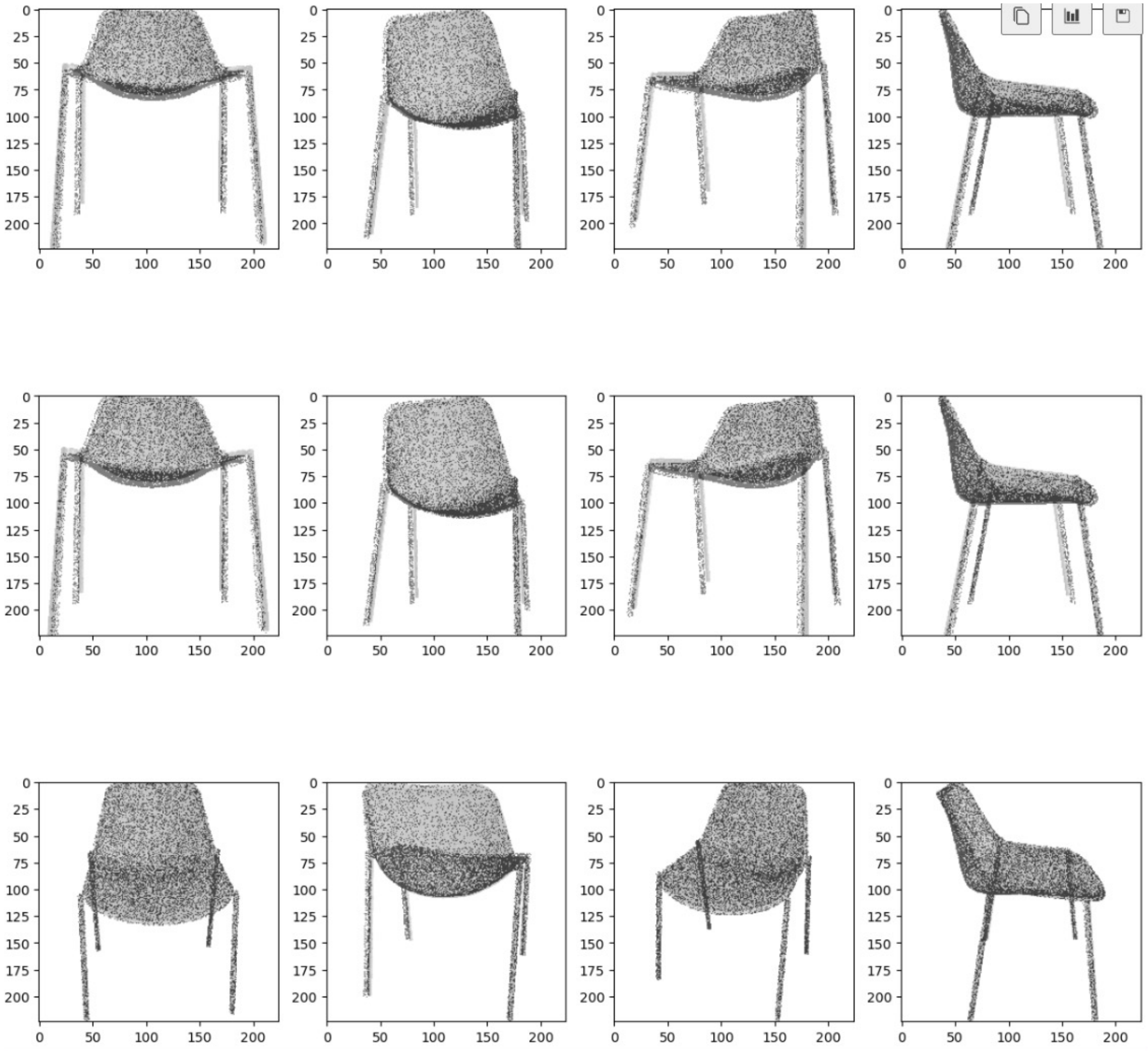
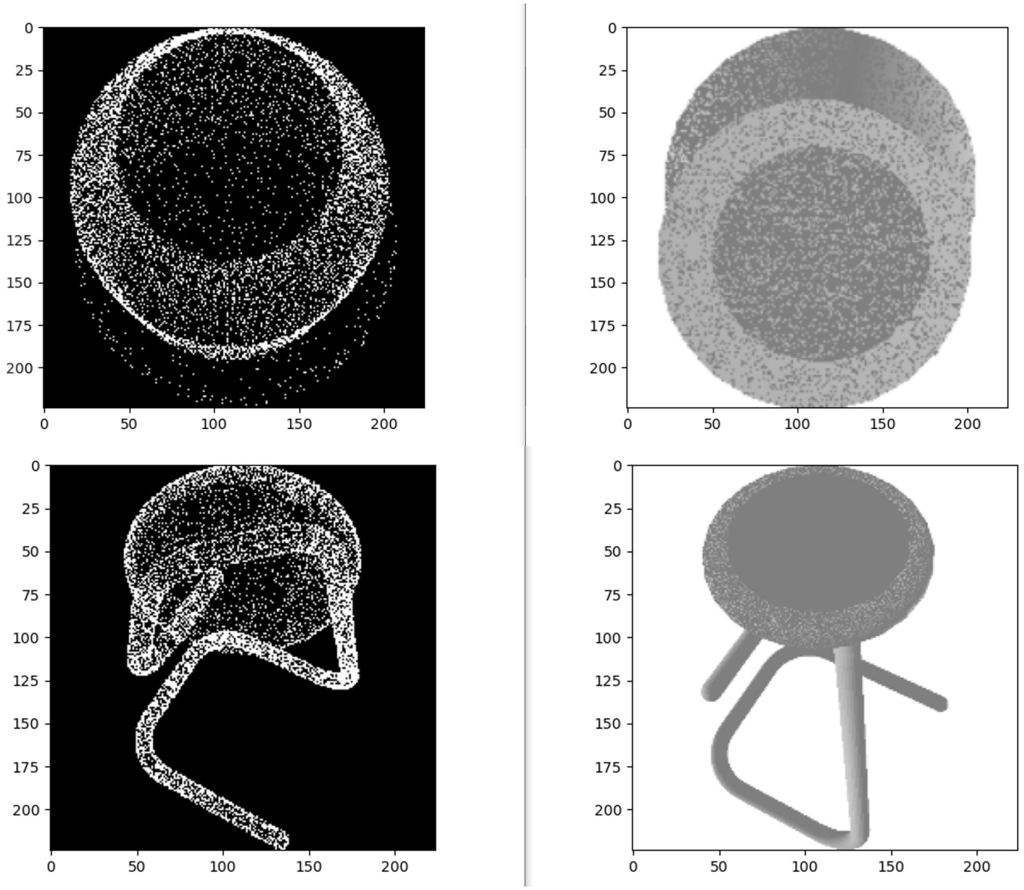
11/13 Report

