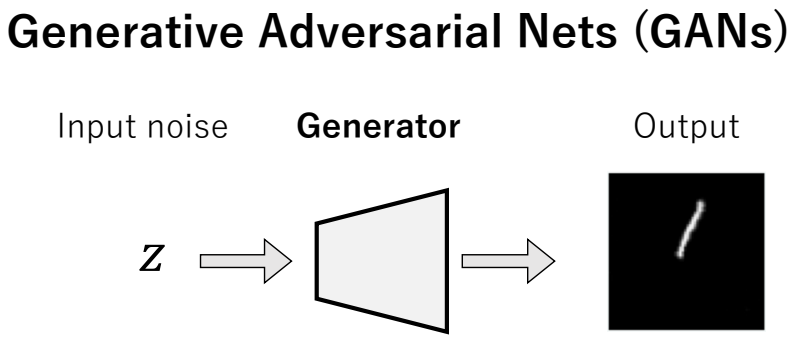
Disentangle, Assemble, and Synthesize: Unsupervised Learning to Disentangle Appearance and Location

Hiroaki Aizawa[1], Hirokatsu Kataoka[2], Yutaka Satoh[2], Kunihito Kato[1]

[1]Gifu University, Japan

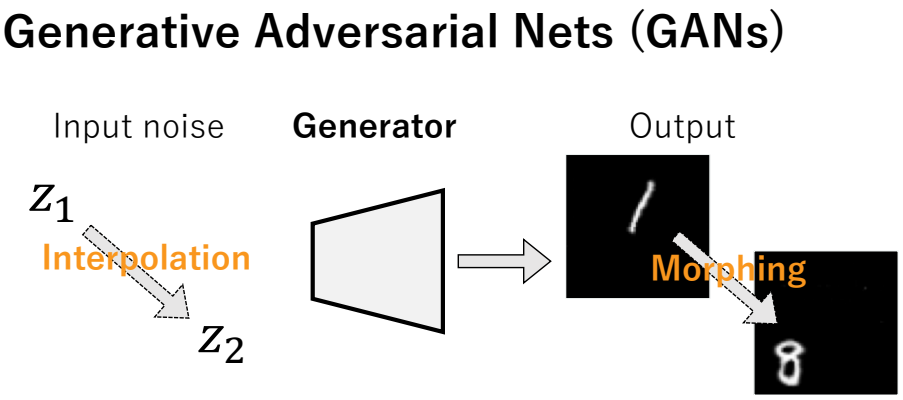
[2]National Institute of Advanced Industrial Science and Technology (AIST), Japan

Background



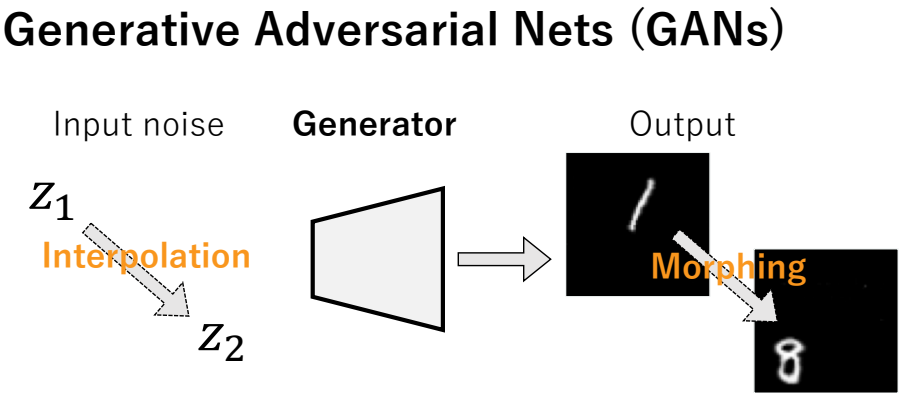
Generative Adversarial Nets (GANs) learn to map points in latent space into image spacce by an adversarial training.

Background



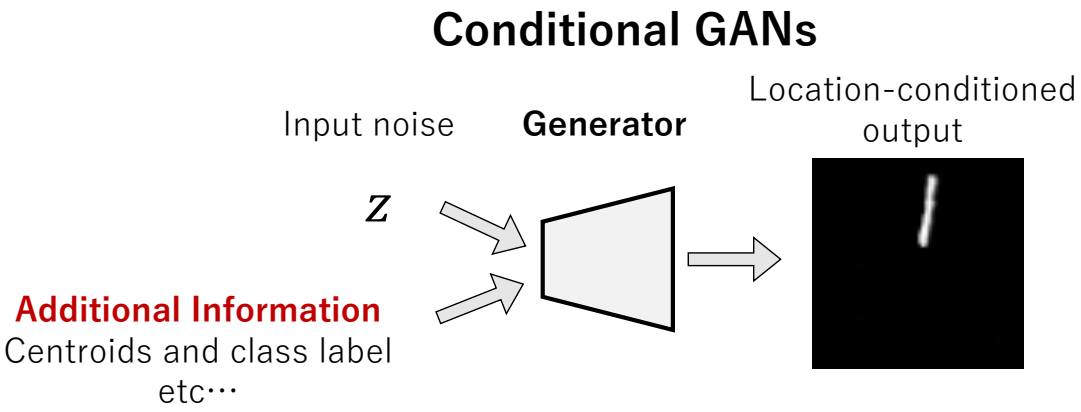
Moreover, the images are morphed from one to another by interpolation between points in latent space.

Background



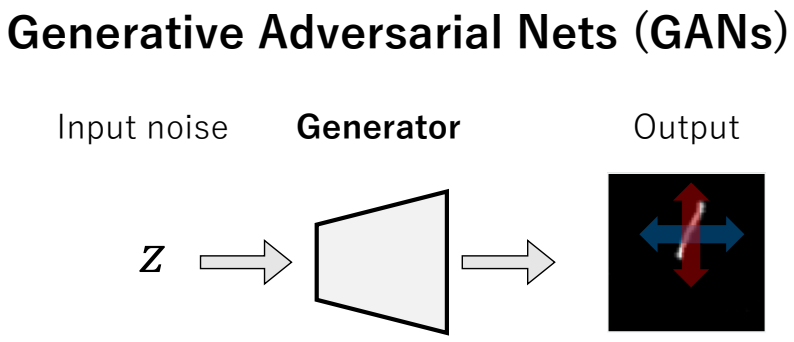
These capabilities provide us many applications for image synthesis and manipulation

Background



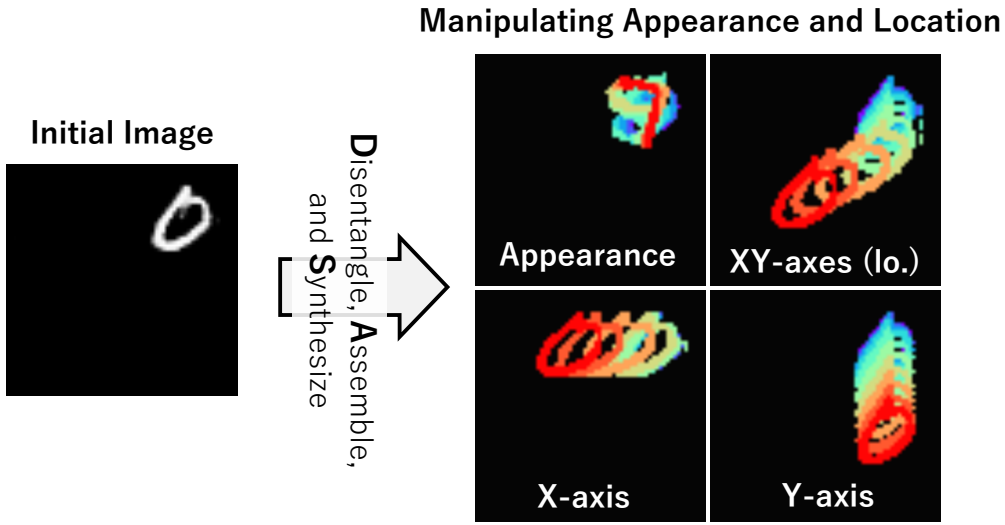
We can design the conditional GANs that gives the centroids and the class labels for controlling the location and the appearance, respectively.

