MammoGAN: High-Resolution Synthesis of Realistic Mammograms

Dimitrios Korkinof[1], Andreas Heindl[1], Tobias Rijken[1], Hugh Harvey[1], Ben Glocker[1,2]

[1] Kheiron Medical Technologies Ltd., London, UK
[2] Department of Computing, Imperial College London, London, UK

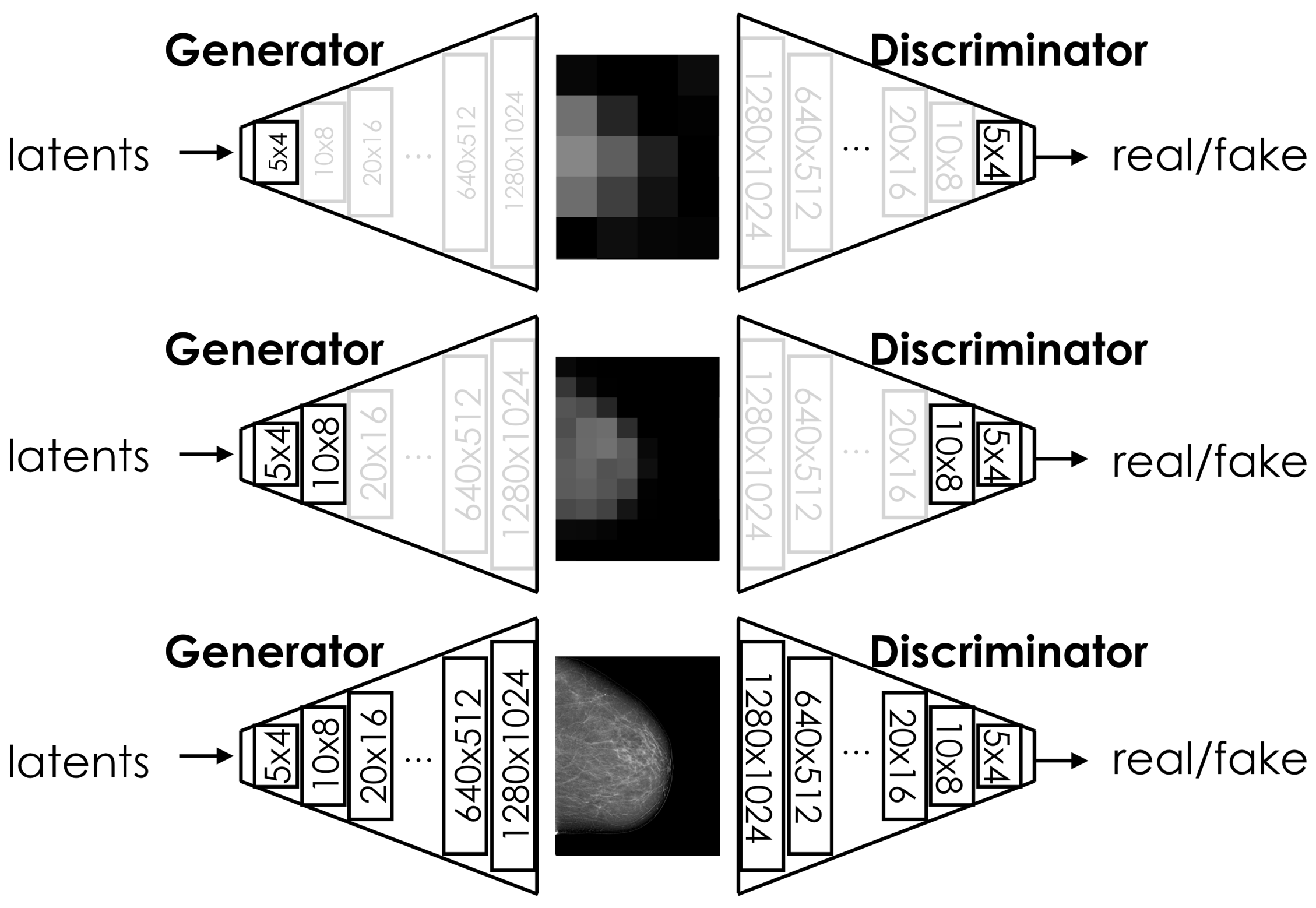

Figure 1: Illustration of progressive GAN training.