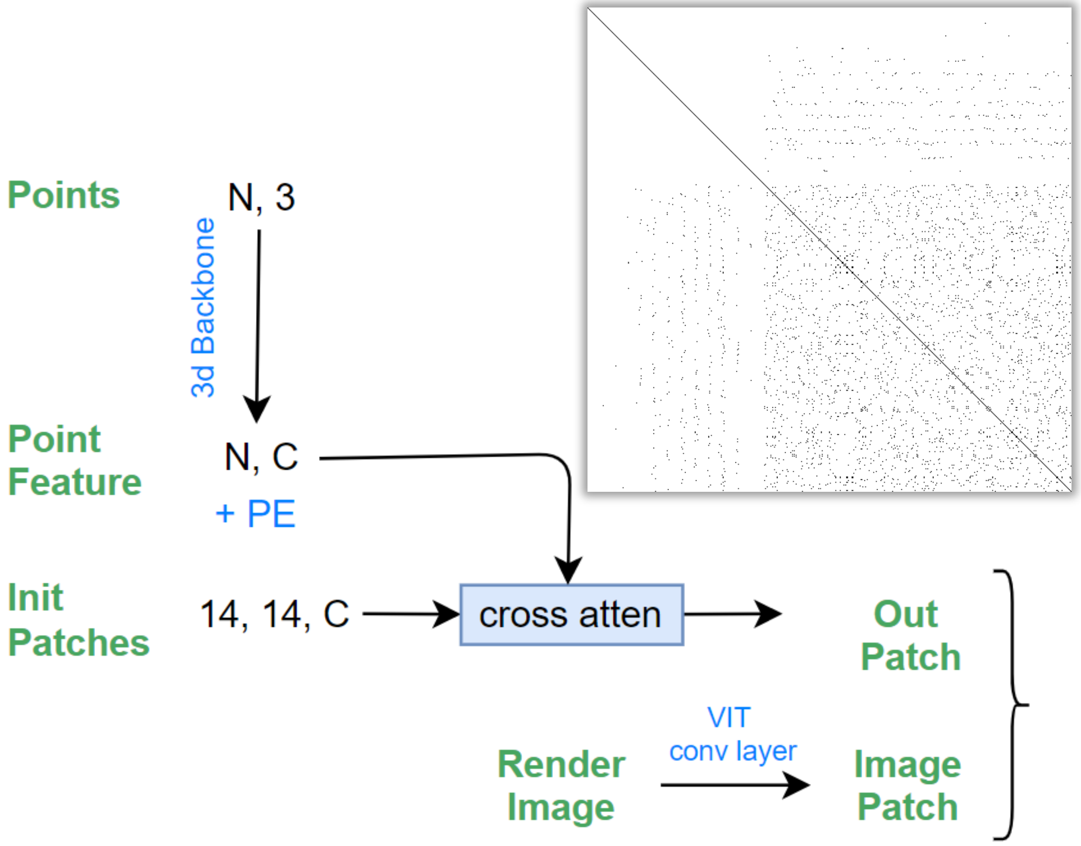
Project Timeline

- 09/19 - PointCLIP performance
  3d feature to 2d feature
- 09/25 - input层面
  P2P → patch → decoder；卡在few-shot达不到perf
- 10/24 - 区分Base & novel类
  decoder → self attention