Our Goal



As the next step for the GANs, we tackle the problem of learning representations that allow us to control only a specific factor in the image for unsupervised image manipulation.

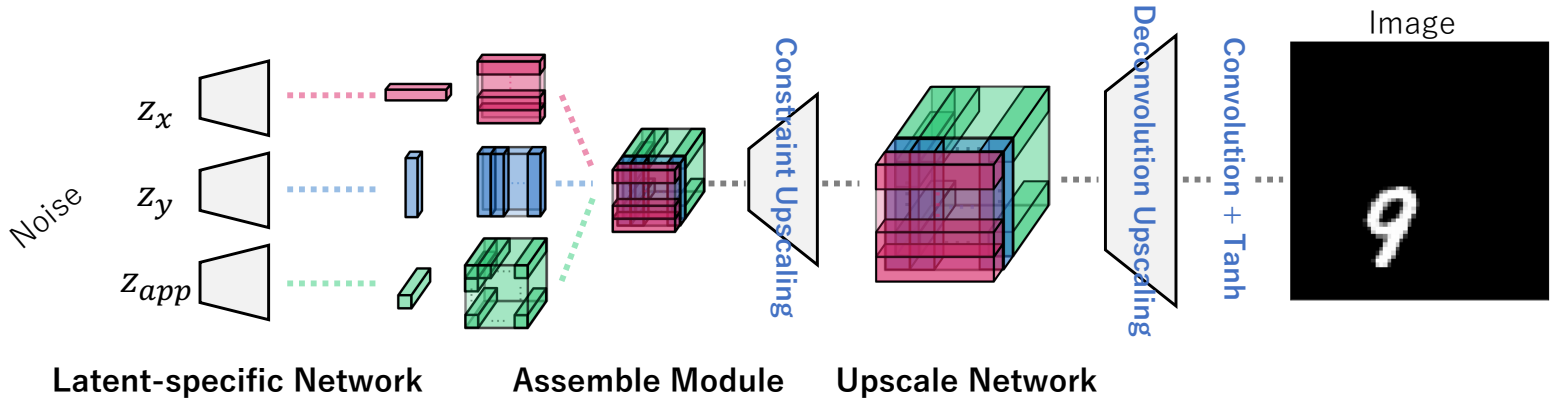
Disentangle-Assemble-Synthesize



To achieve the goal, our **DAS** learns to :

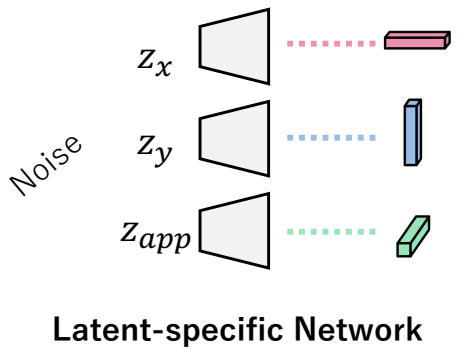
- **D**isentangle appearance, x-axis, and y-axis factors,
- **A**ssemble these representations,
- **S**ynthesize images.

Pipeline



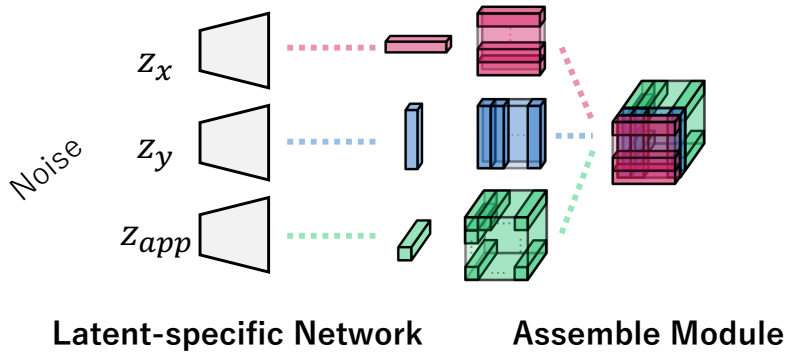
Our DAS consists of a latent-specific network, assemble module, and upscale network.

Pipeline: Latent-specific Networks



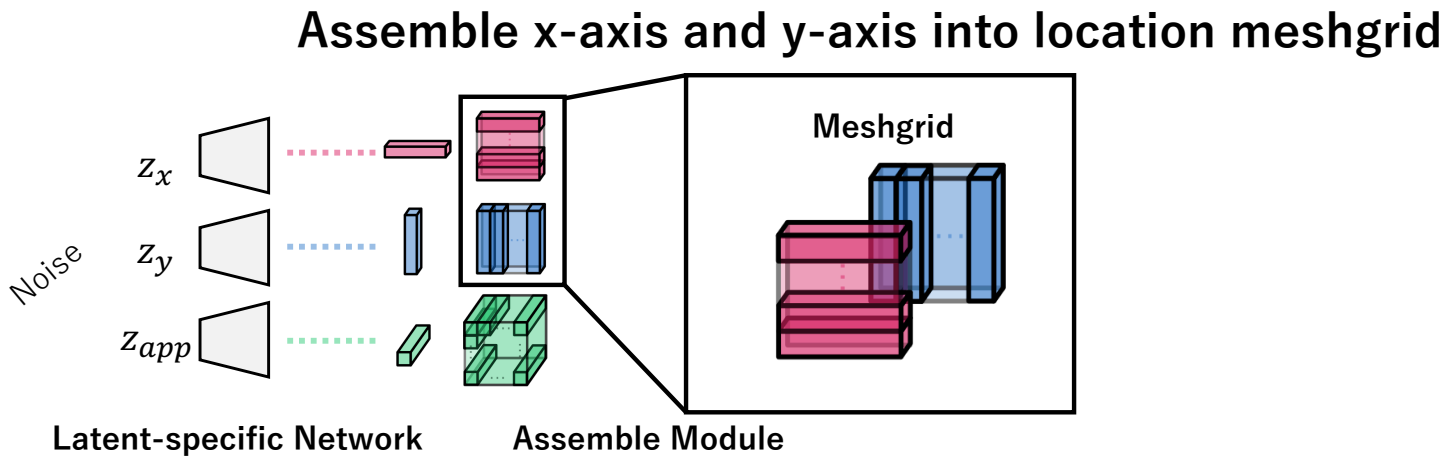
Given the appearance, x-axis, and y-axis noises, each latent-specific network outputs the feature vector of the corresponding factor.