Introduction

We explore whether recent advances in generative adversarial networks (GANs) enable synthesis of realistic medical images that are hard to distinguish from real ones, even by domain experts. High-quality synthetic images can be useful for data augmentation, domain transfer, and out-of-distribution detection.

GAN Resolutions	
WGAN-GP	128x128
StarGAN	128x128
Glow	256x256
CycleGAN	512x512
BigGAN	512x512
PGGAN	1024x1024
Ours	1280x1024

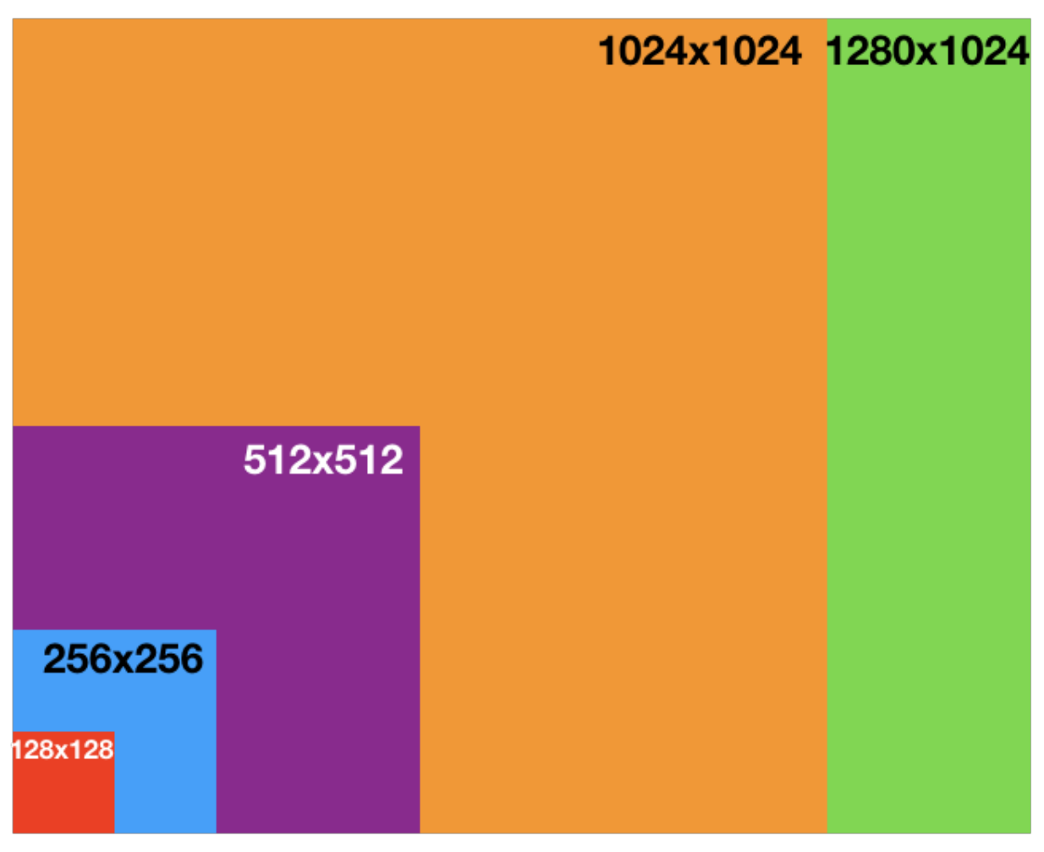

Figure 4: Illustration of various GAN resolutions.

However, generating realistic images is challenging, particularly for Full Field Digital Mammograms (FFDM), due to the high-resolution, textural heterogeneity, fine structural details and specific tissue properties.

