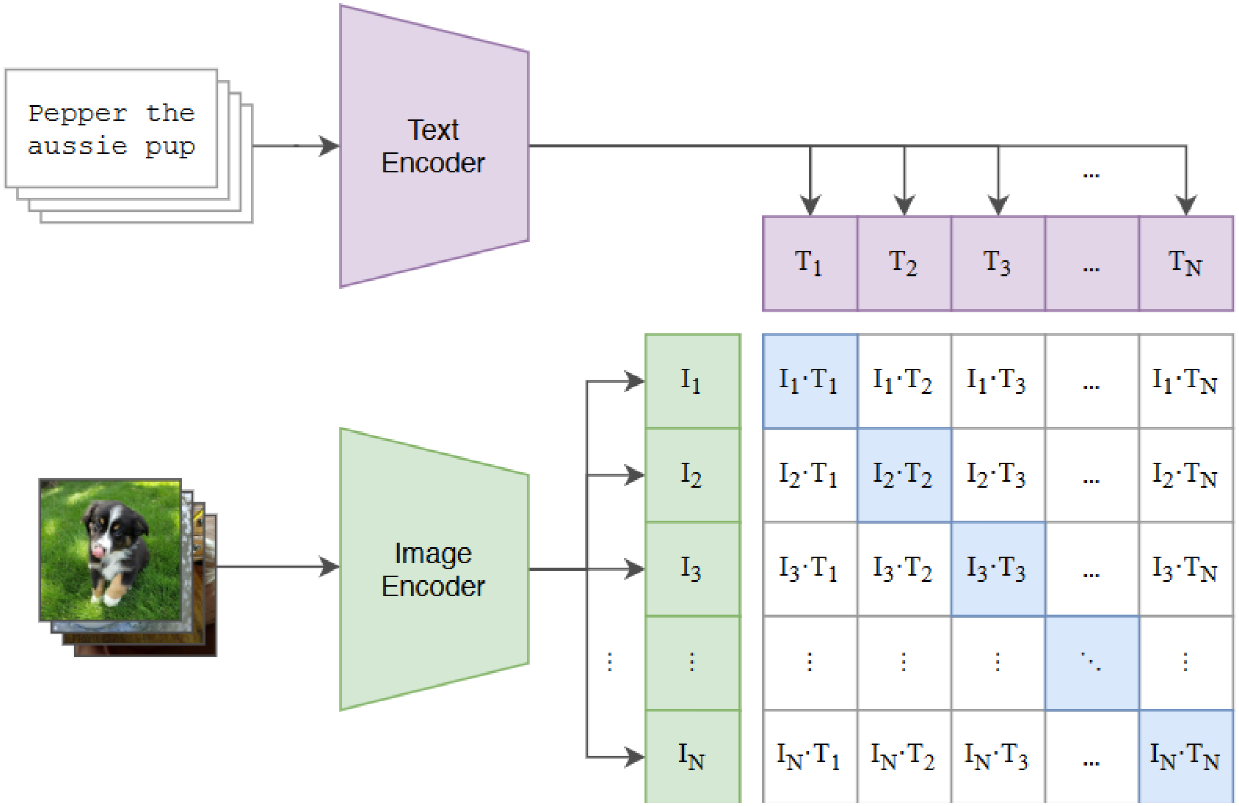
HiFiC architecture[1]

Contrastive Language-Image Pretraining (CLIP)

CLIP features with linear classifiers have better cross-generalization performance

CLIP is trained on large amount of real image-caption pairs collected from the internet.

- Separate text and image encoder
- Minimize cosine similarity between two embedding
- Contrastive loss function
- CLIP image features have good cross-generalization performance[1].