Pipeline: Assemble Module



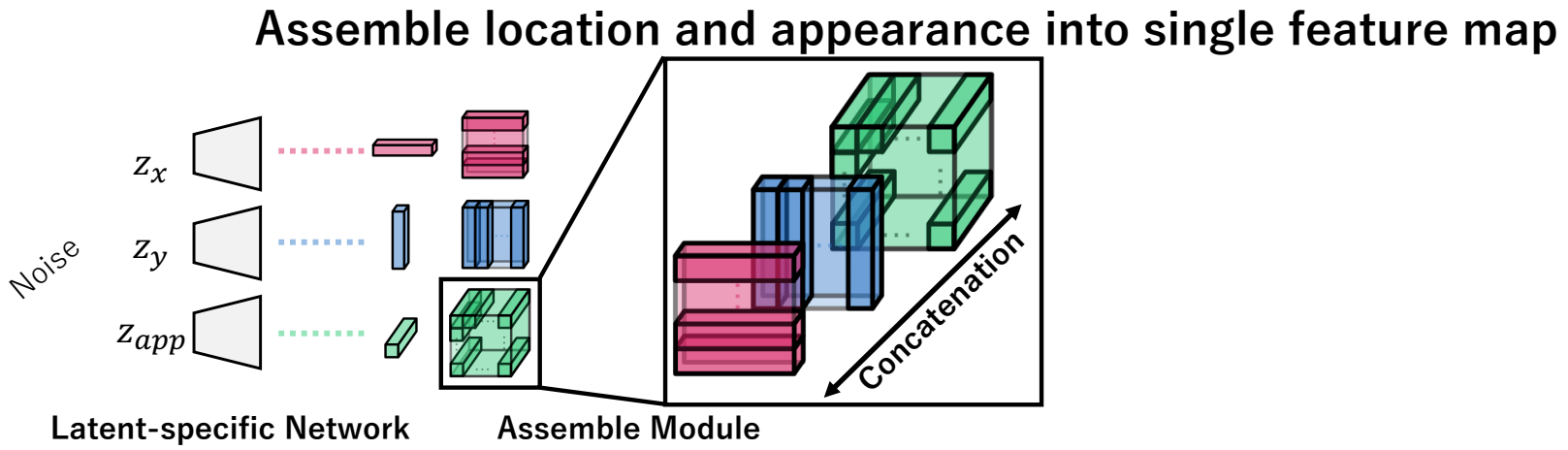
Assemble Module assembles the given set of vector into a structurally constrained feature map for disentanglement.

Pipeline: Location Meshgrid



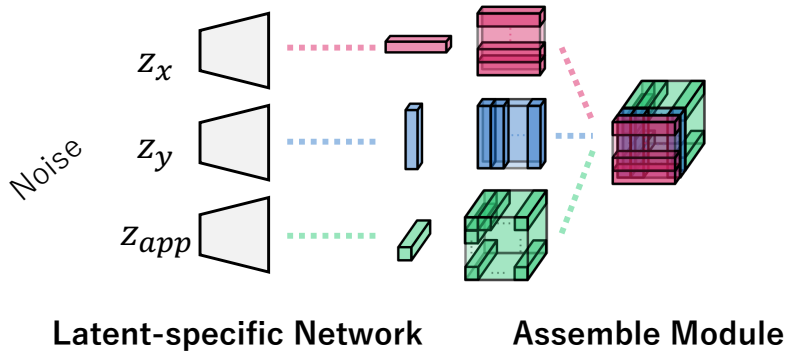
Assemble Module represents the x-axis and y-axis representations as location meshgrid

Pipeline: Structural Constraints



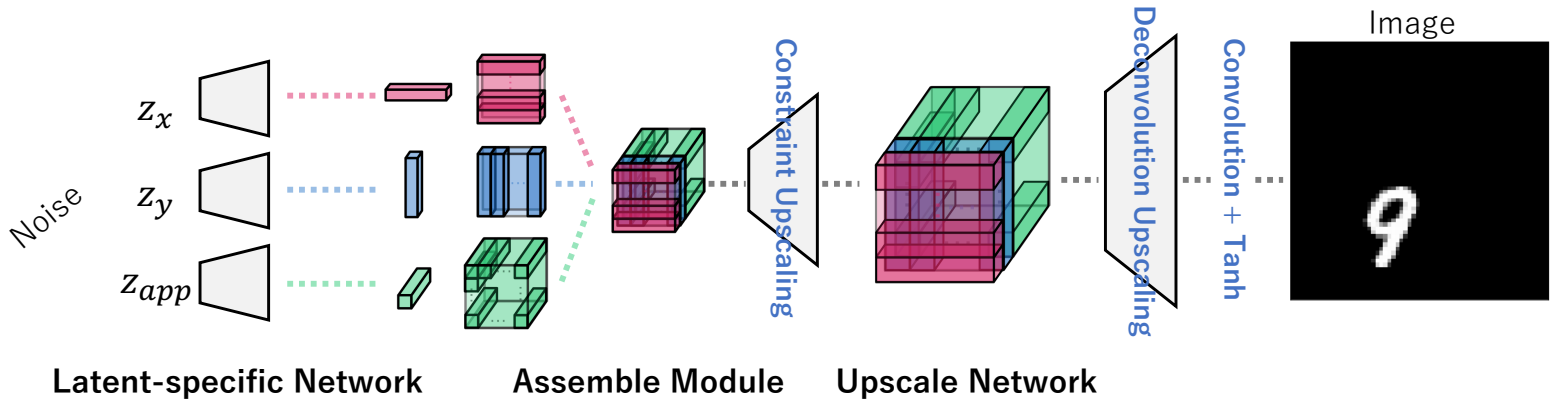
Assemble Module concatenates the location meshgrid with the map tiling the same appearance vector in all positions.

Key idea



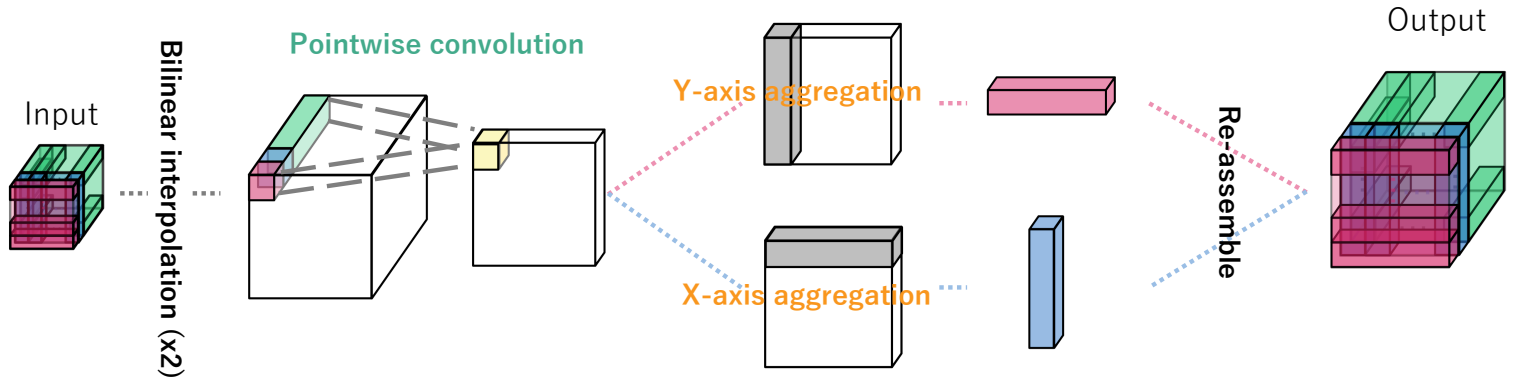
Our idea is to prevent the appearance and the location from interacting with each other by packing them into each position of the single feature map.

Upscale Networks



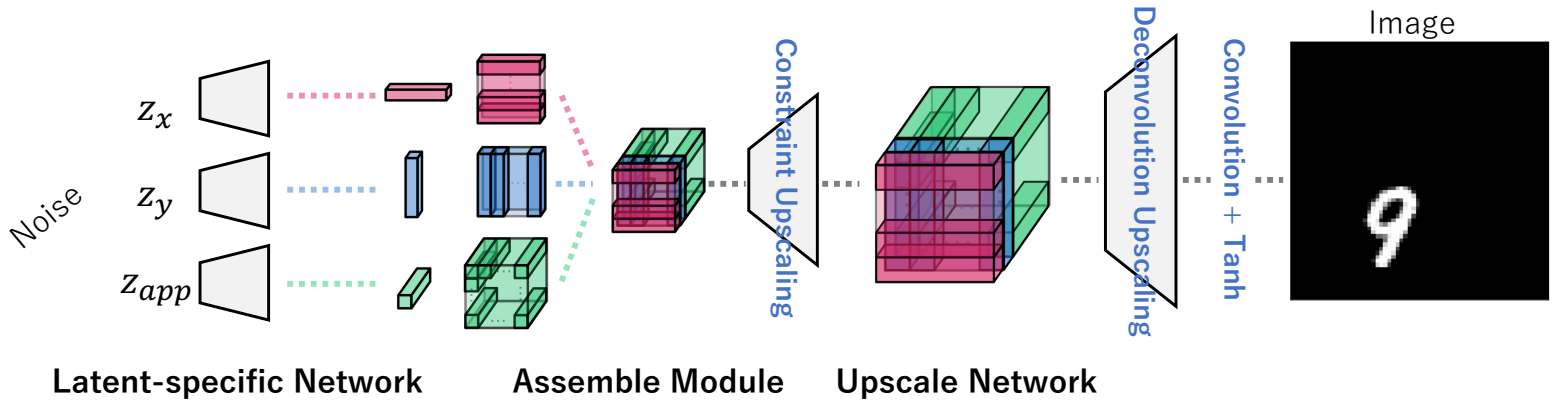
We perform constraint upscaling and deconvolution upscaling to synthesize images.