Online Render & PC







Points

N, 3

3d Backbone

Point Feature

N, C

+ PE

Init Patches

14, 14, C

cross atten

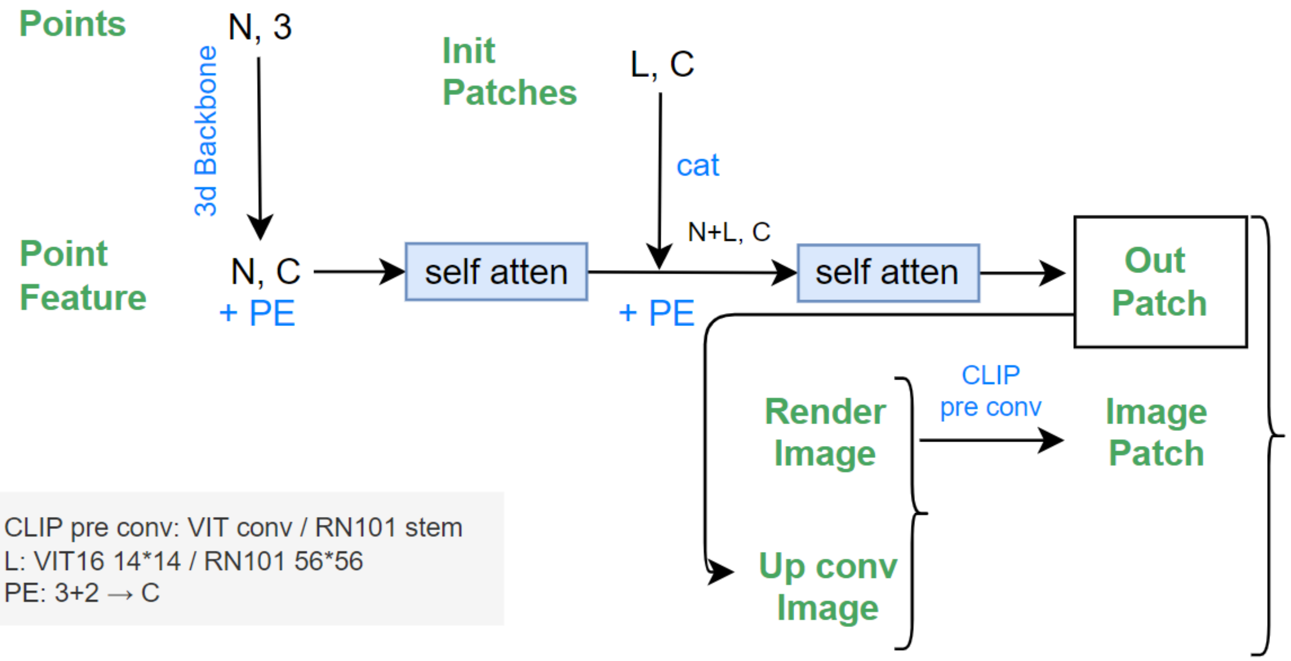
Out Patch

VIT conv layer

Render Image

Image Patch

Decoder



Points

N, 3

3d Backbone

Point Feature

N, C

+ PE

self atten

Init Patches

L, C

cat

N+L, C

+ PE

self atten

Out Patch

Render Image

CLIP pre conv

Image Patch

Up conv Image

CLIP pre conv: VIT conv / RN101 stem
L: VIT16 14*14 / RN101 56*56
PE: 3+2 → C

Self attention

Decoder: switch backbone & trans_dim / init patch / switch pe / up conv to img
SA: vit, rn101 / ablation / adjust loss

- Feature上的相近可能需要一个其他loss

  原图卷积后再反卷积，如果在feature上施加原来的loss，会导致原图也难以恢复

  (discriminator / contrastive loss / cosine with negative）

- 如果point → img feature的一步足够强，在img上直接用l1也足够建出图片
  - point backbone 训练问题
  - 给一些关注edge/structure的引导
  - 3d to 2d的dgcnn-like