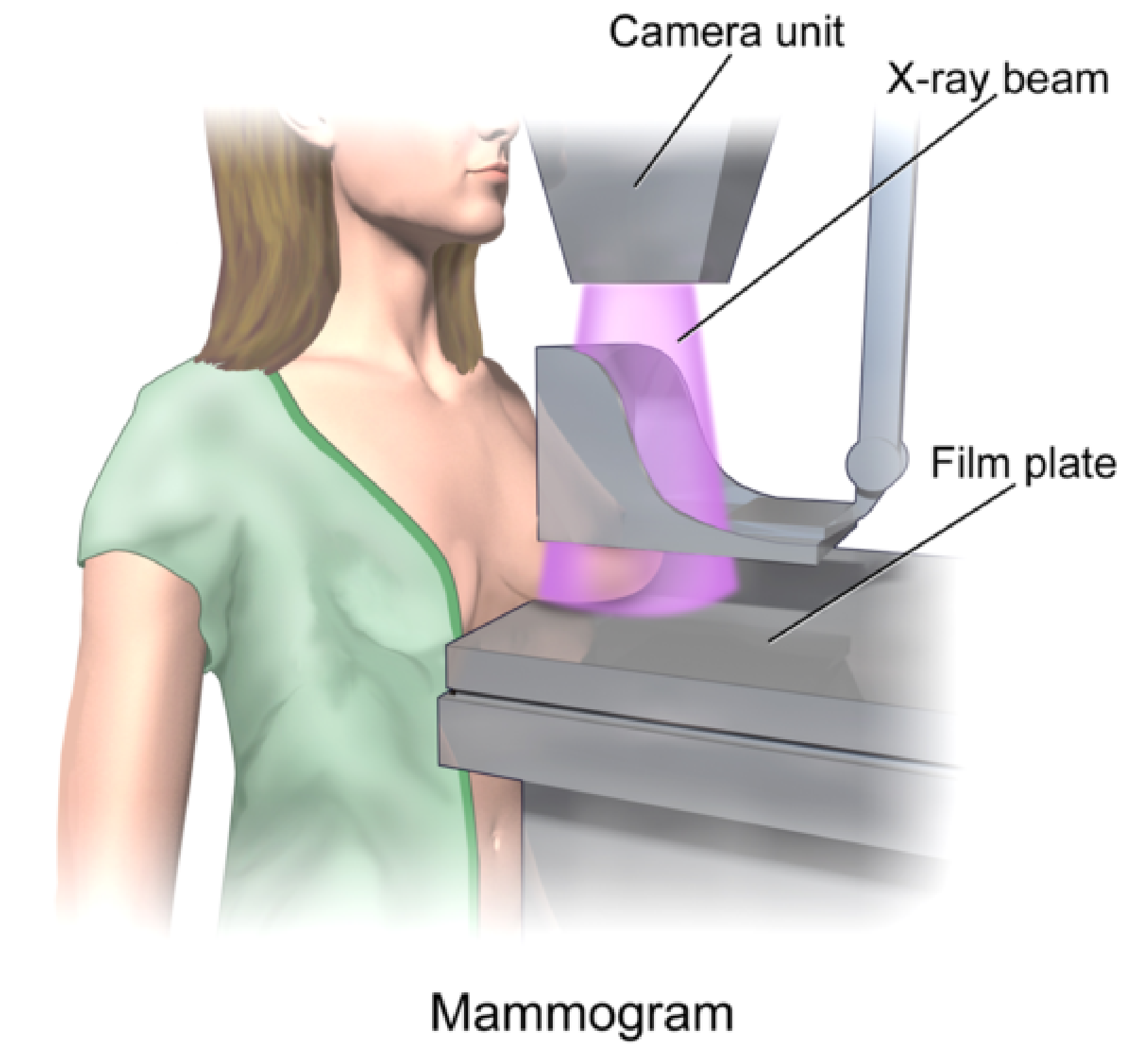

Figure 5: Acquisition of CC image. source: [Blausen, 2014].

Dataset

Mammograms are soft tissue breast X-rays acquired in two standard views, CC and MLO (Fig. 6).

We used a dataset of 1 million mammograms in total.

We excluded images with large foreign bodies (pacemakers, implants, etc.) and post-operative cases (metal clips, etc.).

All images were resized to 1280x1024 pixels preserving their aspect ratio.