Constraint Upscaling



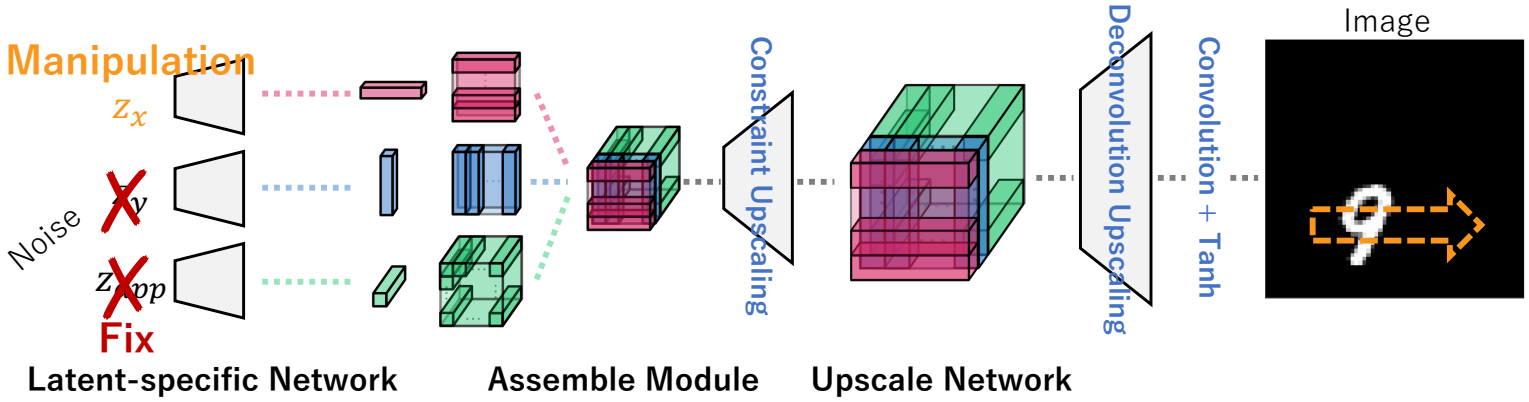
We upscale the constrained feature by pointwise convolution and spatial aggregation along with the given axis direction while maintaining the structural property.

Deconvolution Upscaling



We upscale by a vanilla deconvolution that ignores the property until output size.

How to manipulate images



We can manipulate the image by interpolation of the target factor while fixing the other factors

Visual Results on Translated MNIST

Random sampling by DAS

Interpolation of the appearance

Interpolation of the location

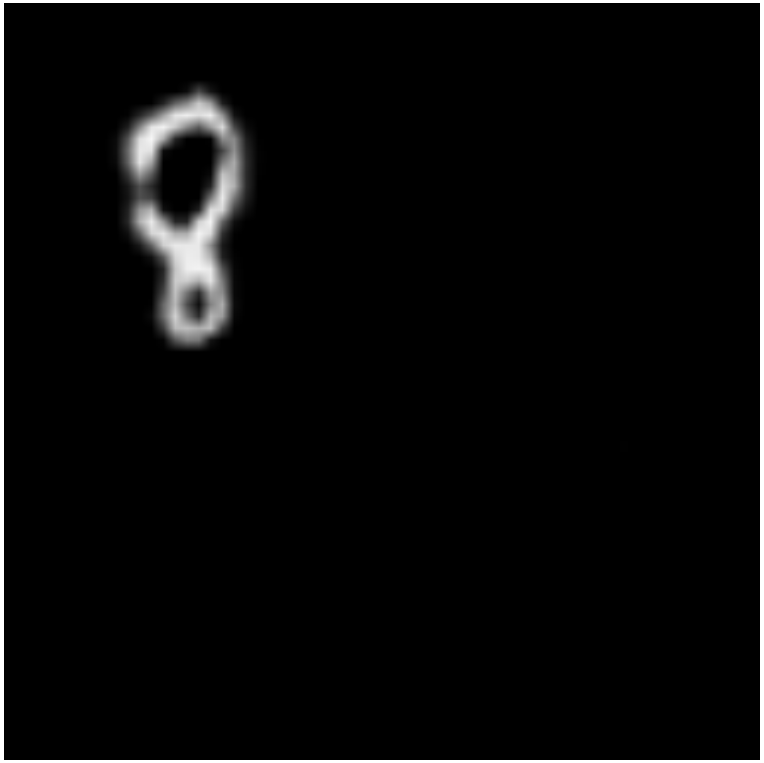
Random sampling by DAS

Interpolation of the appearance

Interpolation of the location

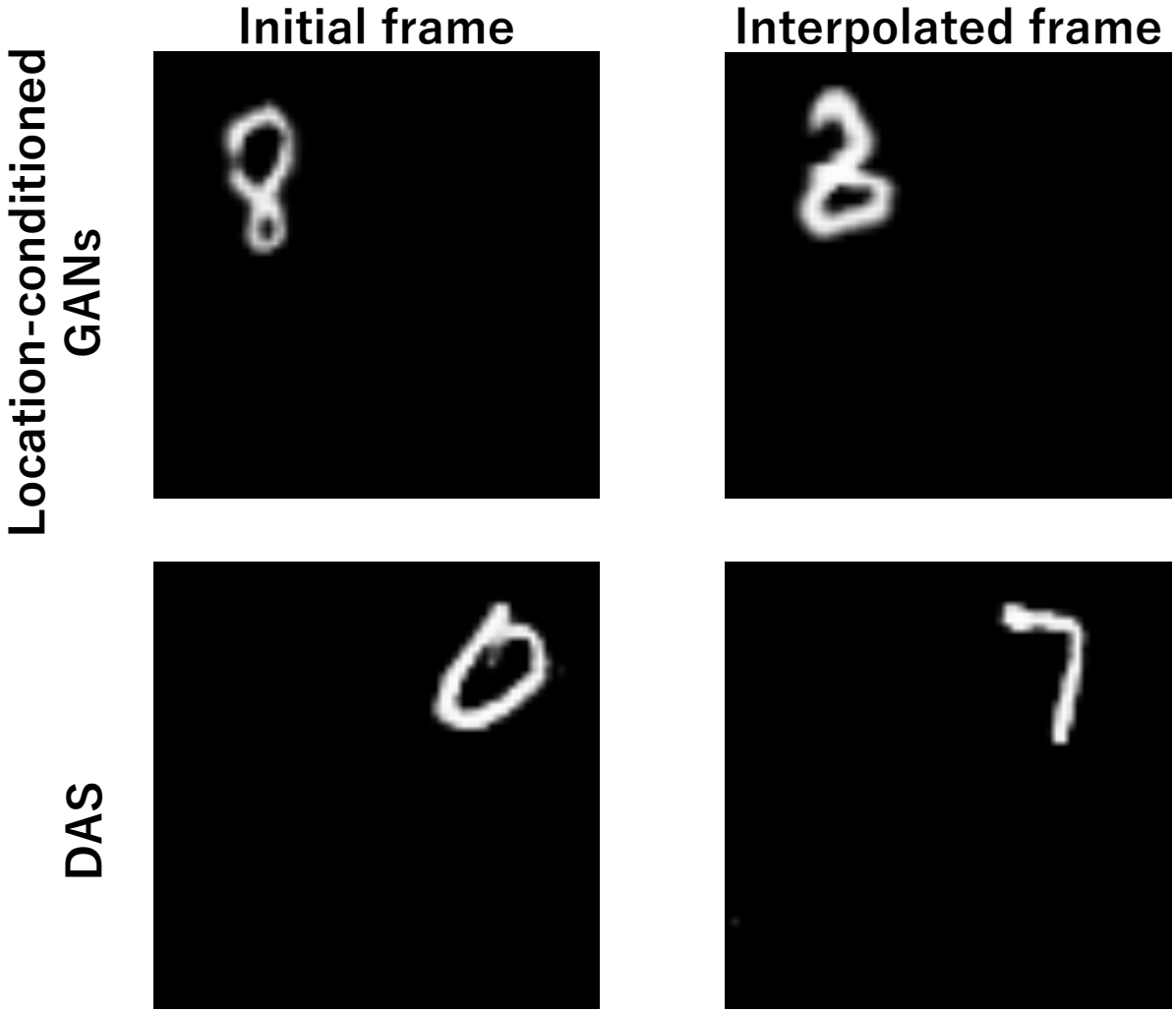
Interpolation of appearance while fixing the location