- Self-supervised methods on features

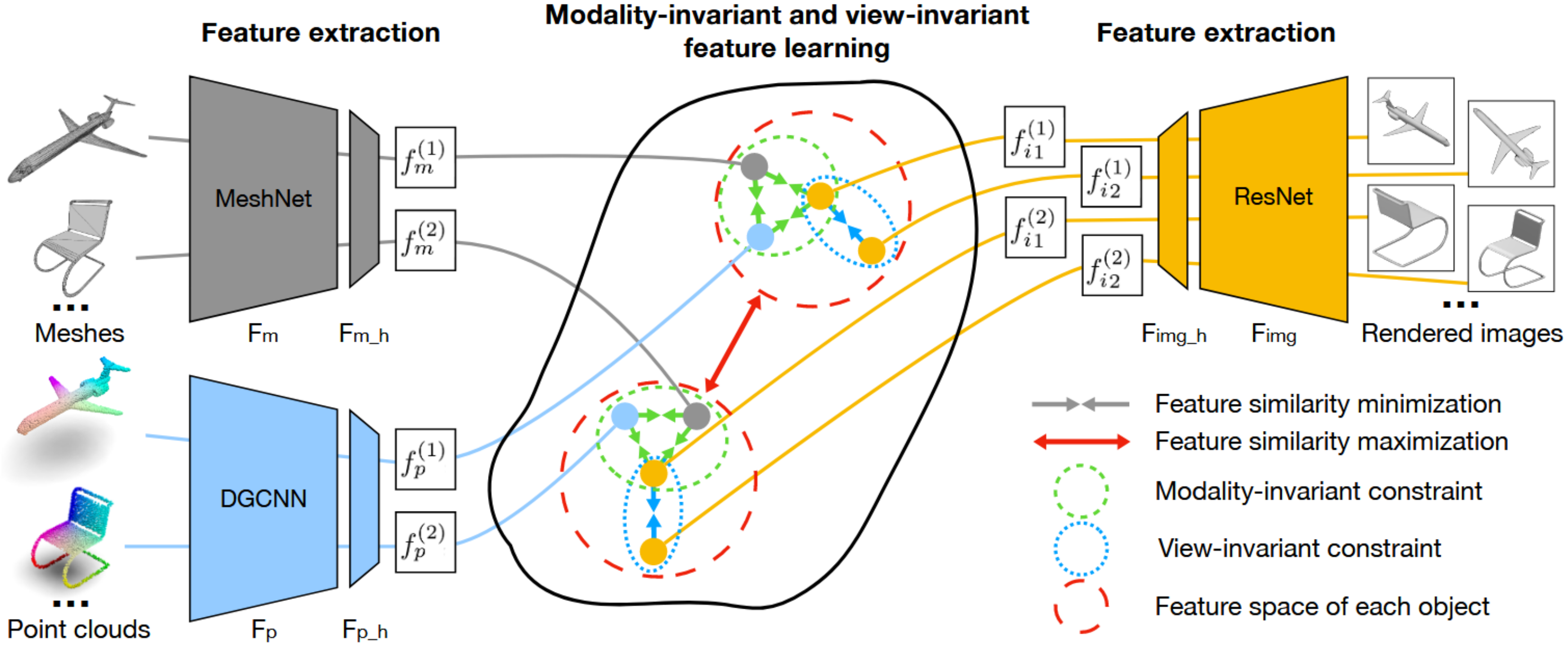
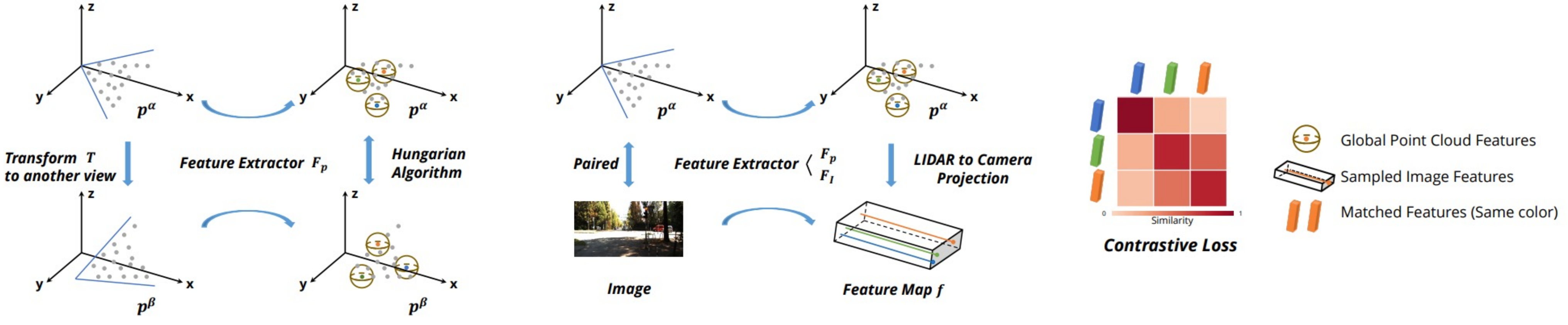


Figure 2. An overview of the proposed self-supervised modal- and view-invariant feature learning framework. Mesh, point cloud, and multi-view image features are extracted by MeshNet, DGCNN, ResNet, and corresponding projection heads, respectively. With contrastive learning to minimum the feature similarity of positive pairs and maximum the feature similarity of negative pairs under modality- and view-invariant constraints, the modal- and view-invariant features can be learned with the proposed heterogeneous framework in the same universal space.

Self-supervised Modal and View Invariant Feature Learning, Jing et al.
ICLR 2022 Conference Withdrawn Submission
Previous work 2021 CVPRW