CLIP architecture

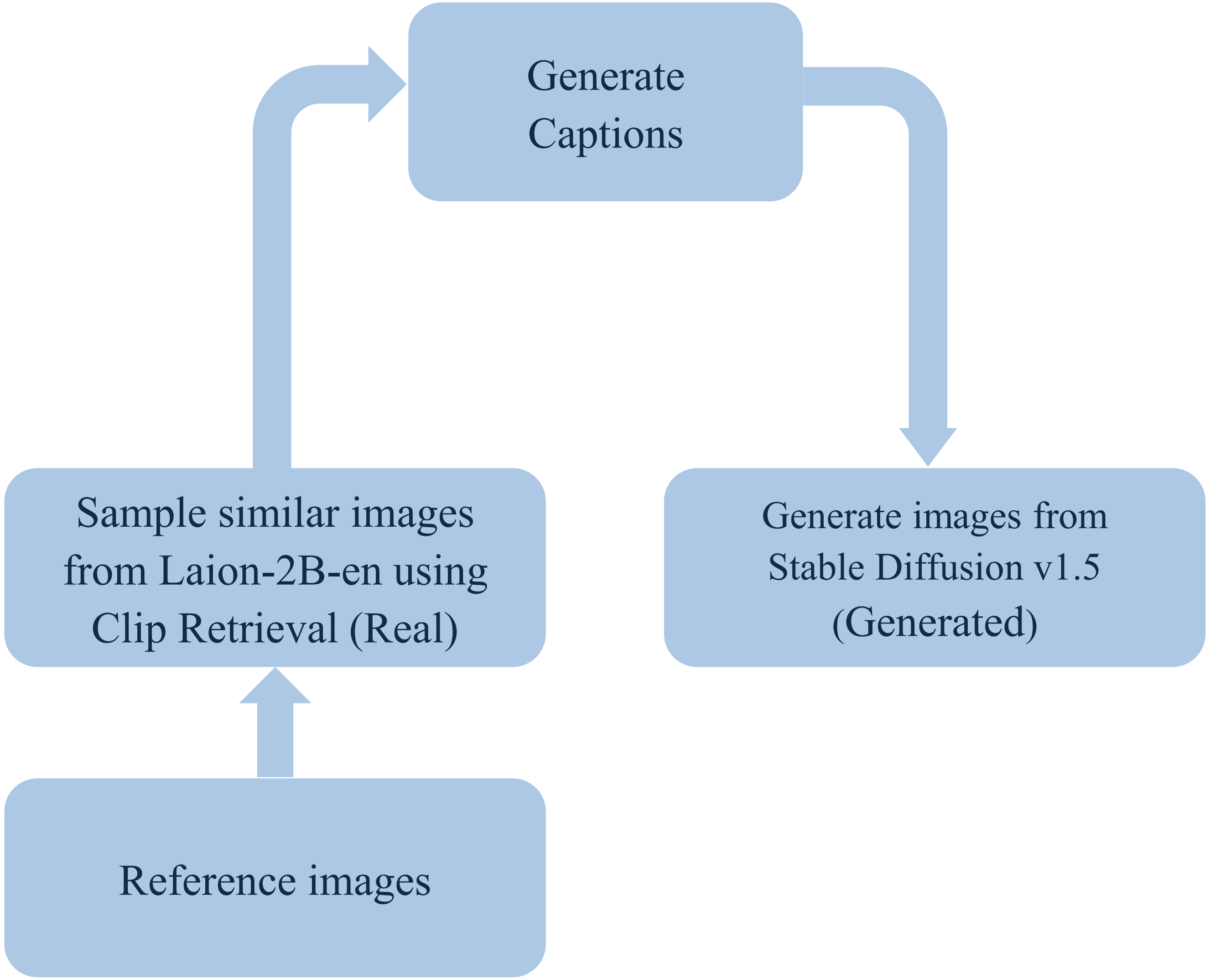
Dataset Creation

Dataset Creation

Real and stable diffusion generated images are conditioned on common captions

Dataset creation pipeline

- Select few reference images manually
- Search similar images based on clip embedding of reference images from LAION-2B-en dataset (rom1504.github.io/clip-retrieval/)
- Collected images are assumed ***real***. Generate captions using ***microsoft/git-large-r-coco*** image captioning model
- Use captions to generate images from SD1.5 model. Append special keywords for photorealistic generation
- All images are JPEG compressed
- Remove white background images



A Few Reference Images

Reference images are used to collect real images from the LAION-2B-en dataset

All images are real.