Location-conditioned GANs

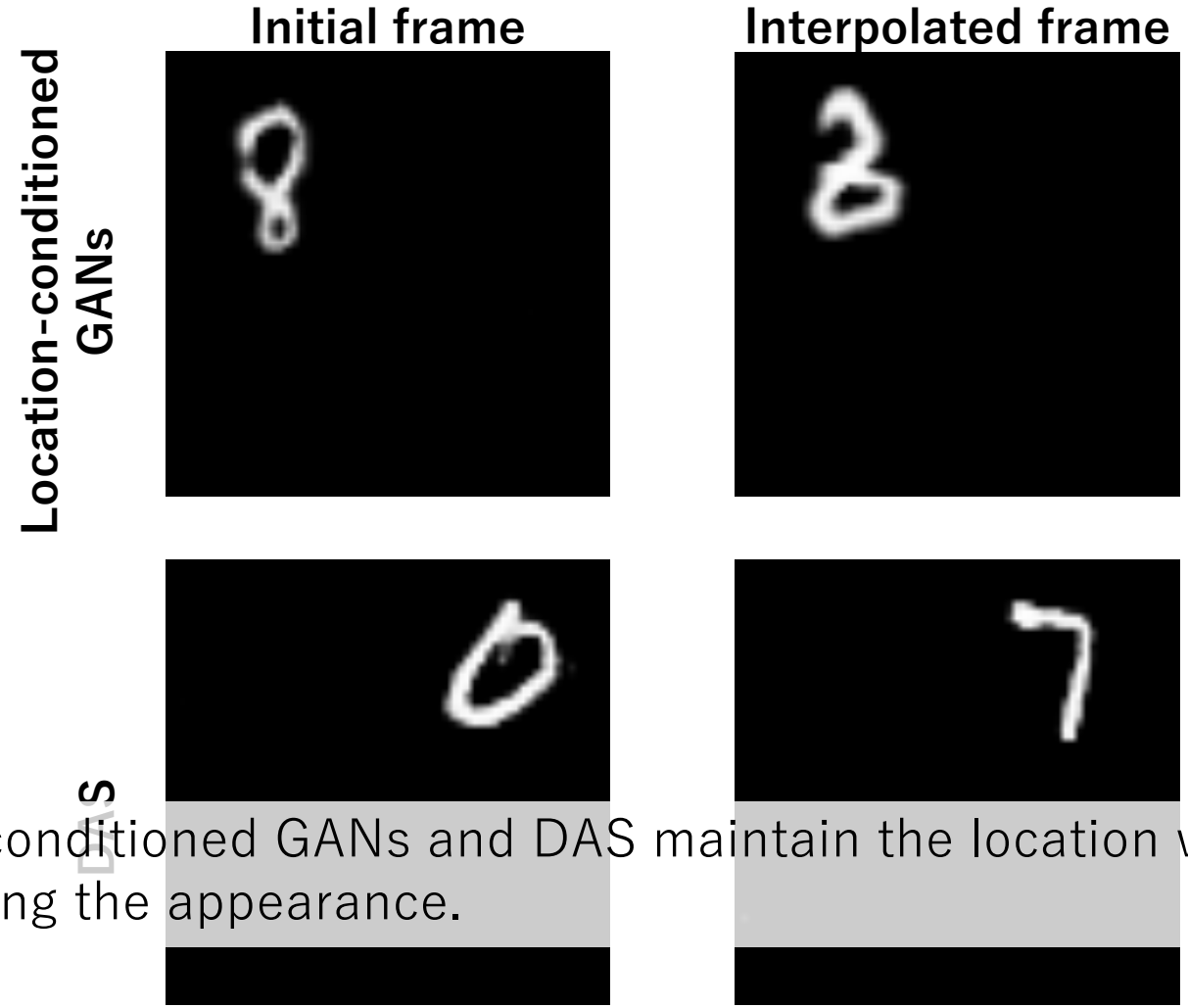


DAS

A comparison between DAS and Conditional GANs

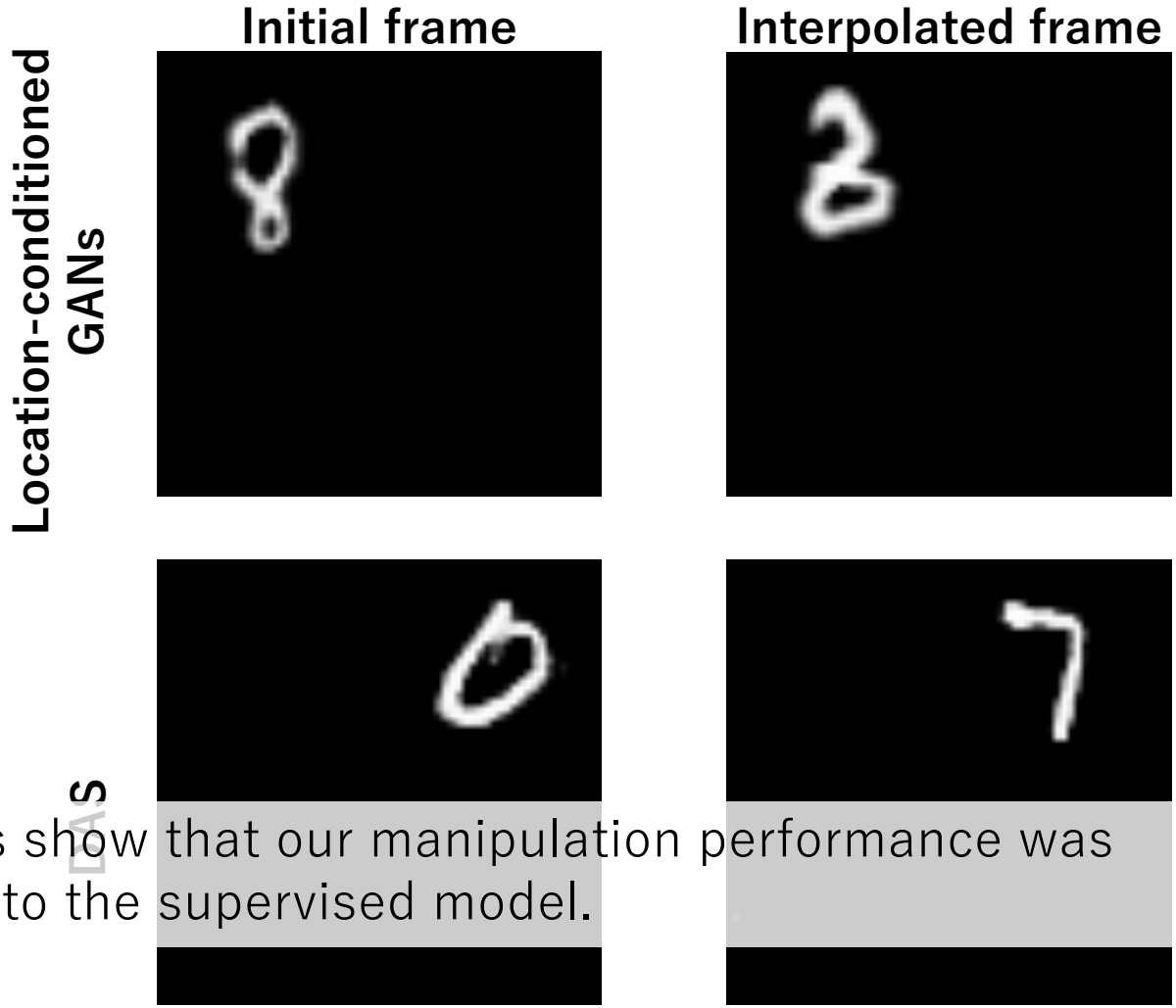


A comparison between DAS and Conditional GANs



Location-conditioned GANs and DAS maintain the location when manipulating the appearance.

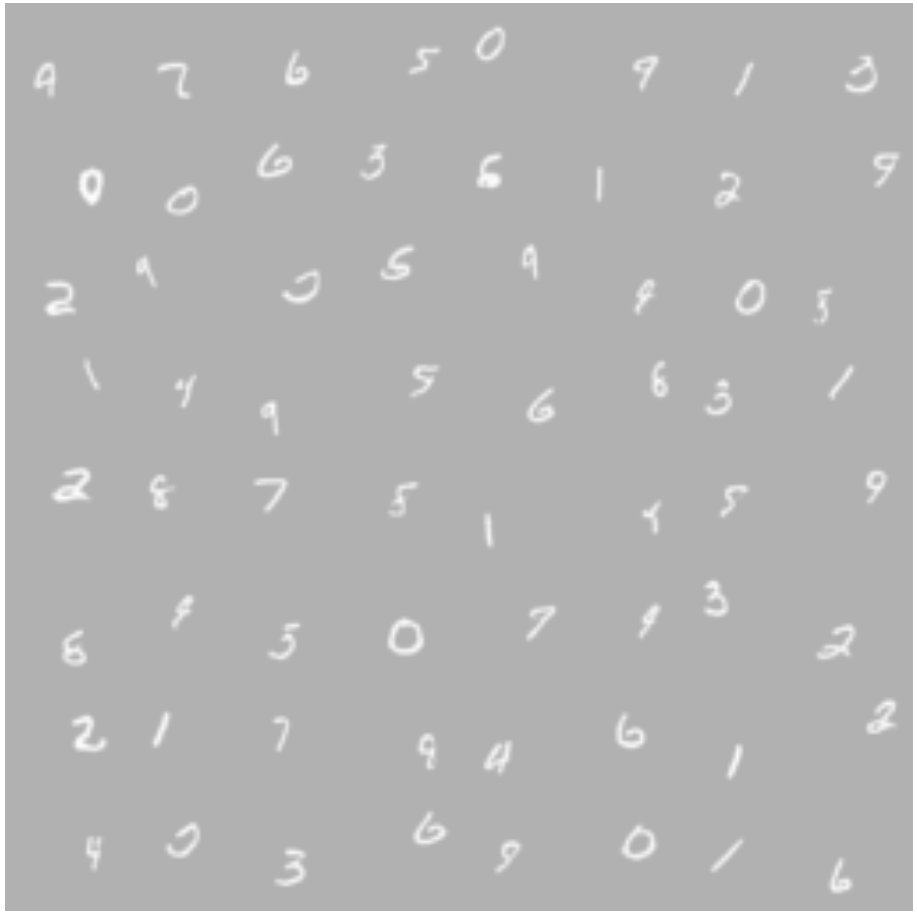
A comparison between DAS and Conditional GANs



The results show that our manipulation performance was equivalent to the supervised model.