- Self-supervised methods on features

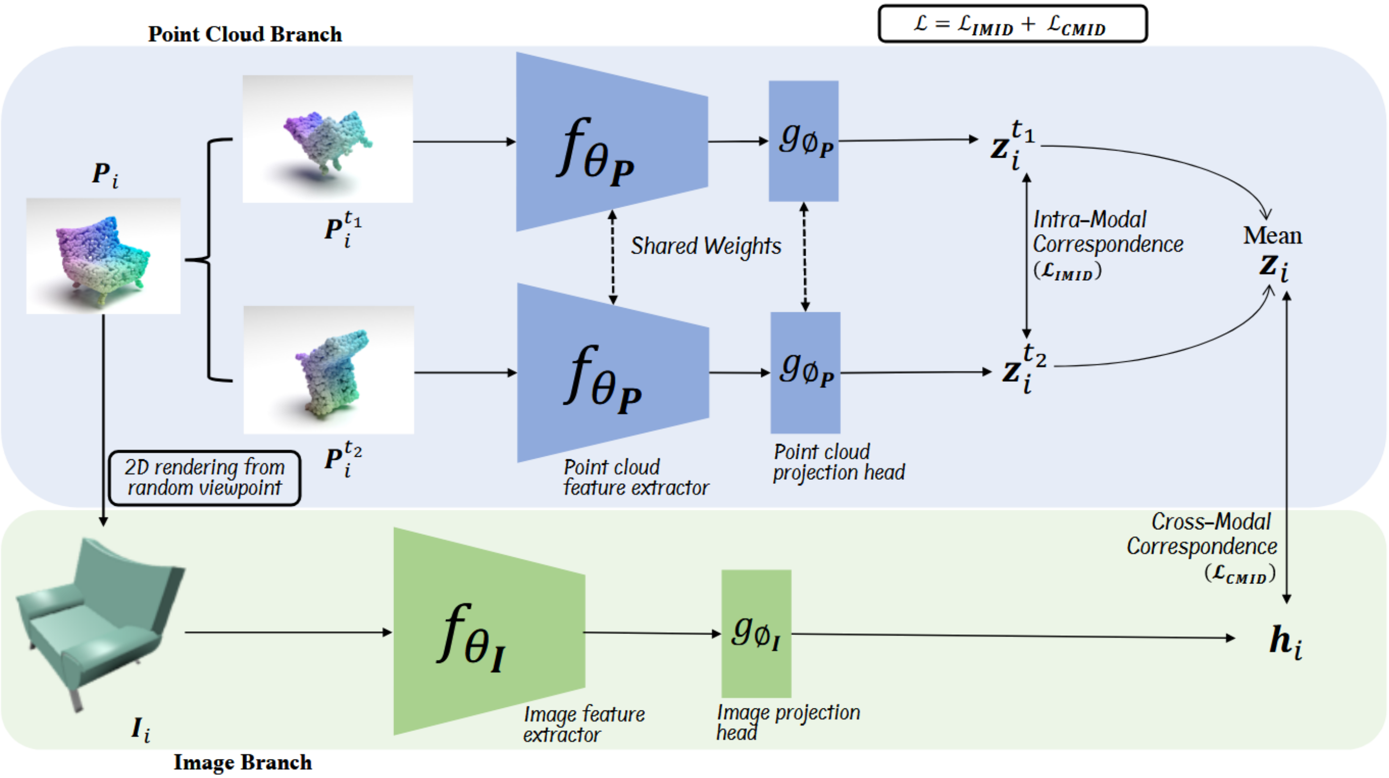


(a) Spatial Perception Module (b) Feature Interaction Module (c) Contrastive Losses

Figure 2: **Framework of SimIPU**. Matched pairs are in the same color. The whole framework is trained in an end-to-end manner. (a) **Intra-Modal Spatial Perception Module**: We utilize set abstraction layers to extract global point cloud features and downsample points (results are in color) from different views. The Hungarian Algorithm is applied to match the downsampled points according to locations. (b) **Inter-Modal Feature Interaction Module**: We adopt a standard ResNet-50 to extract global image features. Projection matrix from point cloud to image plane establish the association between positive pairs. (c) **Contrastive Loss**: Contrastive losses are applied to push closer the distances of matched pair features.