Real and Generated Samples

Real and stable diffusion generated images are conditioned on common captions

A white chair in a white room

An image sampled from LAION-2B-en dataset (assumed real scene)

A stable diffusion v1.5 generated image

Real and Generated Samples

Real and stable diffusion generated images are conditioned on common captions

A bunch of pears sitting on top of a wooden table.

An image sampled from LAION-2B-en dataset (assumed real scene)

A stable diffusion v1.5 generated image

Data Splits For Training and Evaluation

Distribution of real and generated images (Stable Diffusion v1.5) across data splits

Split	Real Images	Generated Images	Total Images
Train	5123*	5501	10624
Validation	665	679	1344
Test	651	719	1370

* Some of the real images were removed from the dataset as they contained objects with backgrounds removed.

Cross-generator Evaluation Dataset

Distribution of real and generated images in cross-generator dataset

Dataset	Images
RAISE-1k	1000

Real Images

Dataset	Images	Architecture
DALL-E2*	1000	open
DALL-E3*	1000	closed
Firefly*	1000	closed
Midjourney* v5	1000	closed
SD v1.3*	1000	open
SD v1.4*	1000	open
SD v1.5	1000	open
SD v2*	1000	open
SD XL*	1000	open

AI-Generated Images

Open: Architectural detail shared

Closed: Architectural detail are not fully disclosed

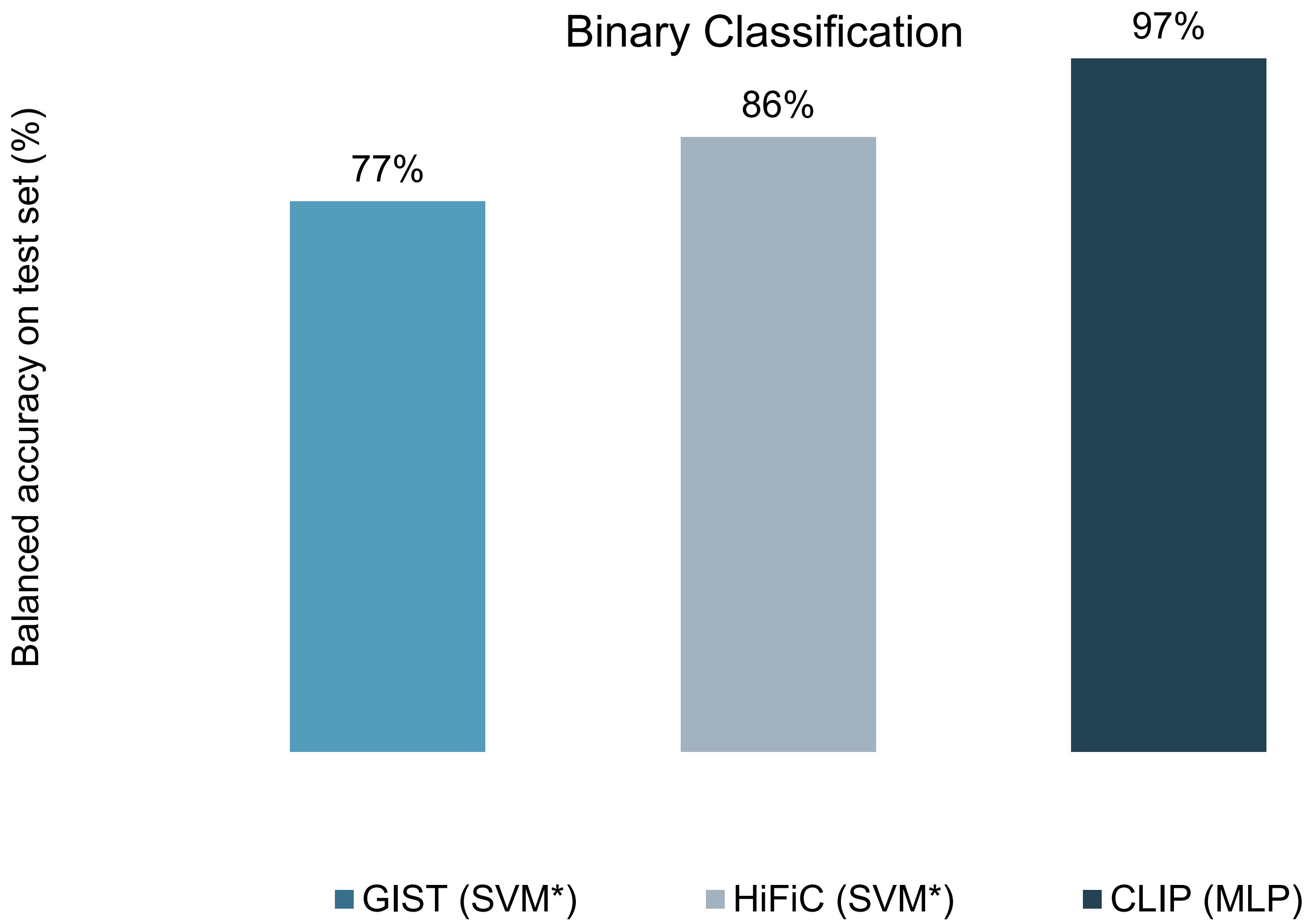
* Images taken from the Synthbuster[1] dataset