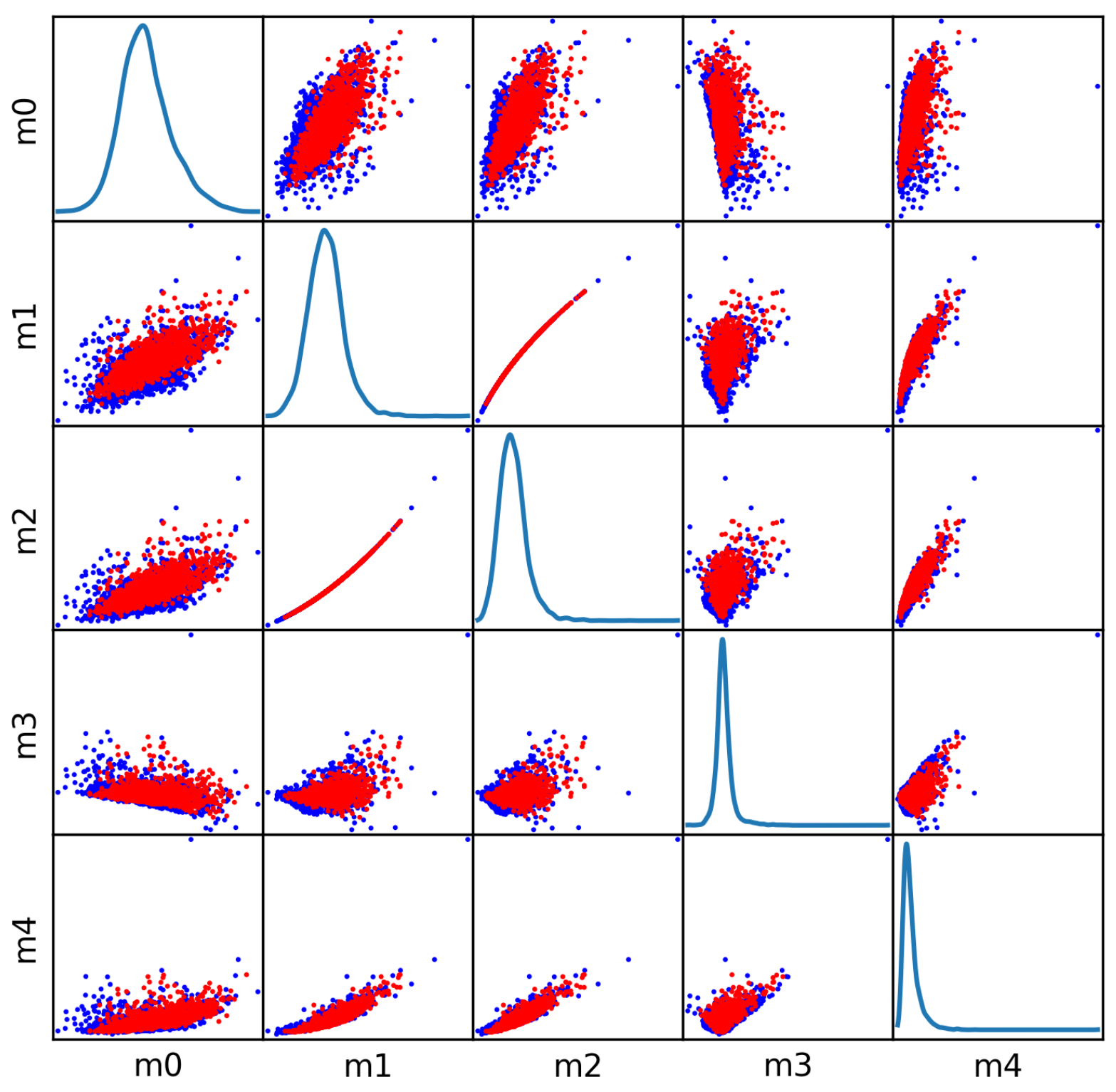

Figure 2: Moments plot (red denoting real and blue synthetic images)

Method

A key development for scaling GANs to higher resolutions is progressive training, as originally proposed in [Karras et al., 2018]. The main concept is to start from a very low resolution, before gradually increasing it as more layers are phased in (Fig.1). We used a Wasserstein objective and gradient penalty.

Despite using progressive training, we still experienced stability issues, which we alleviate by:

- Adding supervised information [Salimans et al., 2016]
- Decreasing the learning rate by 25%
- Increasing the D iterations per each G update
- Doubling the final feature layer and starting depth

Finally, we selected the best network checkpoint based on the sliced Wasserstein distance [Karras et al., 2018].