Simipu: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations.
AAAI2022

- Self-supervised methods on features

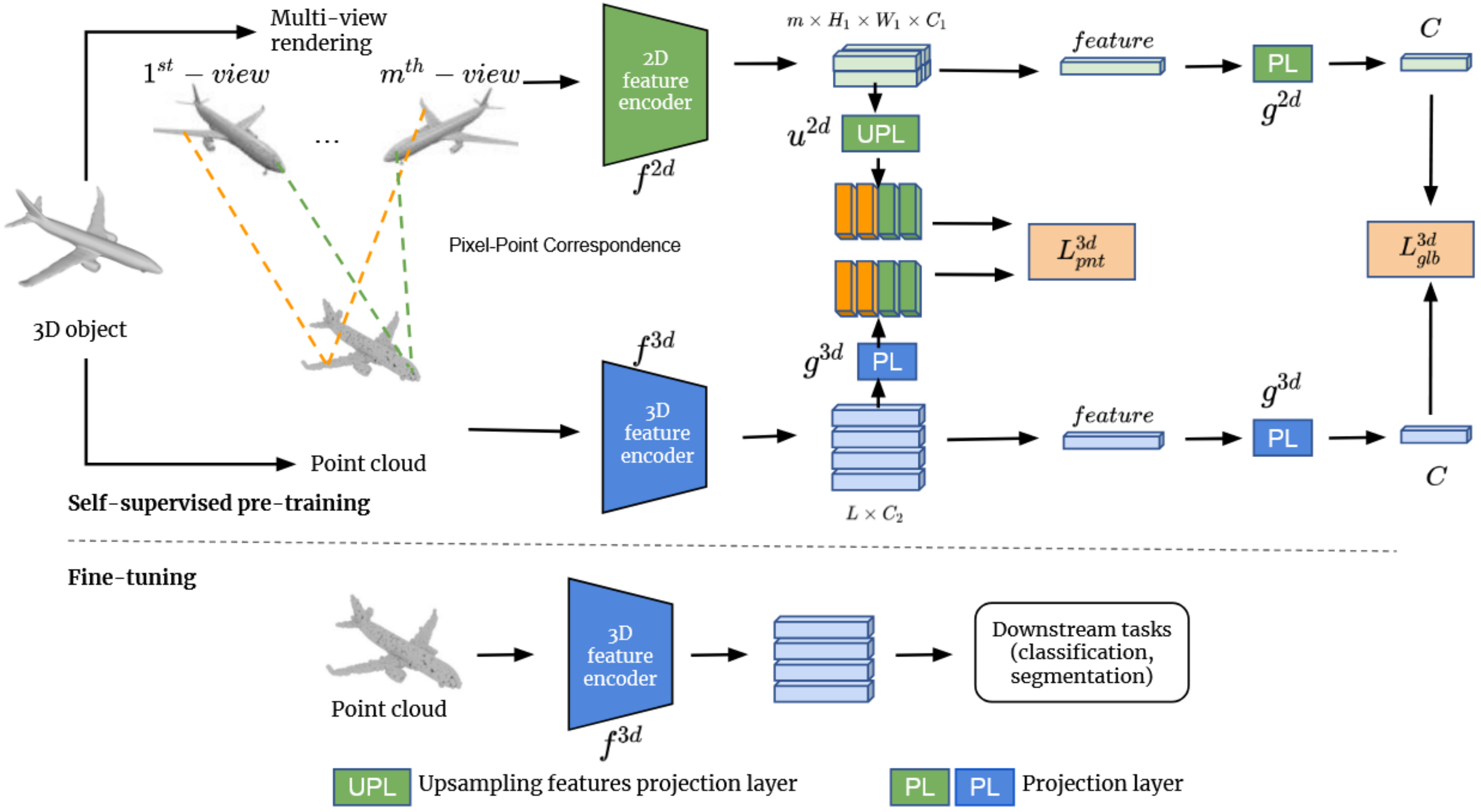


$\mathcal{L} = \mathcal{L}_{IMID} + \mathcal{L}_{CMID}$
Point Cloud Branch
P_i
$P_i^{t_1}$
$P_i^{t_2}$
2D rendering from random viewpoint
f_{θ_P}
g_{ϕ_P}
Shared Weights
Point cloud feature extractor
Point cloud projection head
$z_i^{t_1}$
$z_i^{t_2}$
Intra-Modal Correspondence ($\mathcal{L}_{IMID}$)
Mean
z_i
Cross-Modal Correspondence ($\mathcal{L}_{CMID}$)
I_i
f_{θ_I}
g_{ϕ_I}
Image feature extractor
Image projection head
h_i
Image Branch

Figure 2. The overall architecture of the proposed method (CrossPoint). It comprises of two branches namely: point cloud branch which establishes an intra-modal correspondence by imposing invariance to point cloud augmentations and image branch which simply formulates a cross-modal correspondence by introducing a contrastive loss between the rendered 2D image feature and the point cloud prototype feature. CrossPoint jointly train the model combining the learning objectives of both the branches. We discard the image branch and use only the point cloud feature extractor as the backbone for the downstream tasks.

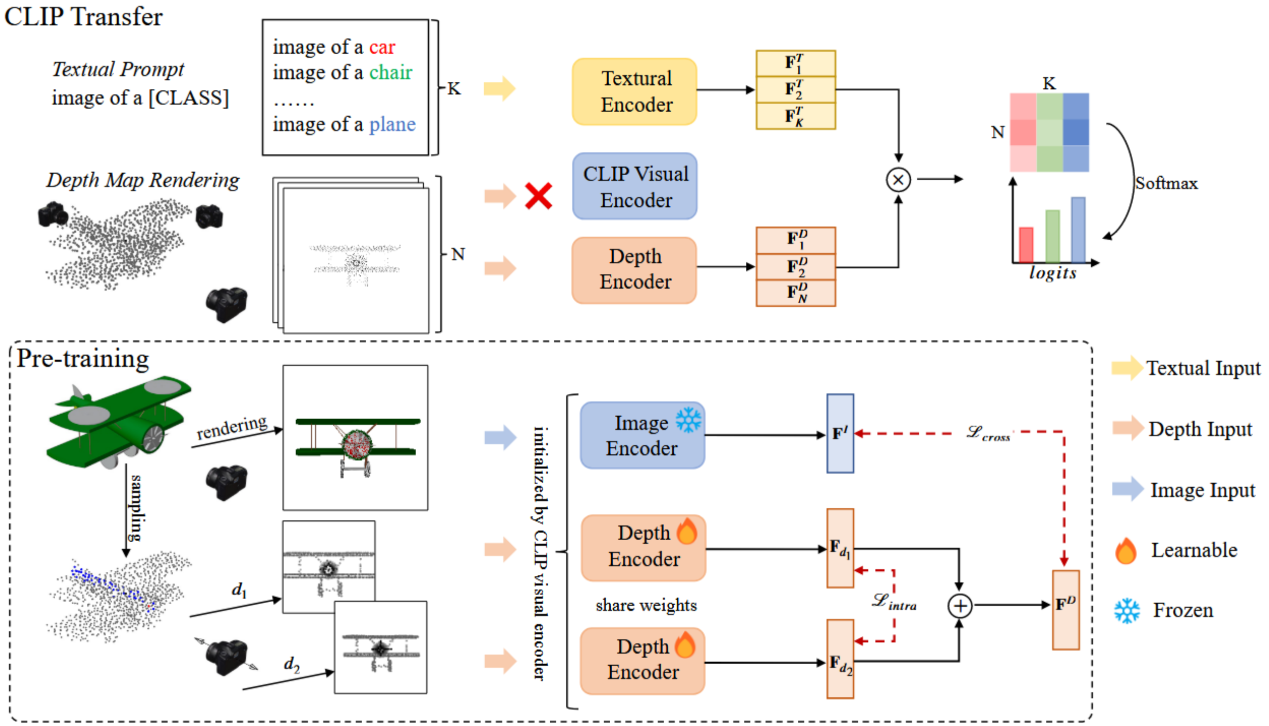
CrossPoint: Self-Supervised Cross-Modal Contrastive Learning for 3D Point Cloud Understanding, Afham et al. CVPR2022