Random sampling by DAS

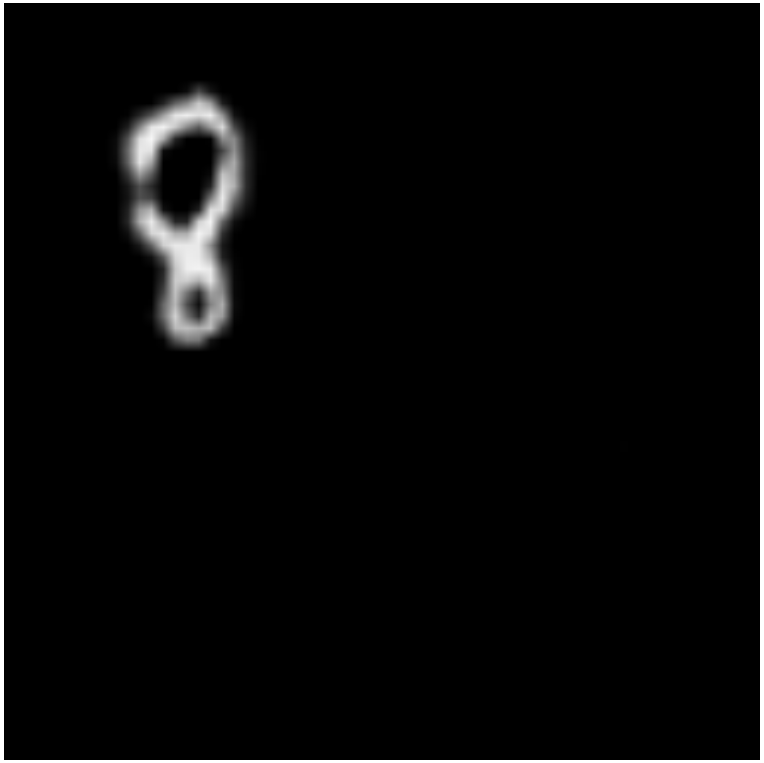
Interpolation of the appearance



Interpolation of the location

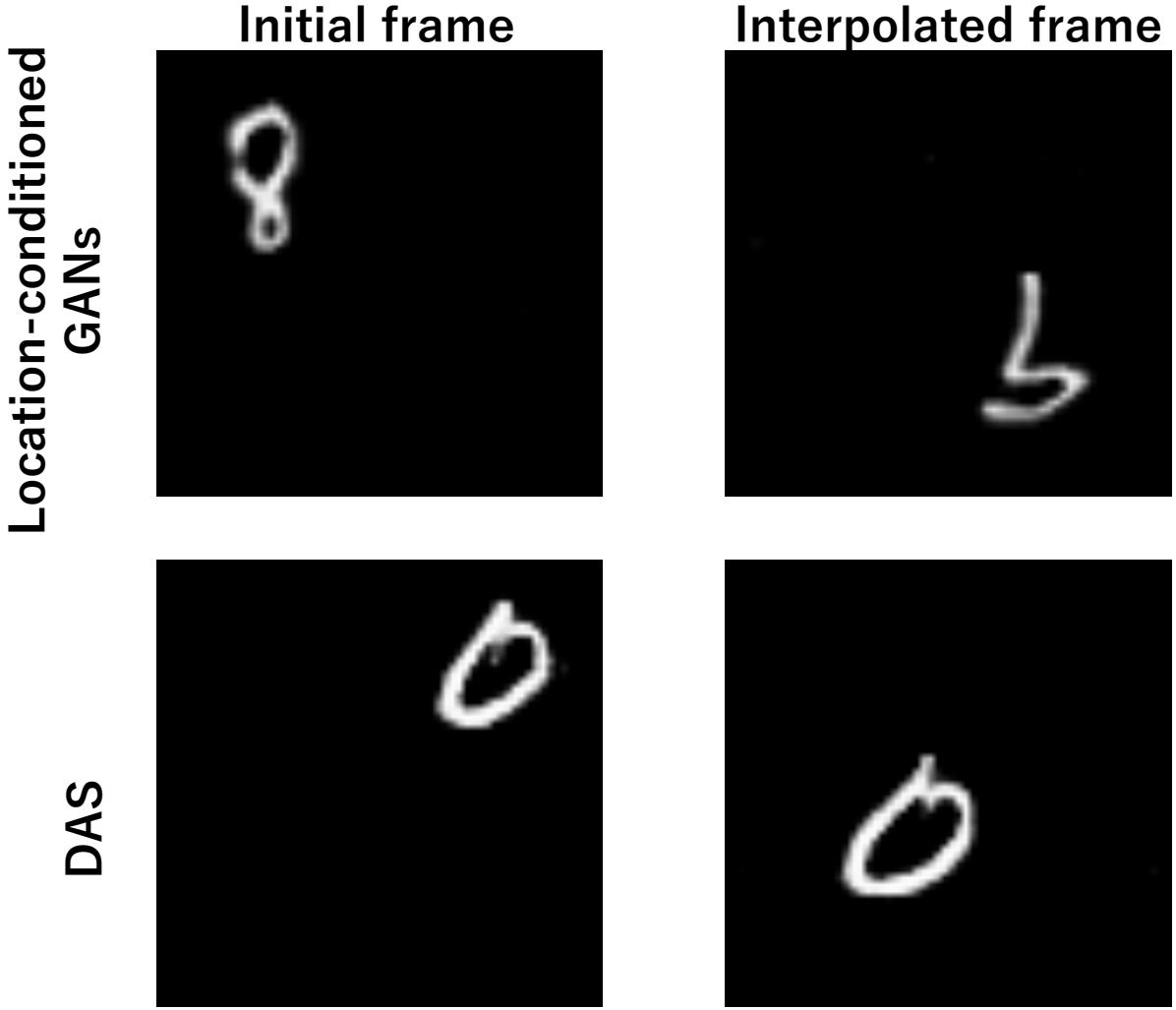
Interpolation of the location while fixing the appearance