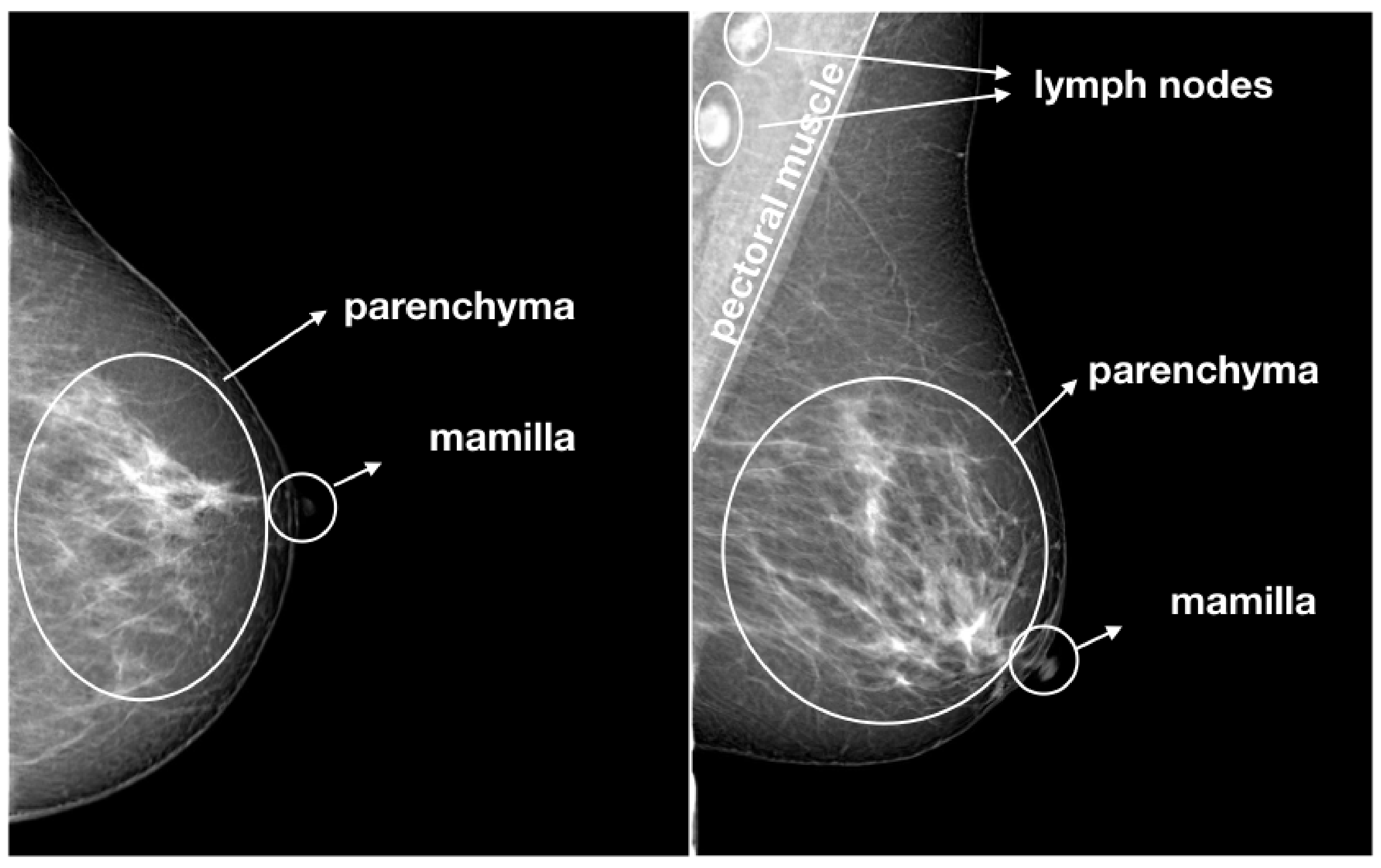

Figure 6: CC and MLO view examples.

Quality Assessment

GAN evaluation metrics (i.e. Frechet inception, sliced Wasserstein etc.) are mostly useful in comparing synthesis methods rather than directly evaluating them.

To assess the quality of the generated images we rely on two methods:

Moments plots: We plot the first five statistical moments to assess similarity between low-level pixel distributions (see Fig. 2).