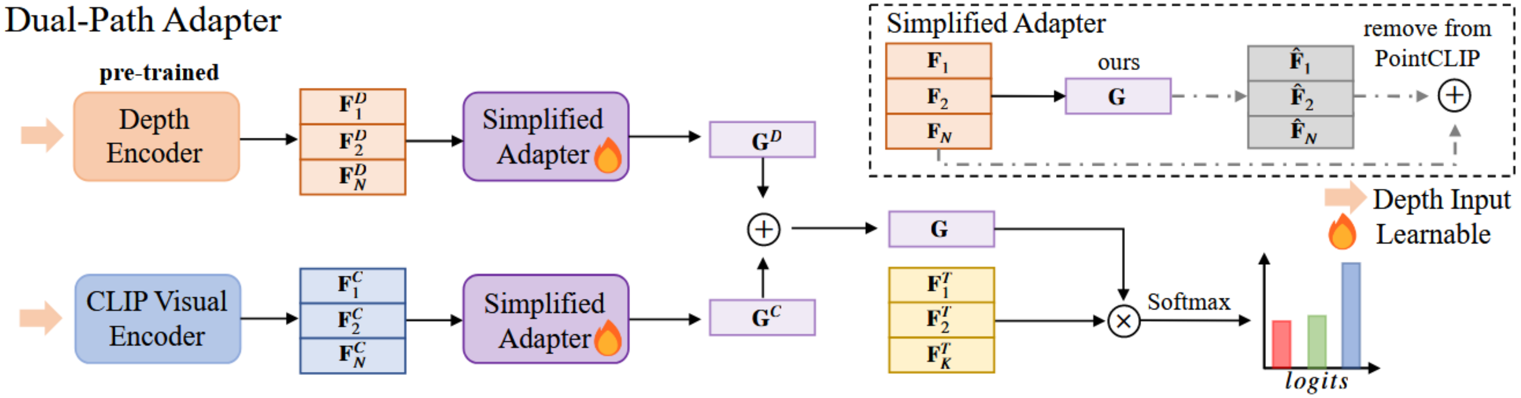
- Self-supervised methods on features



Self-Supervised Learning with Multi-View Rendering for 3D Point Cloud Analysis, Tran et al. ACCV2022

