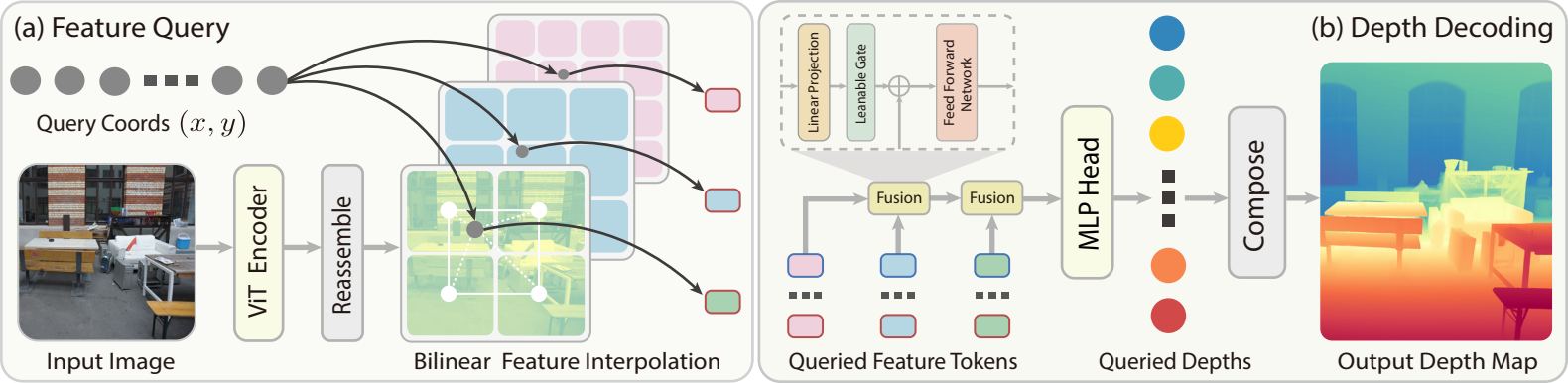
(a) Feature Query
Query Coords $(x, y)$
Input Image
ViT Encoder
Reassemble
Bilinear Feature Interpolation
Linear Projection
Leanable Gate
Feed Forward Network
Fusion
Fusion
Queried Feature Tokens
MLP Head
Queried Depths
Compose
(b) Depth Decoding
Output Depth Map