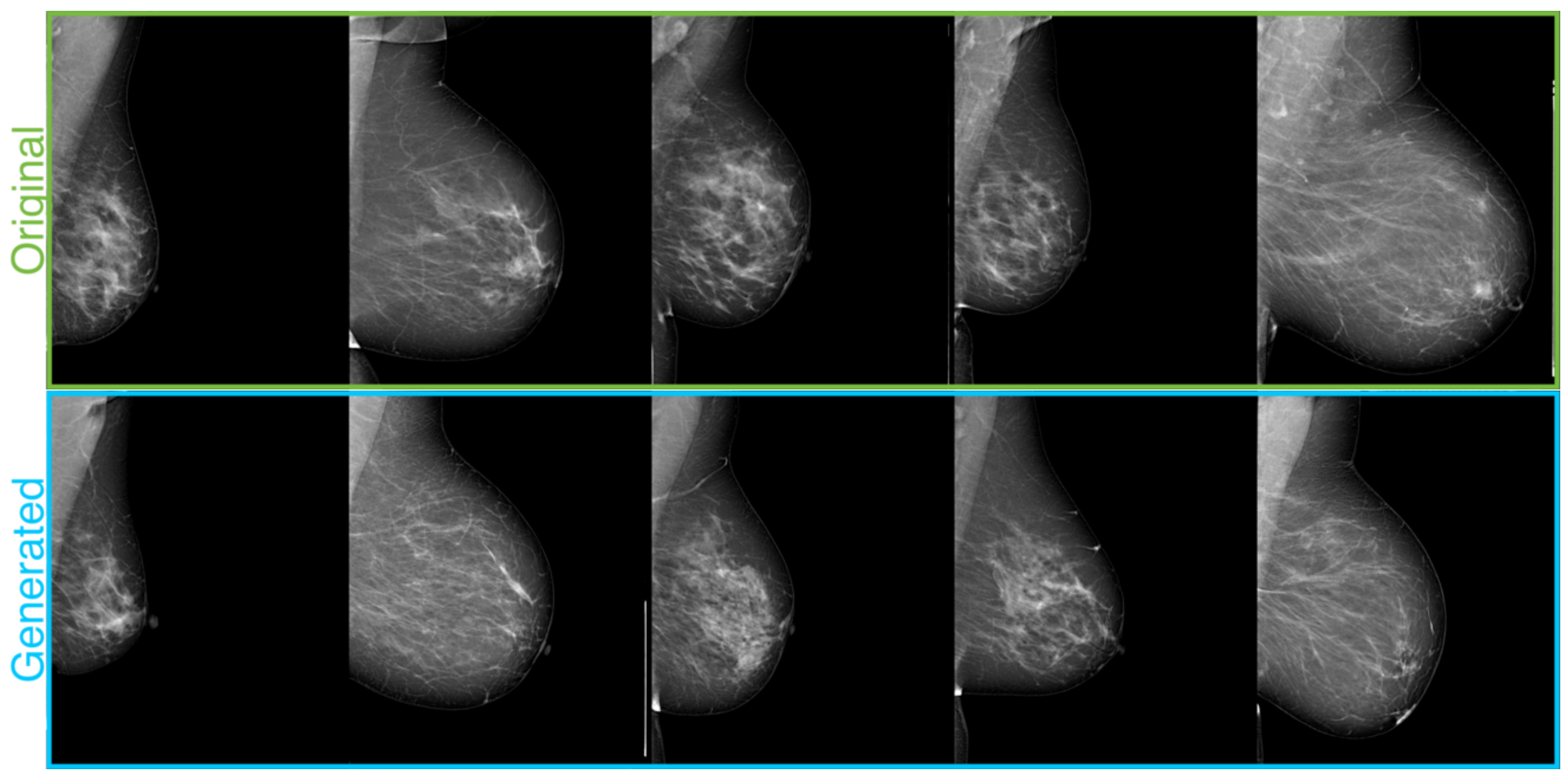

Figure 3: Randomly sampled examples of real and generated MLO views.

User study: We conducted a randomised user study to determine whether synthetic images can be distinguished from real ones as a proxy for perceptual realism.

User Study

We used 1000 randomly sampled synthetic and 1000 real MLO views.

From which, we excluded images with visible artefacts, 13.6% from the synthetic and 2.8% from the real.