Location-conditioned GANs

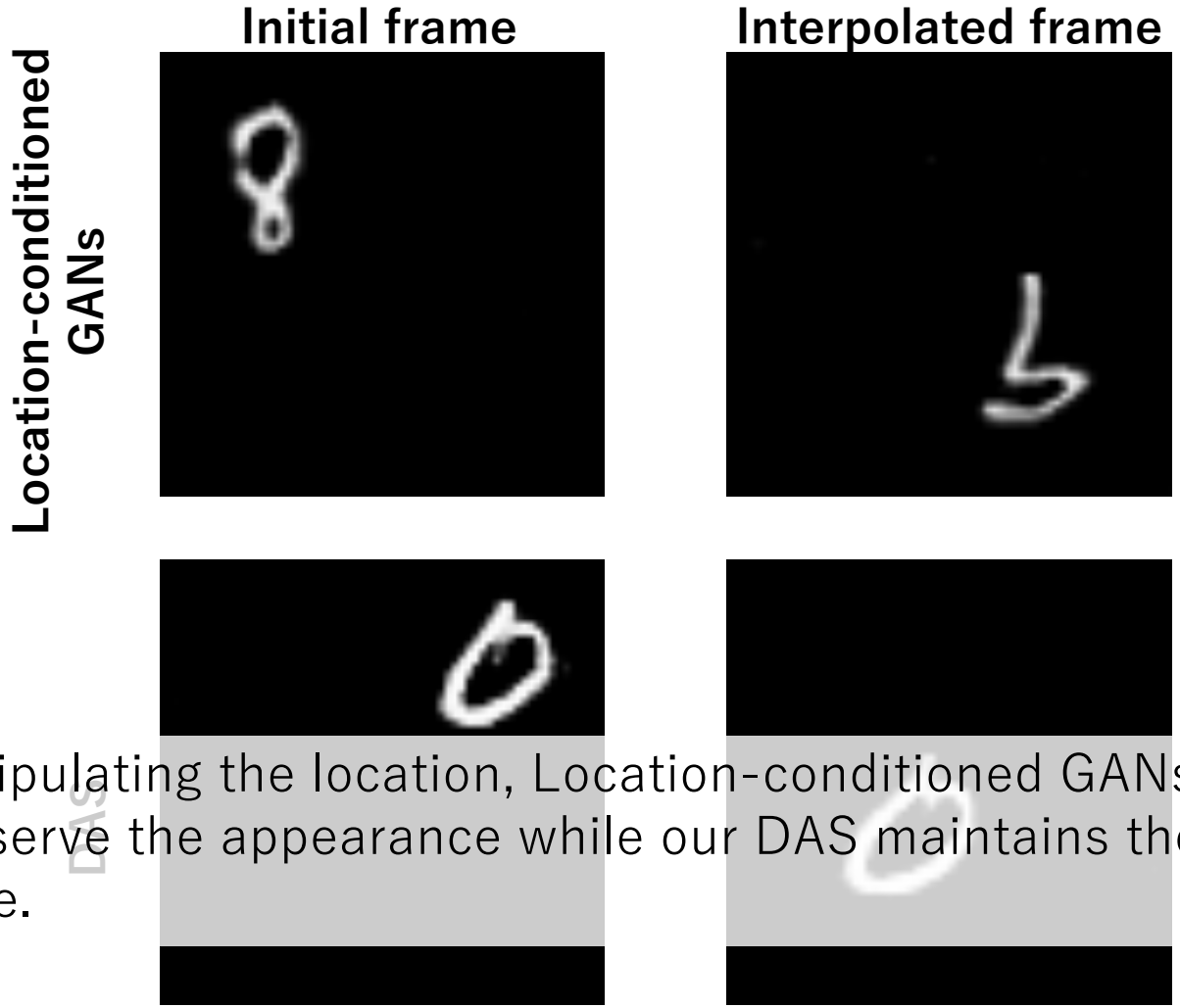


DAS

A comparison between DAS and Conditional GANs

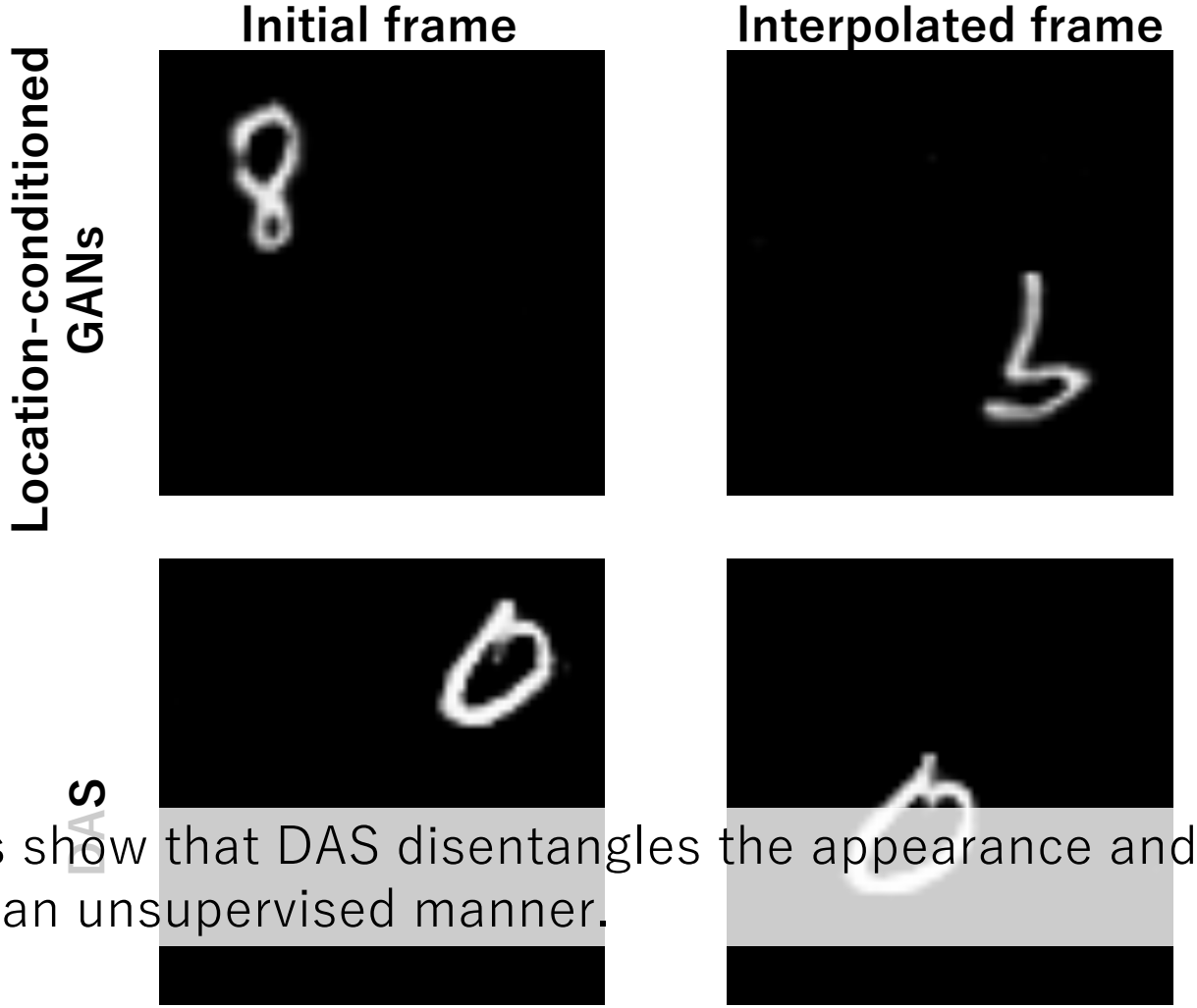


A comparison between DAS and Conditional GANs



When manipulating the location, Location-conditioned GANs do not preserve the appearance while our DAS maintains the appearance.

A comparison between DAS and Conditional GANs

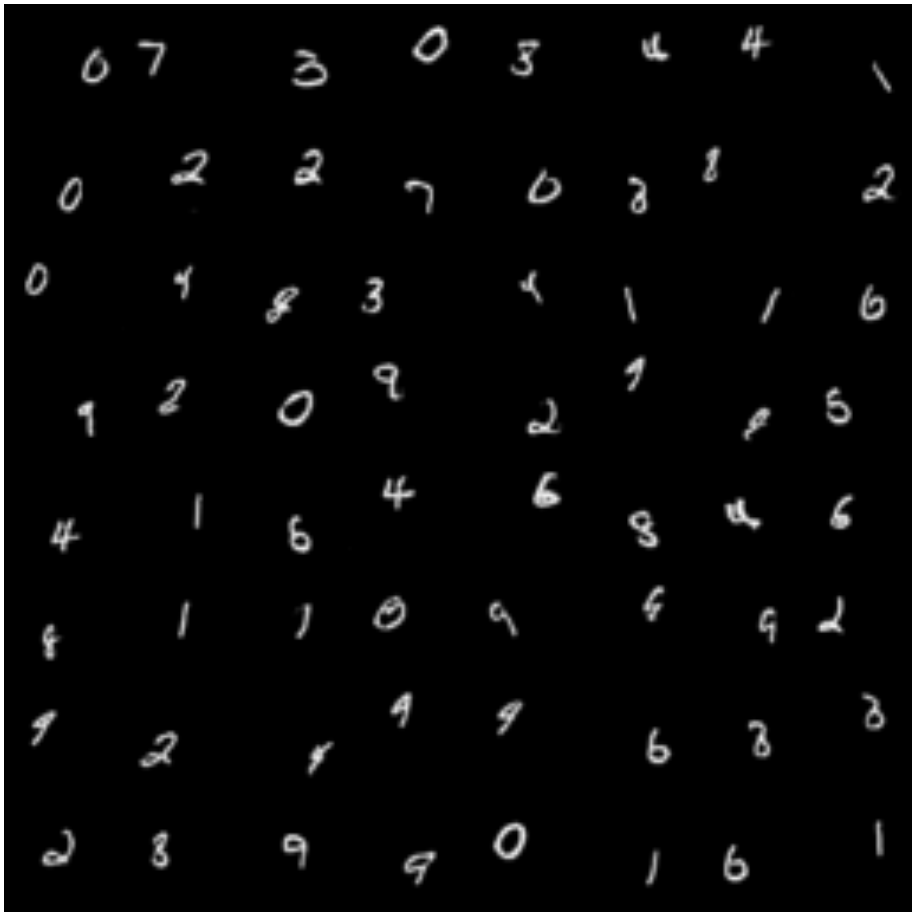


The results show that DAS disentangles the appearance and the location in an unsupervised manner.