Technische Fakultät 1. Q Bammey "Synthbuster: Towards Detection of Diffusion Model Generated Images"

Experiments and Results

Baseline Binary Classification Results

Balanced classification accuracy for different baseline methods on test set



GIST Confusion Matrix

	Pred Real	Pred Gen
Real	496	151
Gen	167	546

HiFiC

517	130
65	648

CLIP

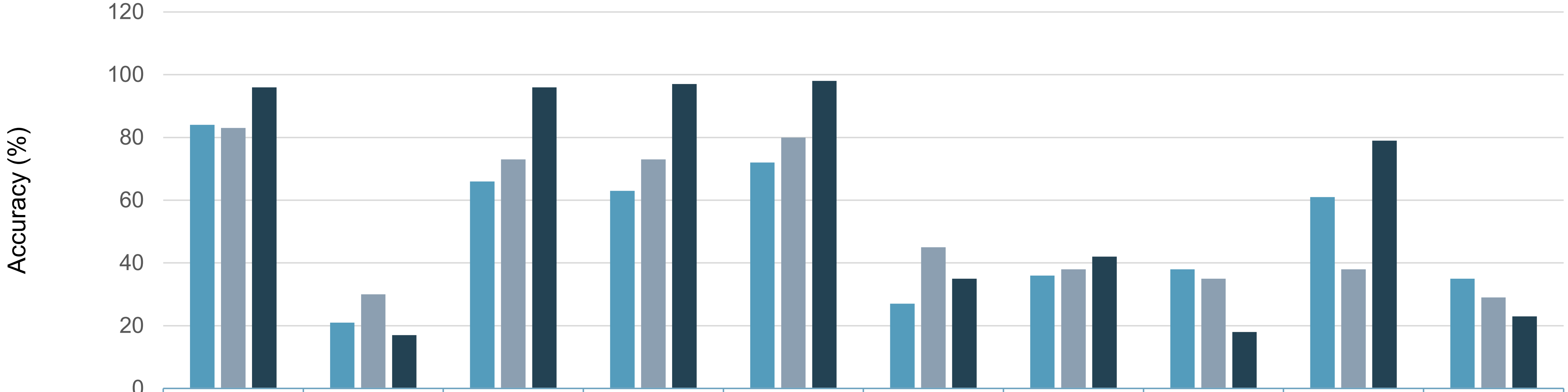
617	21
16	690

* SVM with RBF Kernel

Cross-generator Generalization

Evaluation on cross-generator dataset

Cross-generator performance of baseline methods