We developed a custom tablet app built in Unity with pinch and zoom capability.

Randomly assigned real/synthetic image pairs were presented to participants without time limit.

Results

Qualitative results:

Samples are generally of high quality and variability.

The MLO view is the more difficult of the two.

We observed that 86.4% appeared without artefacts (based on a sample of 1000 MLOs).

Low-level pixel distributions match well in our moments plots (Fig. 2).

Finally, the network resists reproducing calcification and metal markers; non-smooth features that appear as bright spots.

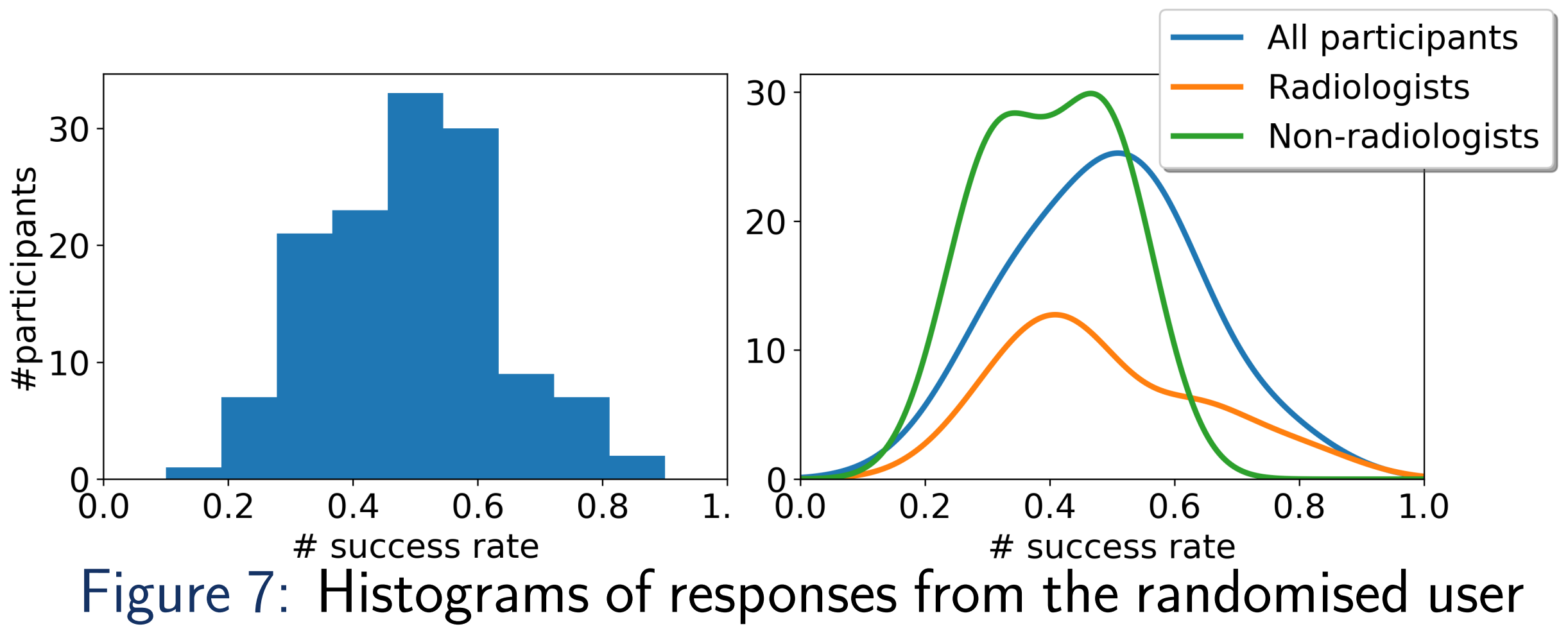

Figure 7: Histograms of responses from the randomised user study - consistent with $\sim Bin\,(n=10, \pi=0.5)$.

User study results:

Our user study was conducted during the RSNA conference with a total of 117 participants.

Participants were asked to assess 10 randomly-paired images with no time limit.

55 participants were radiologists (82% board certified, 60% specialised in breast radiology).

A chi-square test yields $p = 0.999$, a strong indication in favour of the hypothesis that the responses were random $\sim Bin\,(n=10, \pi=0.5)$.