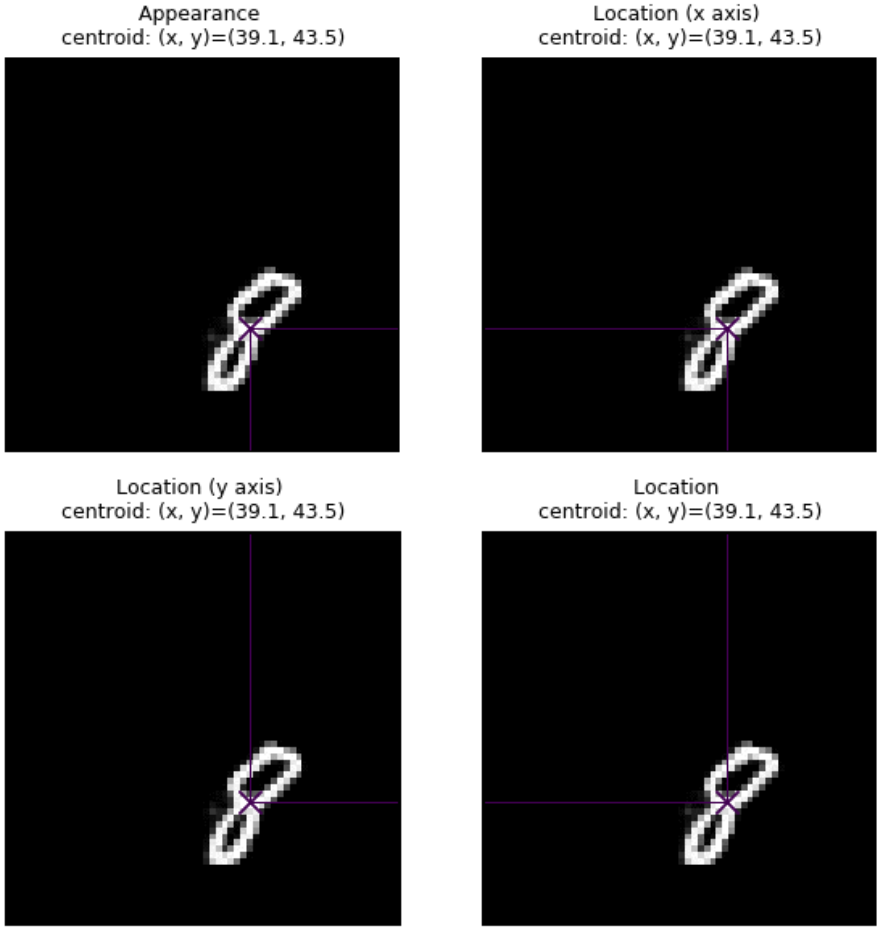
Random sampling by DAS

Interpolation of the x-axis direction

Interpolation of the y-axis direction



Detailed interpolation result



Our Contributions

- Our DAS learns to disentangle the appearance, the x–axis, and the y-axis factors, assemble them, and then synthesize images.

- Our DAS learns an explainable, compositional, manipulatable, and disentangled representation, opposite to GAN