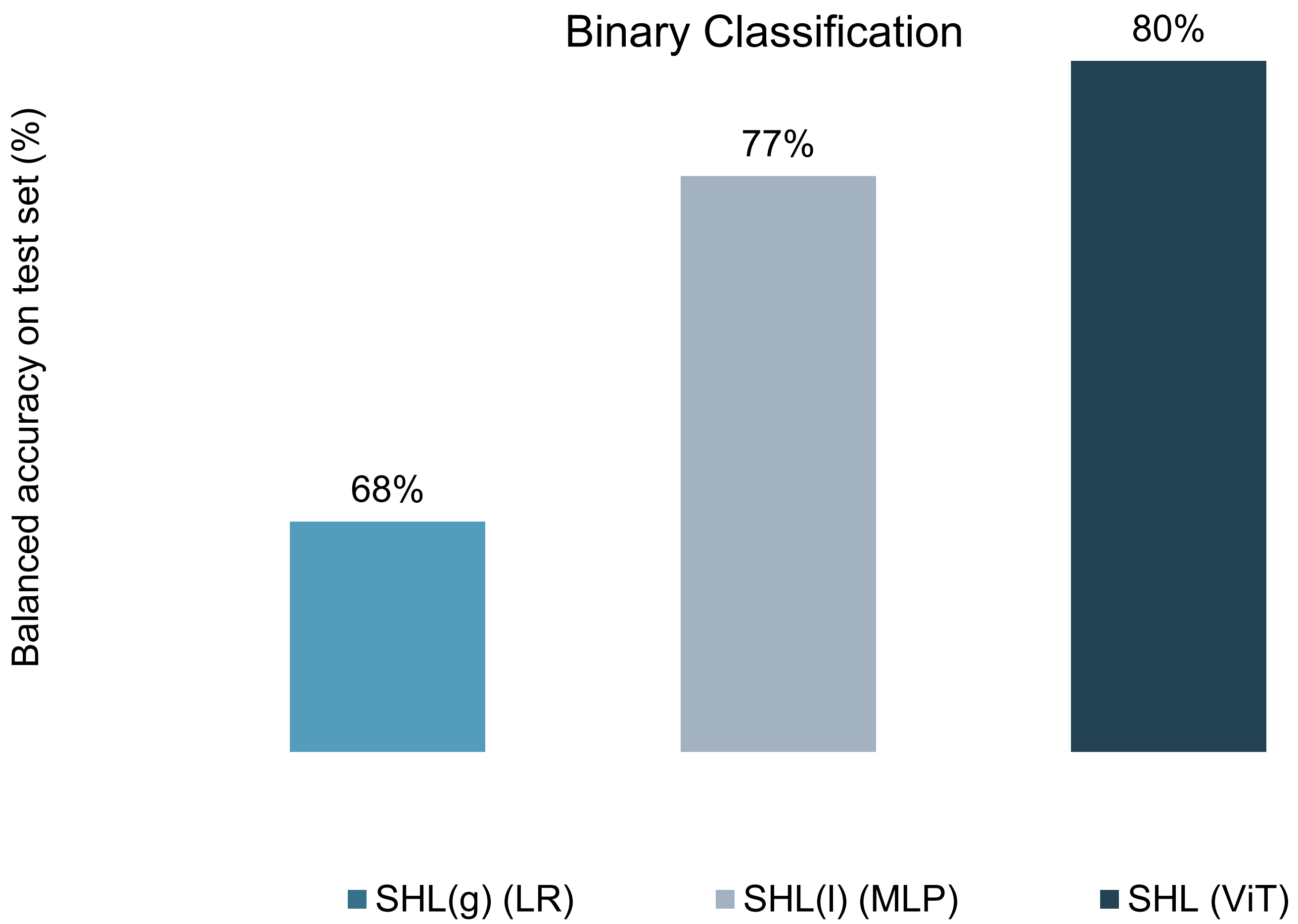


	RAISE-1k	Firefly	SD v1.3	SD v1.4	SD v1.5	SD v2.0	SD XL	DALL-E2	DALL-E3	Midjourney-v5
GIST (SVM)	84	21	66	63	72	27	36	38	61	35
HiFiC (SVM)	83	30	73	73	80	45	38	35	38	29
CLIP (MLP)	96	17	96	97	98	35	42	18	79	23

GIST (SVM) HiFiC (SVM) CLIP (MLP)

Binary Classification Results for Proposed Method

Balanced classification accuracy for proposed methods on test set



SHL(g) Confusion Matrix

	Pred Real	Pred Gen
Real	401	246
Gen	189	524

SHL(l)

378	260
52	654

SHL (ViT)