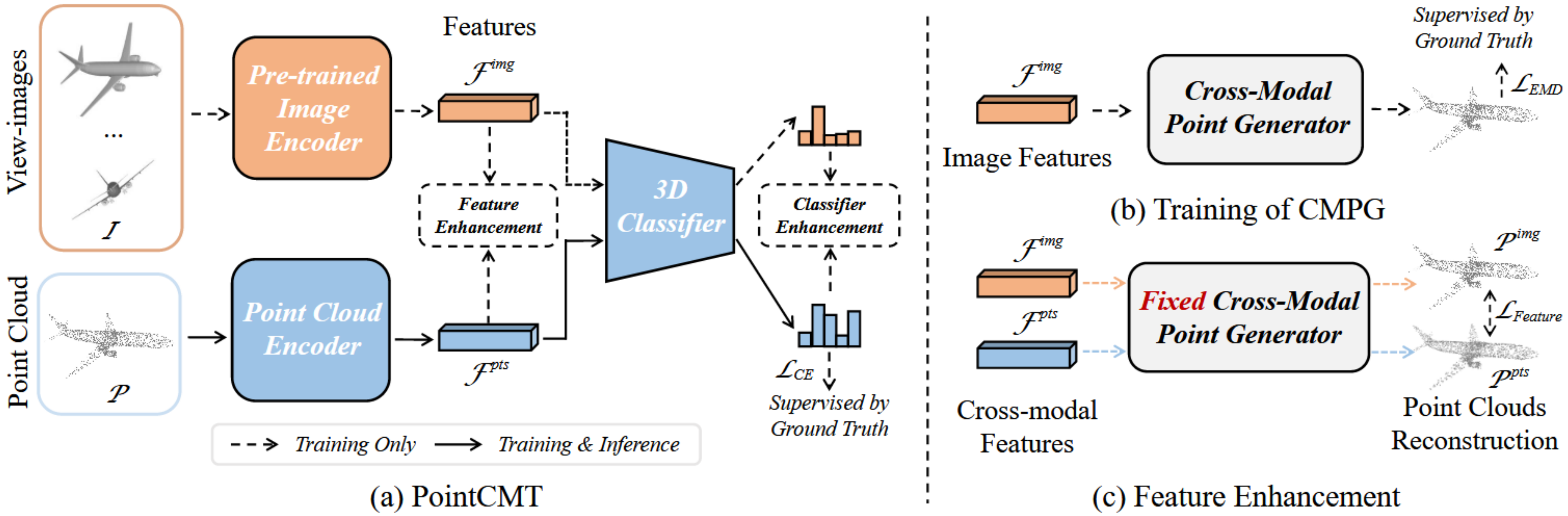
- Distillation





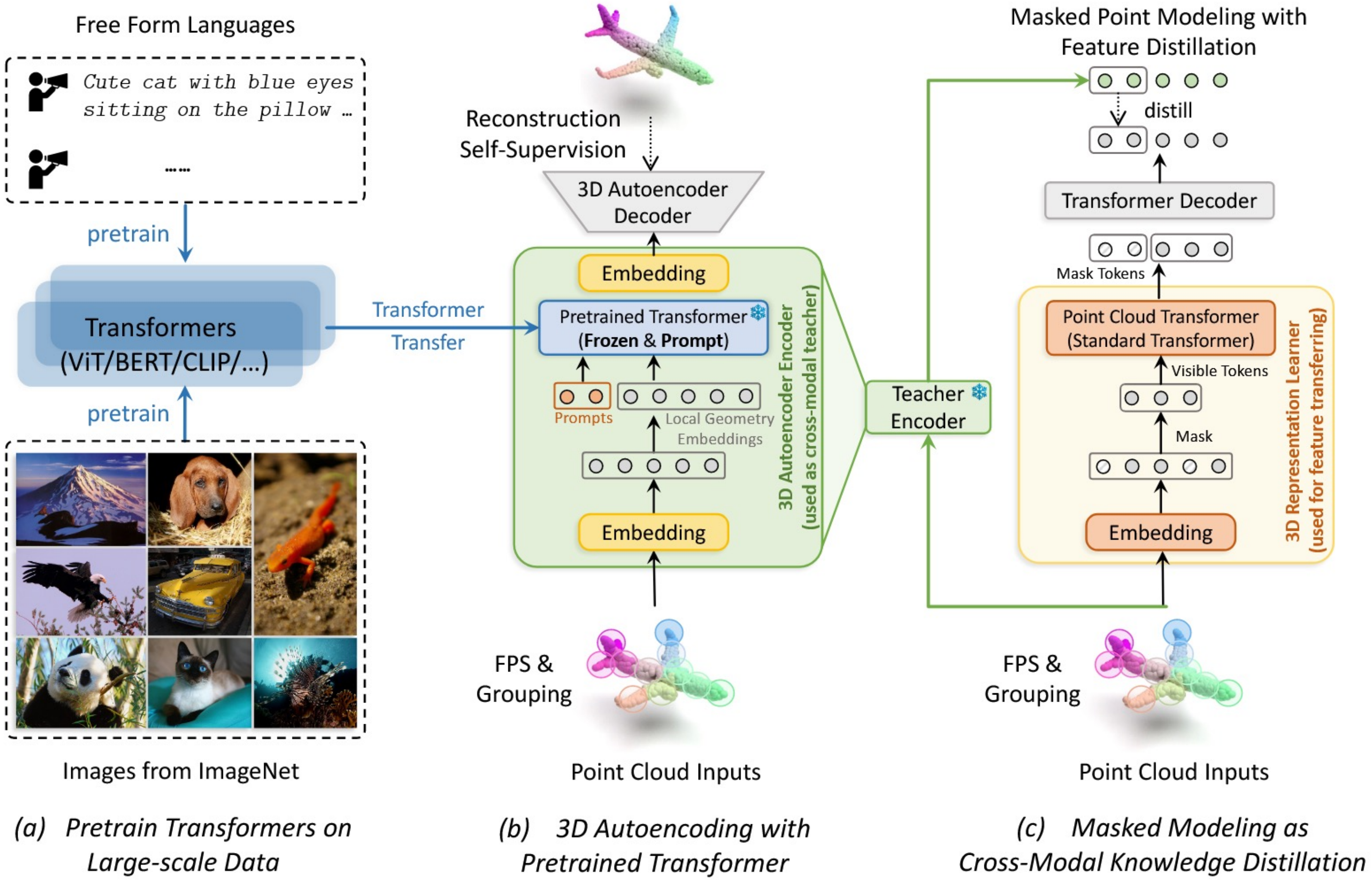
CLIP2Point: Transfer CLIP to Point Cloud Classification with Image-Depth Pre-training，Huang et al.
Arxiv preprint 2022

- Distillation



Let Images Give You More: Point Cloud Cross-Modal Training for Shape Analysis
NIPS2022

- Distillation



Free Form Languages
Cute cat with blue eyes sitting on the pillow …
……
pretrain
Transformers (ViT/BERT/CLIP/...)
pretrain
Images from ImageNet
Transformer Transfer
Reconstruction Self-Supervision
3D Autoencoder Decoder
Embedding
Pretrained Transformer (Frozen & Prompt)
Prompts
Local Geometry Embeddings
Embedding
3D Autoencoder Encoder (used as cross-modal teacher)
FPS & Grouping
Point Cloud Inputs
Teacher Encoder
Masked Point Modeling with Feature Distillation
distill
Transformer Decoder
Mask Tokens
Point Cloud Transformer (Standard Transformer)
Visible Tokens
Mask
Embedding
3D Representation Learner (used for feature transferring)
FPS & Grouping
Point Cloud Inputs

(a) Pretrain Transformers on Large-scale Data

(b) 3D Autoencoding with Pretrained Transformer

(c) Masked Modeling as Cross-Modal Knowledge Distillation

Autoencoders as Cross-Modal Teachers: can pretrained 2D image transformers help 3d representation learning?
Under review ICLR 2023