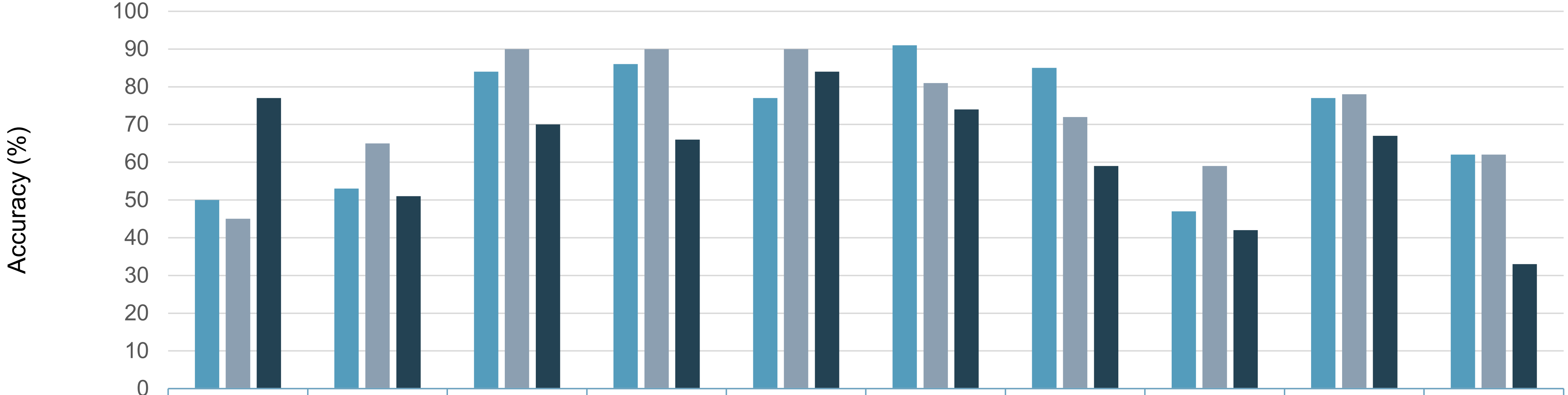
469	181
68	642

* SVM with RBF Kernel

Cross-generator Generalization

Evaluation on cross-generator dataset

Cross-generator performance of our proposed methods

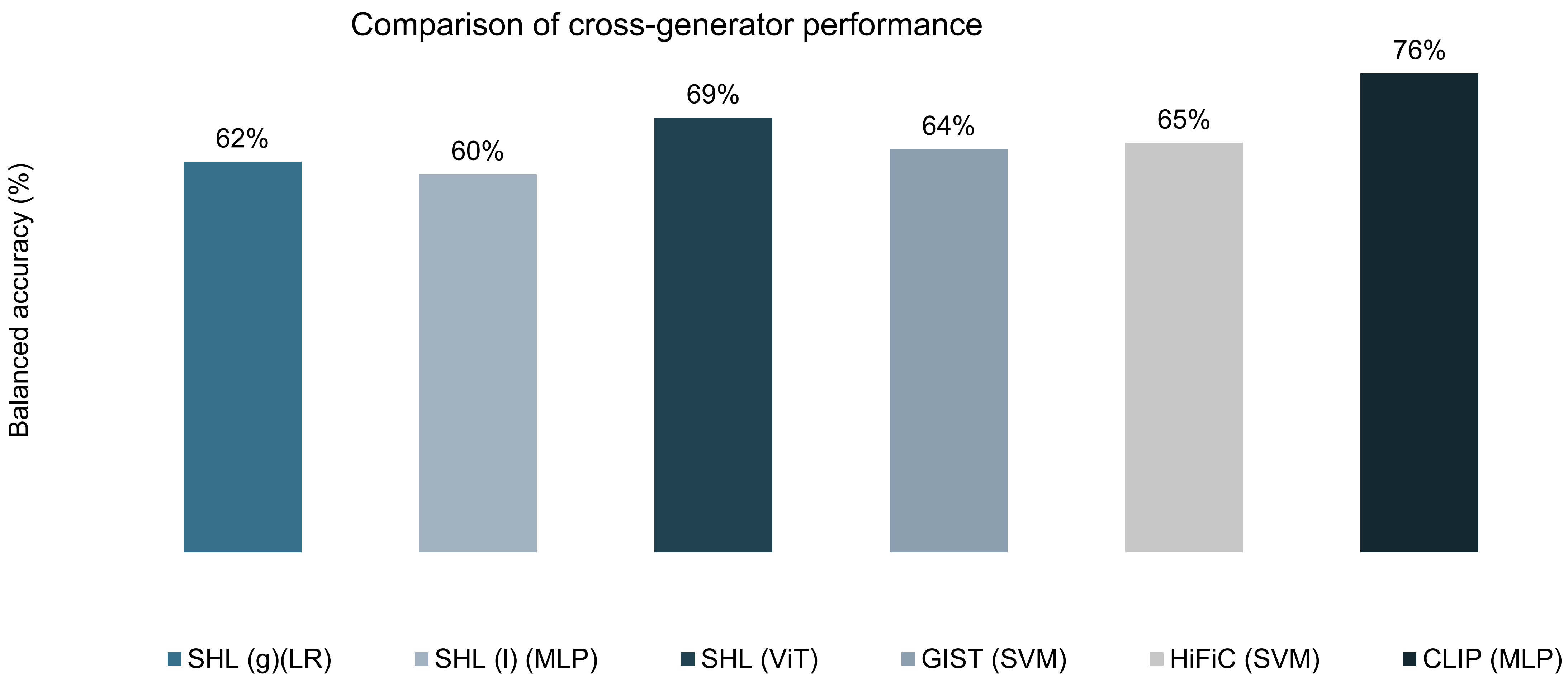


	RAISE-1k	Firefly	SD v1.3	SD v1.4	SD v1.5	SD v2.0	SD XL	DALL-E2	DALL-E3	Midjourney-v5
SHL(g) (LR)	50	53	84	86	77	91	85	47	77	62
SHL(l) (MLP)	45	65	90	90	90	81	72	59	78	62
SHL (ViT)	77	51	70	66	84	74	59	42	67	33

SHL(g) (LR) SHL(l) (MLP) SHL (ViT)

Comparison Cross-generator Performance

Comparison between baseline methods and proposed method for cross-generator performance



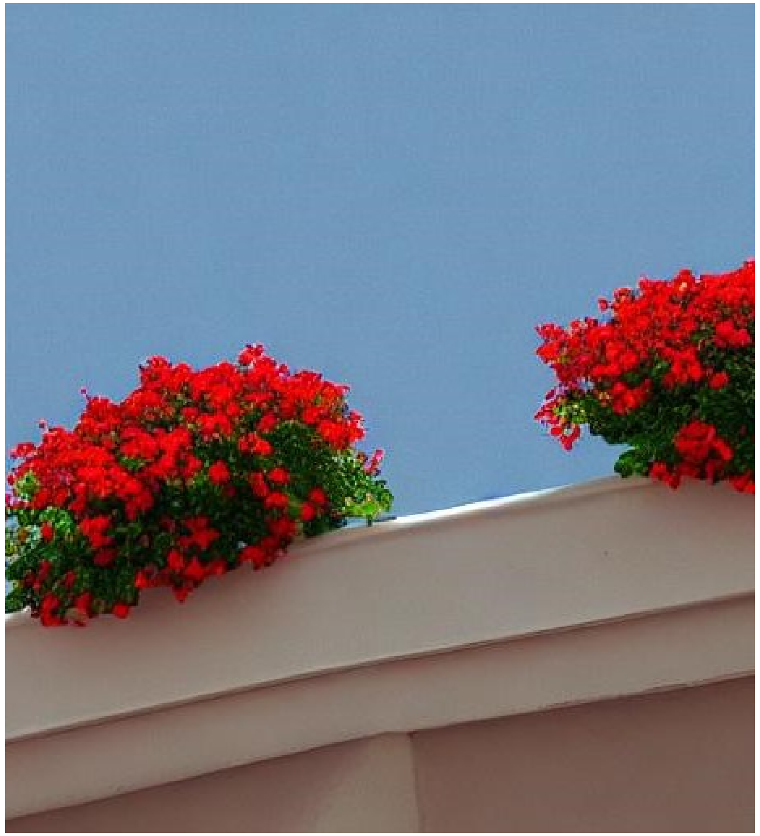
Qualitative Analysis of SHL (ViT)

Some Misclassified Examples

A few generated images taken from a smaller subset (50 images, manually picked) of test set generated images

Generated but predicted real

Generated but predicted real



Generated but predicted real

Some Misclassified Examples

A few real images taken from a smaller subset (50 images, manually picked) of test set real images

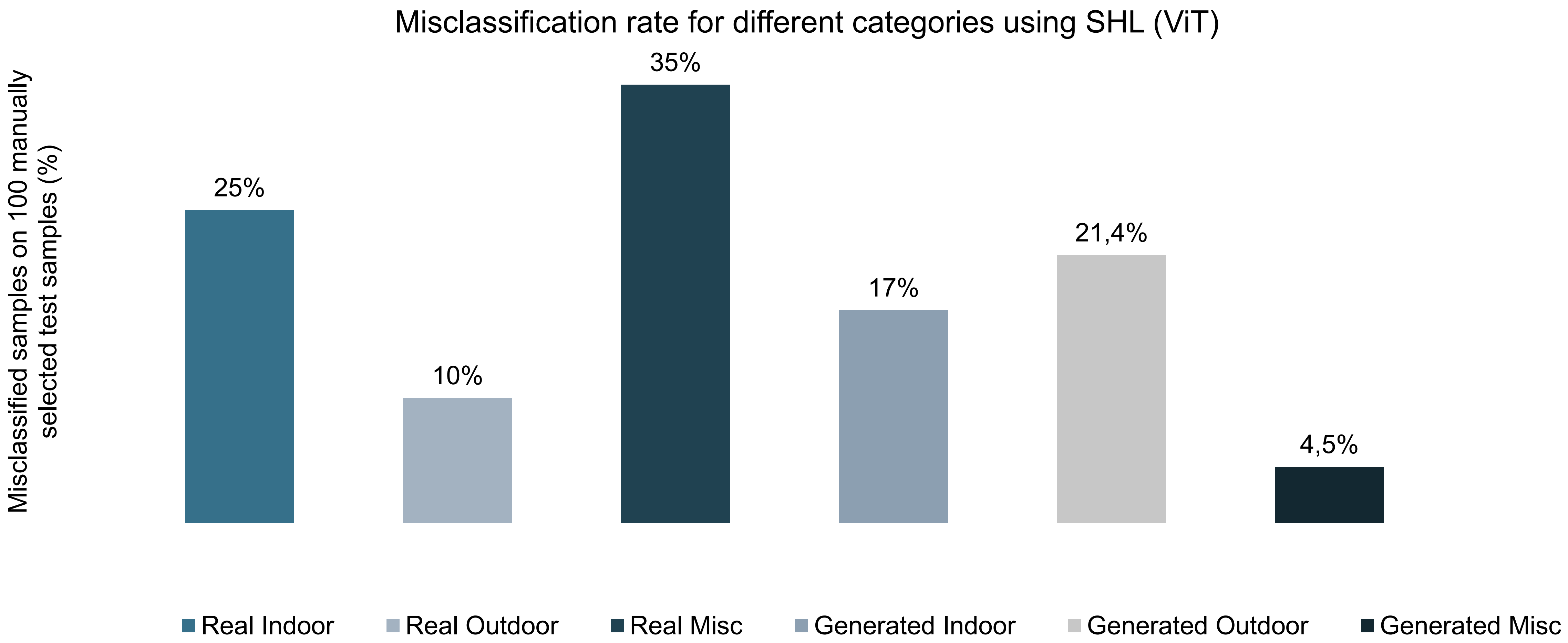
Real but predicted generated

Real but predicted generated

Real but predicted generated

Categorization of Misclassified Samples

Evaluation on manually picked 100 images from test set (Nearly 17 images for each category)



Conclusion and Future Directions

Conclusion and Future Outlook

Conclusion

- Spherical harmonics representation of scene illumination can be used to detect real and generated images
- ViT architecture can capture illumination inconsistencies
- Our proposed method has ability to generalize on unseen image generators.

Future outlook:

- Increase the training set
- Include more diverse generators in cross-generator evaluation dataset
- Experiment with recently proposed better inverse rendering pipelines

Thank you for your attention!